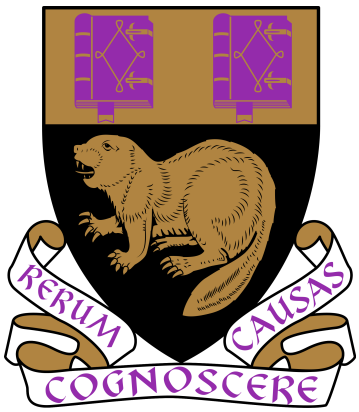
Statistical Inference for Some Choice Models

A thesis presented for the degree of

Doctor of Philosophy



Kaifang Zhou

Department of Statistics

The London School of Economics and Political Science

United Kingdom

March 2023

This thesis is dedicated to
the memory of my beloved grandfather
Zhenchang Zhou

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

Statement of inclusion of previous work

I can confirm that a version of Chapter 2.3 and parts of Chapter 2.7.8 were the results of the previous study for a Master's degree I undertook at the London School of Economics. The corresponding proofs for the previous work are included in Chapter 2.7.10.

Statement of co-authored work

Chapter 1 was jointly co-authored with my supervisor, Professor Milan Vojnović. A version of this chapter was submitted to The Annals of Applied Probability for review.

Chapter 2 was jointly co-authored with my supervisor, Professor Milan Vojnović, and Professor Se-Young Yun from KAIST. A version of this chapter was published in the Operations Research journal. Results from Chapter 2.3 were presented at AISTATS 2020 and published in the conference proceedings.

Acknowledgments

Completing this thesis has been a remarkable journey, and I owe a debt of gratitude to everyone who has helped me along the way.

First and foremost, I would like to express my heartfelt appreciation to my supervisor, Professor Milan Vojnović, for his unwavering support, insightful guidance, and patient encouragement. His vast knowledge, expertise, and willingness to provide constructive feedback have played an instrumental role in shaping the direction of my research, refining my methodology, and enhancing the overall quality of this thesis. I would also like to convey my deep gratitude to Professor Se-Young Yun for his valuable contributions to our research discussions and papers. Additionally, I extend my sincere thanks to my second supervisor, Dr. Yining Chen, for his continuous encouragement and guidance, which have been crucial throughout my academic journey.

Special thanks to my examiners, Dr. Yunxiao Chen and Dr. Ayalvadi Ganesh, for their meticulous assessment of my work and helpful comments they have provided. Their expertise and guidance have significantly enriched the quality of this thesis.

I thank the London School of Economics and Political Science, the Department of Statistics, and the Economic and Social Research Council (ESRC) for their generous support in providing me with the ESRC 1+3 PhD Scholarship. This scholarship has been instrumental in enabling me to undertake this research, and I am truly grateful for the opportunity.

Many thanks go to the staff in the Department of Statistics for their invaluable support and assistance throughout my PhD journey. I am especially grateful to Penny Montague and Imelda Noble, who were always there to help me with any questions or concerns I had. Additionally, I am thankful to Dr. James Abdey and Dr. Marcos Barreto for providing me with excellent teaching opportunities. Being their teaching assistant was a great experience that I will always cherish. I would also like to extend a

special thank you to Professor Angelos Dassios for his support and encouragement at the beginning of my PhD journey. His guidance and advice were crucial in helping me get started on the right foot.

I am truly grateful to my friends and colleagues in Col 7.03, JingHan Tee (especially!), Jose Manuel Pedraza Ramirez, Anica Kostic, Sahoko Ishida, Davide de Santis, Sasha Tsimbalyuk, Gianluca Giudice, Wonbong Jang, for their camaraderie, support, and stimulating conversations. Their feedback, encouragement, and willingness to share their own experiences have been invaluable in keeping me motivated, inspired, and on track.

I am so lucky to have my dearest friends, Angela Jiang, Xinxin Zhang, Xinyi lu, Ke Wang, Alicia Chiang, Sisi Feng, Daidai who are always to be my side. Their friendship has been a beacon of light during the darkest moments of my academic journey, and their positive energy has lifted me up when I felt discouraged or overwhelmed.

I thank my parents, Jinqin Zhou and Xuehong Chen, my grandparents, Youjia Chen and Xiuying Guo for unconditional love, sacrifice, and belief in me. I also want to thank my brother, Wenkai Zhou, for being a great friend to me since our childhood. Though we fought a lot when we were young, we eventually made a pact to be lifetime friends, and we have honored that commitment ever since.

I would like to thank my loyal and loving pawfriends, Maple and Hazel. They have been by my side every day, providing me with companionship, comfort, and endless joy. Their wagging tails and playful antics have brightened my darkest days, and their quiet presence has been a soothing balm during the long hours of research and writing.

Last but not the least, I will not forget to thank my lifetime partner, Alexander Lye, who has been my unwavering source of support, love, and encouragement. He has been with me every step of the way, offering words of encouragement when I needed them the most, listening to my worries and doubts, and cheering me on when I succeeded. I am blessed to have him by my side.

Abstract

This thesis comprises two chapters that study the statistical inference problems for two types of choice models, namely, the discrete voter model and the Bradley-Terry models, respectively.

In Chapter 1, we consider a discrete-time voter model process on a set of nodes, each being in one of two states, either 0 or 1. In each time step, each node adopts the state of a randomly sampled neighbour according to sampling probabilities, referred to as node interaction parameters. We study the maximum likelihood estimation of the node interaction parameters from observed node states for a given number of realizations of the voter model process. We present parameter estimation error bounds by interpreting the observation data as being generated according to an extended voter process that consists of cycles, each corresponding to a realization of the voter model process until absorption to a consensus state. We present new bounds for all moments and a probability tail bound for consensus time. We also present a sampling complexity lower bound for parameter estimation within a prescribed error tolerance for the class of locally stable estimators.

In Chapter 2, we study the popular methods for inference of the Bradley-Terry model parameters, namely the gradient descent and MM algorithm, for maximum likelihood estimation and maximum a posteriori probability estimation. This class of models includes the Bradley-Terry model of paired comparisons, the Rao-Kupper model of paired comparisons allowing for tie outcomes, the Luce choice model, and the Plackett-Luce ranking model. We propose a simple modification of the classical gradient descent and MM algorithm with a parameter rescaling performed at each iteration step that avoids the observed slow convergence issue that we found in our previous work (Vojnovic et al. [2020]). We study the convergence rates of accelerated gradient descent and MM Algorithms for Bradley-Terry models. We also produce some experimental results using synthetic and real-world data to show that significant efficiency gains can be obtained by our new proposed method.

Contents

Introduction

Choice models are mathematical and statistical models that provide a framework for studying individual decision-making and understanding how individuals make choices based on different factors, such as their preferences, beliefs, and information. These models are widely used in various fields. For example, companies use choice models in marketing to understand consumers' decision-making process when facing several options. By applying choice models, companies can better understand how different factors, such as price, quality, and branding, affect consumers' preferences and ultimately influence their purchase decisions. Similarly, in psychology, choice models are used to understand how individuals make choices based on their beliefs, values, and emotions. By understanding how individuals make choices, psychologists can develop interventions that help people make better decisions and improve their overall well-being. In social networks, choice models are used to understand how individuals interact and make choices to adopt opinions from their neighbors in a network. These models provide insights into how ideas and information spread among individuals and how individuals' opinions are influenced by their neighbors. Additionally, in online gaming, choice models can be used to provide rankings of players. These models consider various factors, such as player skills, strategies, and performance, to provide a fair and accurate ranking of players. This information is helpful for matchmaking, tournament organization, and player evaluation.

The voter model is an instance of choice models used to understand how entities make choices when updating their opinions. The voter model process focuses on the spread of opinions within a population, where entities make choices to update their opinions based on the observed opinions that are chosen by other entities. The concept of the voter model, introduced in Holley and Liggett [1975], represents a continuous-time Markov process where each individual is in one of two possible states, either state 0 or 1. In this model, individuals adopt the state of a randomly sampled neighbor at random time instances through independent Poisson processes associated with individuals or links connecting them. The voter model is classified as an interacting particle system (Liggett [1985]). In Granovsky and Madras [1995], the noisy voter model was proposed as an extension of the classic voter model by including spontaneous flipping of states from 0 to 1, and vice versa, for each node. The discrete-time voter model is similar to the continuous-time model, except that node states are updated synchronously at discrete time steps. The discrete-time voter model has been studied under different assumptions about which nodes update their states at discrete time points, such as those presented in Nakata et al.

[1999], Hassin and Peleg [2001], and Cooper and Rivera [2016]. Studies of dynamics and learning in social and economic networks have been pursued from different perspectives, including dynamical systems, stochastic processes, and statistical perspectives, e.g. see Jackson [2008], Kolaczyk [2009], and Easley and Kleinberg [2010]. The key research questions include understanding the long-run behavior of the underlying random dynamical system, time to convergence to a consensus state when such a limit behavior arises, and statistical inference of model parameters from observed data, e.g. inferring node interaction rate parameters from observed node states over time.

The Bradley-Terry model and its generalizations are choice models used for evaluating choice preferences based on ranking scores computed by observed data from various comparison outcomes. The Bradley-Terry model (Bradley and Terry [1952]), introduced by Bradley and Terry in 1952, considers paired comparisons with win-lose outcomes and provides a way to estimate the relative strength of players in a game. Other generalizations of Bradley-Terry models, such as the Rao-Kupper model (Rao and Kupper [1967]) for win-lose-draw outcomes, the Luce choice model (Luce [1959]) for choices from comparison sets, the Plackett-Luce ranking model for full ranking outcomes (Plackett [1975]), as well as group comparisons (Huang et al. [2006b, 2008]) have also been developed. Assigning ranking scores to items based on observed comparison data is a problem that arises in many applications, including information search, social opinion aggregation, electronic commerce, and online gaming platforms. Recently, ranking models have also been applied to evaluate machine learning algorithms. The key problem is to efficiently compute ranking scores that accurately reflect the strength of skills, relevancies, or preferences and to predict ranking outcomes using the estimated parameters of a statistical model of ranking outcomes.

The outline of this thesis will be as follows. This thesis comprises two chapters that study the statistical inference problems for the discrete voter model processes and Bradley-Terry models, respectively.

In Chapter 1, we consider a discrete-time voter model process on a set of nodes, each being in one of two states, either 0 or 1. In each time step, each node adopts the state of a randomly sampled neighbor according to sampling probabilities, referred to as node interaction parameters. A detailed introduction of the voter model and its related work will be given in Chapter 1.1. The key contributions from us to this chapter are summarised as follows. We study the maximum likelihood estimation of the node interaction parameters from observed node states for a given number of realizations of the

voter model process. In Chapter 1.4, we present parameter estimation error bounds by interpreting the observation data as being generated according to an extended voter process that consists of cycles, each corresponding to a realization of the voter model process until absorption to a consensus state. We also present a sampling complexity lower bound for parameter estimation within a prescribed error tolerance for the class of locally stable estimators. To obtain these results, the consensus time of a voter model process plays an important role. We present new bounds for all moments and a probability tail bound for consensus time in Chapter 1.3. Proofs and additional results are included in Chapter 1.5.

In Chapter 2, we study popular methods for inference of the Bradley-Terry model parameters, namely the gradient descent and MM algorithm, for maximum likelihood estimation and maximum a posteriori probability estimation. This class of models includes the Bradley-Terry model of paired comparisons, the Rao-Kupper model of paired comparisons allowing for tie outcomes, the Luce choice model, and the Plackett-Luce ranking model. A more detailed discussion of the Bradley-Terry model and its related work will be included in Chapter 2.1. A summary of our prior results from Vojnovic et al. [2020] will be given in Chapter 2.3. The main contributions from us to this chapter are as follows. In Chapter 2.4, we propose a simple modification of the classical gradient descent and MM algorithm with a parameter rescaling performed at each iteration step that avoids the observed slow convergence issue that we found in our previous work (Vojnovic et al. [2020]). We study the convergence rates of accelerated gradient descent and MM algorithms for Bradley-Terry models. In Chapter 2.5, we also provide some experimental results using synthetic and real-world data to demonstrate the identified slow convergence issue of the classic gradient descent and MM algorithm and show that significant efficiency gains can be obtained by our newly proposed method. Proofs and additional results are included in Chapter 2.7.

Chapter 1

Dynamics and Inference for Voter Model Processes

1.1 Introduction

The mathematical models known as *interacting particle systems* have been studied in different academic disciplines, with a canonical application to modeling opinion formation in social networks, where individuals interact pairwise and update their state in a way depending on their previous states Aldous [2013]. Models of opinion formation in social networks, e.g. DeGroot [1974], were introduced to study how consensus is reached in a network where individuals update their opinions based on their personal preferences and observed opinions of their neighbors. Threshold models of collective behavior Granovetter [1978] assume individuals update their opinions according to a threshold rule, with an individual adopting a new state only if the number of its neighbors who adopted this state exceeds a threshold value.

In this chapter, we consider the classic interacting particle system known as the voter model. The voter model was introduced in Holley and Liggett [1975] as a continuous-time Markov process, under which each individual is in one of two possible states, either 0 or 1. In this model, each individual adopts the state of a randomly sampled neighbor at random time instances according to independent Poisson processes associated with individuals or links connecting them. This voter model is an instance of an interacting particle system Liggett [1985]. A *noisy voter model* was introduced in Granovsky and Madras [1995], which is obtained from the classic voter model by adding spontaneous flipping of

states from 0 to 1, and 1 to 0, to each node. The discrete-time voter model is defined analogously to the continuous-time voter model, but with node states updated synchronously at discrete time steps. The discrete-time voter model was studied under different assumptions about which nodes update their states at discrete time points, e.g. Nakata et al. [1999], Hassin and Peleg [2001], and Cooper and Rivera [2016]. This model is particularly useful for understanding discrete changes in opinions and behaviors within a population. It finds applications in fields like social science, political science, and sociology, aiding in the study of the evolution of discrete opinions and behaviors among interacting agents. The discrete voter model captures the essence of opinion shifts while considering the granular nature of decision-making, making it a valuable tool for analyzing various societal and behavioral phenomena.

Studies of dynamics and learning in social and economic networks have been pursued from different perspectives, including dynamical systems, stochastic processes, and statistical perspectives, e.g. see Jackson [2008], Kolaczyk [2009], and Easley and Kleinberg [2010]. The key research questions include understanding the long-run behavior of the underlying random dynamical system, time to convergence to a consensus state when such a limit behavior arises, and statistical inference of model parameters from observed data, e.g. inferring node interaction rate parameters from observed node states over time.

We study dynamics and inference for the discrete-time *voter model*, defined as a Markov chain $\{X_t\}_{t\geq 0}$ with state space $\{0,1\}^n$, where X_t represents states of nodes at time t, updated such that node states X_{t+1} are independent conditional on X_t, with marginal distributions

$$X_{t+1,u} \mid X_t \sim \mathrm{Ber}(a_u^\top X_t) \text{ for } t \geq 0 \text{ and } u \in \{1,\ldots,n\} \tag{1.1.1}$$

where $a_u^\top$ is the u-th row of a stochastic matrix A, the initial state X_0 is assumed to have distribution μ, and $\mathrm{Ber}(p)$ denotes Bernoulli distribution with mean p. A matrix is said to be a *stochastic matrix* if it has real, non-negative elements and all row sums equal to 1. Intuitively, $a_{u,v}$ is the probability of node u sampling node v in a time step.

The voter model can be equivalently defined as a random linear dynamical system with $X_0 \sim \mu$ and

$$X_{t+1} = Z_{t+1}X_t, \text{ for } t \geq 0 \tag{1.1.2}$$

where $Z_1, Z_2, \ldots$ are independent and identically distributed (i.i.d.) $n \times n$ random stochastic matrices, with elements of value 0 or 1, and $\mathbb{E}[Z_1] = A$.

The voter model has $C = \{\mathbf{0}, \mathbf{1}\}$ as absorbing states and all other states are transient. Statistical inference for the voter model asks to estimate parameter A from $m \geq 1$ independent sample paths of the voter model process. For the analysis of parameter estimation, it is convenient to consider an *extended voter process* that consists of cycles, each of which corresponding to a realization of the voter model process with initial state sampled according to given initial state distribution and ending at hitting a consensus state. Such an extended voter process is defined as

$$X_{t+1} = Z_{t+1} X_t \mathbb{I}_{\{X_t \notin C\}} + \xi_{t+1} \mathbb{I}_{\{X_t \in C\}} \tag{1.1.3}$$

where ξ_t is an i.i.d. sequence of random vectors taking values in $\{0,1\}^n$ according to distribution μ.

The voter model defined by (1.1.1) and equivalently by (1.1.2) is defined such that all nodes update their states in every time step. We will also consider an *asynchronous discrete-time voter model* under which in each time step exactly one node updates its state. The asynchronous voter model dynamics is defined by

$$X_{t+1,u} \mid X_t \sim \mathrm{Ber}(a_u^\top X_t) \text{ if } u = I_t \text{ and } X_{t+1,u} = X_{t,u} \text{ if } u \in \{1, \ldots, n\} \setminus \{I_t\} \tag{1.1.4}$$

where I_t are i.i.d. random variables according to uniform distribution on $\{1, \ldots, n\}$. The asynchronous discrete-time voter model also obeys the random linear dynamical system recursive equation (1.1.2) but with $Z_1, Z_2, \ldots$ being i.i.d. $n \times n$ random stochastic matrices with elements of value 0 or 1 such that $\mathbb{E}[Z_t \mid I_t = u] = e_u a_u^\top + \sum_{v \neq u} e_v e_v^\top$ where e_w denotes the n-dimensional standard basis vector, with the w-th element equal to 1 and other elements equal to 0.

The discrete-time *ϵ-noisy voter model* is defined by $X_0 \sim \mu$, and

$$X_{t+1,u} \mid X_t \sim \mathrm{Ber}(f(a_u^\top X_t)) \text{ for } t \geq 0 \text{ and } u \in \{1, \ldots, n\} \tag{1.1.5}$$

where f is some given function $f : [0,1] \to [\epsilon, 1-\epsilon]$, and $\epsilon \in [0, 1/2]$. Under certain conditions on f and $0 < \epsilon < 1/2$, the ϵ-noisy voter model is an ergodic stochastic process. This is important for statistical inference of the model parameters as they can be inferred from a single, sufficiently long random realization of the stochastic process. This is in contrast to the voter model which requires

several realizations of the voter model process for inference of the model parameters.

A special case of an ϵ-noisy voter model is *the linear ϵ-noisy voter model* defined by taking $f(x) = \epsilon + (1-2\epsilon)x$. Note that the linear ϵ-noisy voter model corresponds to the voter model when $\epsilon = 0$. For the linear ϵ-noisy voter model, we have

$$X_{t+1} = D(Q_{t+1})Z_{t+1}X_t + R_{t+1} \tag{1.1.6}$$

where (Q_t, R_t) is an i.i.d. sequence of n dimensional vectors with independent elements with distribution $\mathbb{P}[(Q_{t,u}, R_{t,u}) = (q,r)] = p(q,r)$ with $p(0,0) = p(0,1) = \epsilon$ and $p(1,0) = 1-2\epsilon$, for all $u \in \{1,\ldots,n\}$, and $D(x)$ denoting diagonal matrix with diagonal elements x. The random linear dynamical system (1.1.6) can be seen as a randomly perturbed version of the random linear dynamical system (1.1.2). The role of this perturbation is significant, making an absorbing Markov chain to an ergodic Markov chain.

Statistical inference for the voter model process is a challenging task because the number of informative node interactions for parameter estimation vanish as node states converge to a consensus state. For a node interaction to be informative for the estimation, it is necessary that the node has neighbors with mixed states—having some neighbors in state 0 and some in state 1. For example, consider asynchronous discrete-time voter model, where at each time step a single, random node observes the state of a randomly picked neighbor and updates its state, with node interactions restricted to a path connecting n nodes. Assume that initially k nodes on one end of the path are in state 1 and other nodes are in state 0. Then, we can show that the expected number of nodes participating in at least one informative interaction until absorption to a consensus state is $k(\log(n/k) + \Theta(1))$ when $k = o(n)$, we show this in Section 1.5.17. For the given voter model instance, typically, informative interactions will be observed only for a small fraction of nodes in each realization of the voter model process. In general, for the voter model inference, both the matrix of pairwise interaction rates A and the initial state distribution μ play an important role.

We present new results on statistical inference for the voter model. This is achieved by using a framework that allows to study inference for absorbing stochastic processes, by learning from several realizations of the underlying stochastic process. This is different from existing work on statistical inference for stationary autoregressive stochastic processes, akin to the aforementioned ϵ-noisy voter model. In order to study statistical inference for absorbing stochastic processes, we need certain

properties of the hitting time of an absorbing state. Specifically, for the voter model, we need bounds on the expected value and a probability tail bound of consensus time. We present new results for the latter two properties of the consensus time, which may be of general interest. Before summarizing our contributions in some more detail, we review related work.

1.1.1 Related work

Prior work on voter models is mostly concerned with dynamics of voter processes on graphs, studying properties such as hitting probabilities and time to reach an absorbing state. Several seminal works studied voter model in continuous time, where interactions between vertices occur at events triggered by independent Poisson processes associated with vertices or edges, e.g. Cox [1989b], Liggett [1985], Oliveira [2012]. The discrete-time voter model was first studied in Nakata et al. [1999] and Hassin and Peleg [2001] under assumption that A is a stochastic matrix such that $a_{u,v} > 0$ if and only if $a_{v,u} > 0$ and the support of A corresponds to the adjacency matrix of a nonbipartite graph. Hassin and Peleg [2001] found a precise characterization of the hitting probabilities of absorbing states, showing that $\lim_{t\to\infty} \mathbb{P}[X_t = \mathbf{1}] = 1 - \lim_{t\to\infty} \mathbb{P}[X_t = \mathbf{0}] = \pi^\top x$, for any initial state x, where π is the stationary distribution of A, i.e. a unique distribution π satisfying $\pi^\top = \pi^\top A$.

The consensus time of voter model process was studied in previous work under various assumptions. In an early work, Cox [1989b] studied coalescing random walks and voter model consensus time on a torus. Most of works studied voter model process with node interactions defined by a graph $G = (V, E)$ where V is the set of vertices and E is the set of edges. In Hassin and Peleg [2001], using duality between voter model process and coalescing random walks, it was shown that the expected consensus time is $O(m(G)\log(n))$, where $m(G)$ is the worst-case expected meeting time of two random walks on graph G. It's worth noting that the meeting time naturally serves as a clear lower bound for coalescing time, hence also a lower bound on the consensus time. Kanade et al. [2019] showed that $m(G) = O((nd_{\max}/\Phi(G))\log(d_{\max}))$, for a lazy random walk, where $d_{\max}$ is the maximum node degree and $\Phi(G)$ is graph conductance. This lazy random walk, in each time step remains at the current vertex with probability $1/2$ and otherwise moves to a randomly chosen neighbor. Combined with the result in Hassin and Peleg [2001], this implies the expected consensus time bound $\tilde{O}((nd_{\max})/\Phi(G))$. Berenbrink et al. [2016] studied a voter model process where in each time step, every vertex copies the state of a randomly selected neighbor with probability $1/2$, and, otherwise, does not change its state (corresponding to the previously defined lazy random walk). In

this setting, they showed that the expected consensus time is $O((d(V)/d_{\min})/\Phi(G))$, where $d_{\min}$ is the minimum node degree and $d(V)$ is the sum of node degrees. Our bound on the expected consensus time is competitive to the best previously known bound up to at most a logarithmic factor in n. Aldous and Fill [2002], Cooper et al. [2010] showed that if the states of vertices are initially distinct, the voter model process takes $\Theta(n)$ expected steps to reach consensus on many classes of expander graphs with n vertices. Cooper et al. [2013] established bounds on the expected consensus time for the voter model process on general connected graphs that depend on an eigenvalue gap of the transition matrix of random walk on the graph and the variance of the degree sequence. Cooper and Rivera [2016] showed that the expected consensus time is $O(1/\Psi_A)$ where Ψ_A is a property of A (we discuss in Section 1.3.2). Oliveira and Peres [2019] established various results on hitting times of lazy random walks on graphs. Unlike to the aforementioned previous work, which found bounds on the expected consensus bound or bounds that hold with a constant probability, our results allow us to derive high probability bounds.

The problem of inferring node interaction parameters from observed node states was studied in an early work by Netrapalli and Sanghavi [2012] for classic epidemic models, and more recently by Pouget-Abadie and Horel [2015a] for independent cascade model and some stationary voter model processes, as well as by Gomez-Rodriguez et al. [2016] for some continuous-time network diffusion processes. None of these works considered statistical inference for an absorbing voter model process. A recent line of work studied statistical estimation for sparse autoregressive processes, including vector autoregressive processes Basu and Michailidis [2015], Hall et al. [2019, 2016], Zhu et al. [2017], Zhu and Pan [2020], sparse Bernoulli autoregressive processes Pandit et al. [2019], Mark et al. [2019], Katselis et al. [2019], and network Poisson processes Mark et al. [2019]. These works established convergence rates for the parameter estimation problem. All these works are concerned with stationary autoregressive processes and thus do not apply to inference of absorbing stochastic processes, such as the voter model we study. Our work provides a framework to study statistical inference for absorbing stochastic processes, which may be of independent interest for autoregressive stochastic processes. Another related work is on identification of discrete-time linear dynamical systems with random noise, e.g. recent works by Simchowitz et al. [2018] and Jedra and Proutiere [2019]. We present a new lower bound on the sampling complexity for estimation of voter model parameters by studying the random linear dynamical system that governs the evolution of the voter model process.

1.1.2 Summary of our contributions

We show an upper bound on the expected consensus time

$$\mathbb{E}^0[\tau] \leq \frac{1}{\Phi_A} \log\left(\frac{1}{2\pi^*}\right)$$

where $\pi^* = \min\{\pi_v : v = 1, \ldots, n\}$, π is the stationary distribution of A, and Φ_A is a parameter of A. The upper bound is tight in the sense that there exist voter model instances that have the expected consensus time matching the upper bound up to a poly-logarithmic factor in n. The upper bound is obtained by a Lyapunov function analysis and follows from the exponential moment bound $\mathbb{E}[e^{\theta\tau}] \leq 1/(2\pi^*)$ that holds for any θ such that $(1 - \Phi_A)e^{\theta} \leq 1$. This exponential moment bound allows us to bound the consensus time with high probability. Specifically, for any $\delta \in (0, 1]$, $\tau \leq (\log(1/(2\pi^*)) + \log(1/\delta))/\Phi_A$ with probability at least $1 - \delta$. In particular, this implies a high probability bound, $\tau \leq (\log(1/(2\pi^*)) + \log(n))/\Phi_A$, which holds with probability at least $1 - 1/n$. Moreover, by using the aforementioned exponential moment bound, we can bound any moment of consensus time. These results are instrumental for the voter model inference problem, and may also be of independent interest.

We developed a methodology for establishing statistical estimation error bounds for absorbing stochastic processes. This is based on using the framework for analysis of M-estimators with decomposable regularizers from high-dimensional statistics along with bounds on the expected length of cycles and probability tail bounds for the length of cycles. To obtain these results, we leverage our bounds on the expected consensus time and probability tail bounds for the consensus time, and use probability tail bounds for some super-martingale sequences.

We show that the parameter estimation error of a voter model with parameter A due to statistical estimation errors, measured by squared Frobenious norm, is

$$\tilde{O}\left(\frac{s}{\alpha^2(1/\Phi_A)\lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^2}\frac{1}{m}\right)$$

with high probability, for sufficiently large number m. Here s is an upper bound on the support of the voter model parameter A, $\lambda_{\min}(\mathbb{E}[X_0 X_0^\top])$ is the smallest eigenvalue of the correlation matrix of the extended voter process with respect to stationary distribution, and $\alpha > 0$ is a lower bound for any non-zero element of A.

We also present a lower bound on the sampling complexity for statistical inference of the voter model parameters, using the framework of locally stable estimators. Roughly speaking, for the voter model with parameter A with every element in its support of value at least $\alpha > 0$, to have the Frobenious norm of the parameter estimation error bounded by ϵ, with probability at least $1 - \delta$, the number of voter model process realizations, m, must satisfy, for every sufficiently small $\epsilon > 0$,

$$m \geq \frac{\alpha}{16} \frac{1}{\epsilon^2 \mathbb{E}^0[\tau] \lambda_{\min}(\mathbb{E}\left[X_0 X_0^\top\right])} \log\left(\frac{1}{2.4\delta}\right).$$

1.1.3 Organization of the chapter

In Section 1.2 we provide additional definitions for model formulation and some mathematical background for the analysis in the model. Section 1.3 contains our main results on the consensus time of the voter model process. Specifically, this includes the exponential moment bound in Theorem 1.3.1 from which we derive a bound on the expected consensus time in Corollary 1.3.2, a bound on any moment of consensus time in Corollary 1.3.3, and a probability tail bound for a sum of independent consensus times in Theorem 1.3.2. Section 1.4 contains our results on statistical estimation, with an upper bound provided in Theorem 1.4.2 and a lower bound in Theorem 1.4.3. Section 1.5 contains missing proofs and some further results.

1.2 Preliminaries

In this section we provide additional details for the model formulation and then provide some background definitions and results that we use in the rest of the chapter.

1.2.1 Model formulation

The voter model is defined as a Markov chain $\{X_t\}_{t\geq 0}$ on the state space $\mathcal{X} = \{0,1\}^n$ with initial state X_0 with distribution μ and the state transitions defined by

$$X_{t+1} = Z_{t+1} X_t, \text{ for } t \geq 0 \tag{1.2.1}$$

where $Z_1, Z_2, \ldots$ is an i.i.d. random sequence of stochastic matrices taking values in $\{0,1\}^{n\times n}$ with independent rows. We can interpret $X_{t,u}$ as the state of vertex $u \in V := \{1, \ldots, n\}$ at time t, which takes value 0 or 1.

The system (1.2.1) is a time-variant linear dynamical system with random i.i.d. linear transformations as defined above.

We use the notation

$$A = \mathbb{E}[Z_1].$$

We assume that A is an aperiodic, irreducible transition matrix. Under these assumptions, A has a stationary distribution π which is unique, given as the solution of global balance equations $\pi^\top = \pi^\top A$. We denote with $a_u^\top$ the u-th row of matrix A. Note that a_u can be interpreted as a probability distribution according to which vertex u initiates pairwise interactions. Any two vertices u and v are said to be *neighbors* if $a_{u,v} > 0$, i.e. if the two vertices interact with a positive probability.

For the voter model, there are two absorbing states $C = \{\mathbf{0}, \mathbf{1}\}$, and all other states are transient. Let $C_0 = \{\mathbf{0}\}$ and $C_1 = \{\mathbf{1}\}$, hence $C = C_0 \cup C_1$. We refer to either of the two absorbing states as a consensus state. Let τ denote the hitting time of a consensus state, we refer to as the *consensus time*, which is defined by

$$\tau = \min\{t \geq 1 : X_t \in C\}.$$

In our analysis, we consider the extended voter process $\{X_t\}_{t\in\mathbb{Z}}$ defined by cycles of individual voter model process realizations. Let $\{T_i\}_{i\in\mathbb{Z}}$ be a point process defined as follows. Let T_i be the time at which the i-th voter model process is in its initial state, and let $S_i = T_{i+1} - T_i$. We assume that X_t for $T_i \leq t < T_{i+1}$, correspond to the states of the i-th voter model process until reaching a consensus state, excluding its final consensus state. Note that, indeed, $S_i = \tau_i$, where τ_i is the consensus time of the i-th voter model process. In some parts of our analysis, we will also consider the alternative definition of the extended voter process which includes the final states of individual voter model processes. In this case, $S_i = \tau_i + 1$. The stationary distributions of the two extended voter processes are different.

We denote with $\mathbb{P}^0[A]$ the probability of an event A conditional on time 0 being a point, i.e. $\{T_0 = 0\}$. We denote with $\mathbb{P}[A]$ the probability of event A under stationary distribution, where 0 is an arbitrary time. In the framework of stationary point processes, the two distributions are referred to as the Palm distribution and stationary distribution, respectively. By the Palm inversion formula, for any measurable function $f : \mathcal{X} \mapsto \mathbb{R}$,

$$\mathbb{E}[f(X_0)] = \frac{\mathbb{E}^0\left[\sum_{t=0}^{S_1-1} f(X_t)\right]}{\mathbb{E}^0[S_1]}. \tag{1.2.2}$$

We will also use the following notation in our analysis of parameter estimation. Let $\{X_t^{(i)}\}_{t\geq 0}$, $i = 1, \ldots, m$ be independent voter model processes, each with initial state distribution μ. Let τ_i denote the consensus time of voter model process i. With a slight abuse of notation, we will sometimes write X_t in lieu of $X_t^{(1)}$ and τ in lieu of τ_1.

1.2.2 Markov chains background

In this section we present some definitions and results from Markov chain theory that we use in our analysis.

Let $\{X_t\}_{t\geq 0}$ be a time homogeneous Markov chain on a state space $(\mathcal{X}, \mathcal{E})$. Let $P(x, A)$, $x \in \mathcal{X}$, $A \in \mathcal{E}$ denote the transition probability and let P denote the corresponding operators on measurable functions mapping $\mathcal{X}$ to $\mathbb{R}$. Let $P_t(x, \cdot)$ denote the transition probabilities at time t. We define the following conditions:

(A1) *Minorization condition*. There exist $S \in \mathcal{E}$, $\epsilon > 0$, and a probability measure ν on $(\mathcal{X}, \mathcal{E})$ such that

$$P(x, A) \geq \epsilon\nu(A)$$

for all $x \in S$ and $A \in \mathcal{E}$.

(A2) *Drift condition*. There exist a measurable function $V : \mathcal{X} \to [1, \infty)$ and constants $\lambda < 1$ and $K < \infty$ satisfying

$$PV(x) := \mathbb{E}[V(X_1) \mid X_0 = x] \leq \begin{cases} \lambda V(x) & \text{if } x \notin S \\ K & \text{if } x \in S. \end{cases}$$

(A3) *Strong aperiodicity condition*. There exists $\tilde{\epsilon} > 0$ such that $\epsilon\nu(S) \geq \tilde{\epsilon}$.

We say that a measurable function $V : \mathcal{X} \to [1, \infty)$ is a *drift function* for P with respect to S, with constants $\lambda < 1$ and $K < \infty$, if it satisfies (A2).

For any given set $S \subset \mathcal{E}$, let us define the hitting time

$$\tau_S = \min\{t > 0 : X_t \in S\}.$$

We say that the set S is an *atom* if $P(x, \cdot) = P(y, \cdot)$ for all $x, y \in S$. In this case, we may assume

$\epsilon = 1$ and $P(x, \cdot) = \nu(\cdot)$ for all $x \in S$.

By Theorem 1.1 in Baxendale [2005], under (A1)-(A3), $\{X_t\}_{t\geq 0}$ has a unique stationary distribution π and $\mathbb{E}_{X\sim\pi}[V(X)] < \infty$. Moreover, there exists $\rho < 1$ depending only on ϵ, $\tilde{\epsilon}$, λ and K such that whenever $\rho < \gamma < 1$, there exists $M < \infty$ depending only on γ, ϵ, $\tilde{\epsilon}$, and K such that

$$\sup_{|g|\leq V} |\mathbb{E}_x[g(X_t)] - \mathbb{E}_\pi[g(X_0)]| \leq MV(x)\gamma^t$$

for all $x \in \mathcal{X}$ and $t \geq 0$, where the supremum is over all measurable functions $g : \mathcal{X} \to \mathbb{R}$ satisfying $|g(x)| \leq V(x)$ for all $x \in \mathcal{X}$. If g is restricted to functions satisfying $|g(x)| \leq 1$ for all $x \in \mathcal{X}$, then we have the standard geometric ergodicity condition $||P_t(x, \cdot) - \pi||_{TV} \leq MV(x)\gamma^t$ where $||\cdot||_{TV}$ denotes the total variation distance. The Markov chain is said to be (M, ρ)-*geometrically ergodic* if $||P_t(x, \cdot) - \pi||_{TV} \leq M\rho^t$ for some $M < \infty$ and $\rho < 1$.

The following is a key lemma (e.g. Lemma 2.2 and Theorem 3.1 Lund and Tweedie [1996], Proposition 4.1 Baxendale [2005]) that we will use in our analysis of consensus time.

Lemma 1.2.1. *Let $\{X_t\}_{t\geq 0}$ be a Markov chain on $(\mathcal{X}, \mathcal{E})$ with transition kernel P, and let $C \in \mathcal{E}$. Suppose that $V : \mathcal{X} \to [1, \infty)$ is a measurable function that satisfies $PV(x) \leq \lambda V(x)$ for all $x \notin C$, for a fixed $\lambda < 1$. Then, for all $x \in \mathcal{X}$,*

$$\mathbb{E}_x[\lambda^{-\tau_C}] \leq V(x).$$

The lemma can be established by some Lyapunov drift arguments.

1.2.3 Miscellaneous definitions

We use different definitions of norms. For every $x \in \mathbb{R}^n$, $||x||_p$ denotes the L_p norm, $||x||_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$. For every matrix $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times m}$, $||X||_{p,q}$ is defined as

$$||X||_{p,q} = (||x_1||_q^p + \ldots + ||x_n||_q^p)^{1/p}.$$

In particular, $||X||_{1,1}$ is the sum of absolute values of elements of X. $||X||_0$ denotes the number of non-zero elements in X, i.e. the *support* of X. The *Frobenius norm* $||X||_F$ is defined as

$$||X||_F = ||X||_{2,2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}^2}.$$

1.3 Consensus time

In this section we show our results on consensus time of the voter model process.

1.3.1 Consensus time bounds

For any vector $a \in \mathbb{R}^n$, let us define

$$V_a(x) = a^\top x(1 - a^\top x).$$

For the voter model with parameter A, $V_{a_u}(X_t)$ is the variance of $X_{t+1,u}$ conditional on X_t. Intuitively, we may interpret $V_{a_u}(X_t)$ as a measure of diversity of states of neighbors of vertex u. Note that $V_{a_u}(X_t) = 0$ if and only if all neighbors of vertex u are in the same state (either 0 or 1).

Lemma 1.3.1. *For every $x \in \{0,1\}^n$ and $t \in \mathbb{Z}$, function V_π satisfies the following expected drift equation:*

$$\mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] = -\sum_{u=1}^{n} \pi_u^2 V_{a_u}(x) + \mathbb{E}^0[V_\pi(X_0)]\mathbb{I}_{\{x \in C\}}.$$

We can interpret the term $\sum_{u=1}^n \pi_u^2 V_{a_u}(x)$ as a weighted sum of variances of Bernoulli distributions with parameters $a_u^\top x$, associated with vertex neighborhood sets. The weights are equal to the squares of the elements of the stationary distribution π. By taking expectation on both sides in the equation in Lemma 1.3.1 with respect to the stationary distribution, under assumption that $\mu(C) = 0$, and applying Palm inversion formula (1.2.2), we get

$$\mathbb{E}[\mathbb{I}_{\{X_0 \in C\}}] = \frac{\mathbb{E}^0[\sum_{t=0}^{\tau} \mathbb{I}_{\{X_t \in C\}}]}{\mathbb{E}^0[\tau] + 1} = \frac{1}{\mathbb{E}^0[\tau] + 1}.$$

Hence, we have

$$(\mathbb{E}^0[\tau]+1)\sum_{u=1}^{n}\pi_u^2\mathbb{E}[V_{a_u}(X_0)]=\mathbb{E}^0[V_\pi(X_0)]. \tag{1.3.1}$$

From (1.3.1), we can observe that the expected consensus time is fully determined by the expected variance of vertex states with respect to the stationary distribution measured by $\sum_{u=1}^{n}\pi_u^2\mathbb{E}[V_{a_u}(X_0)]$ and the variance of the initial state measured by $\mathbb{E}^0[V_\pi(X_0)]$. Intuitively, the smaller the value of $\sum_{u=1}^{n}\pi_u^2\mathbb{E}[V_{a_u}(X_0)]$, the larger the expected consensus time.

The following is a key property of A in our analysis of consensus time

$$\Phi_A=\min\left\{\frac{\sum_{u=1}^{n}\pi_u^2V_{a_u}(x)}{\pi^\top x(1-\pi^\top x)}:x\in\{0,1\}^n,x\notin C\right\}. \tag{1.3.2}$$

Note that $0<\Phi_A\leq 1$. The inequality $\Phi_A>0$ can be shown by contradiction as follows. Suppose $\Phi_A=0$, which is equivalent to $a_u^\top x(1-a_u^\top x)=0$ for all $u\in V$. We can then partition V into two non-empty sets S and $V\setminus S$ such that each u has support of a_u fully contained in either S or $V\setminus S$. This implies that A has a block structure, which contradicts the assumption that A is irreducible and aperiodic. The inequality $\Phi_A\leq 1$ follows from

$$\begin{aligned}\sum_{u=1}^{n}\pi_u^2a_u^\top x(1-a_u^\top x) &\leq \sum_{u=1}^{n}\pi_u a_u^\top x(1-a_u^\top x)\\ &\leq \left(\sum_{u=1}^{n}\pi_u a_u^\top x\right)\left(1-\left(\sum_{u=1}^{n}\pi_u a_u^\top x\right)^2\right)\\ &= \pi^\top x(1-\pi^\top x)\end{aligned}$$

where the first inequality is by non-negativity of the summation terms and $\pi_u\in[0,1]$ for all $u\in V$, the second inequality is by concavity of $x\mapsto x(1-x)$, and the equation is by the global balance equations $\pi^\top=\pi^\top A$.

By Lemma 1.3.1 and definition of Φ_A, we have the following corollary.

Corollary 1.3.1. *For all $x\in\{0,1\}^n$ and $t\in\mathbb{Z}$,*

$$\mathbb{E}[V_\pi(X_{t+1})-V_\pi(X_t)\mid X_t=x]\leq-\Phi_AV_\pi(x)+\mathbb{E}^0[V_\pi(X_0)]\mathbb{I}_{\{x\in C\}}.$$

We next present a bound on the exponential moment of consensus time.

Theorem 1.3.1. *For any $x \in \{0,1\}^n$ such that $x \notin C$, and any $\theta \in \mathbb{R}$ such that $(1-\Phi_A)e^\theta \leq 1$,*

$$\mathbb{E}_x^0[e^{\theta\tau}] \leq \frac{V_\pi(x)}{\min_{z\in\{0,1\}^n\setminus C} V_\pi(z)}.$$

Proof. The theorem follows from the general result for Markov chains satisfying the Lyapunov drift condition stated in Lemma 1.2.1. Recall that $C = \{\mathbf{0}, \mathbf{1}\}$. Let $V(x) := V_\pi(x)/\min_{z\in\{0,1\}^n\setminus C} V_\pi(z)$. Using Corollary 1.3.1, $V(x)$ is a drift function with respect to C, with constants $\lambda = 1-\Phi_A$ and $K = \mathbb{E}^0[V(X_0)]$. By Lemma 1.2.1, we have $\mathbb{E}_x[(1-\Phi_A)^{-\tau}] \leq V(x)$. Combining with the condition $(1-\Phi_A)e^\theta \leq 1$, the claim of the theorem follows. □

We have the following upper bound for the expected consensus time.

Corollary 1.3.2. *For every $x \in \{0,1\}^n$ such that $x \notin C$,*

$$\mathbb{E}_x^0[\tau] \leq \frac{1}{\Phi_A}\log\left(\frac{1}{2\pi^*}\right)$$

where $\pi^* = \min\{\pi_v : v = 1,\ldots,n\}$.

Proof. By Theorem 1.3.1, taking θ such that $(1-\Phi_A)e^\theta = 1$, and Jensen's inequality, we have

$$\mathbb{E}_x^0[\tau] \leq \frac{1}{\log\left(\frac{1}{1-\Phi_A}\right)}\log\left(\frac{V_\pi(x)}{\min_{z\in\{0,1\}^n\setminus C} V_\pi(z)}\right).$$

The corollary follows from the last inequality and combining with the following facts (a) $V_\pi(x) \leq 1/4$ for all $x \in \{0,1\}^n$, (b) $V_\pi(z) \geq \pi^*(1-\pi^*) \geq \pi^*/2$, for all $z \in \{0,1\}^n \setminus C$, and (c) $1/\log(1/(1-\Phi_A)) \leq 1/\Phi_A$. □

From (1.3.1), we obtain

$$\mathbb{E}^0[\tau] \geq \frac{4\mathbb{E}^0[V_\pi(X_0)]}{||\pi||_2^2} - 1.$$

By Corollary 1.3.2, $\mathbb{E}^0[\tau] \leq \log(1/(2\pi^*))/\Phi_A$. Hence, if $\Phi_A = \Omega(||\pi||_2^2)$ and $\mathbb{E}^0[V_\pi(X_0)] = \Omega(1)$, the upper bound is tight within a factor logarithmic in $1/\pi^*$. For instance, for the complete graph case, i.e. when $a_{u,u} = 0$ and $a_{u,v} = 1/(n-1)$ for all $u \neq v$, we have $\Phi_A = (1/n)(1+o(1))$ (we show

this in Section 1.3.2). In this case, we have the upper bound $\mathbb{E}^0[\tau] = O(n\log(n))$, which is within a factor logarithmic in n of the lower bound $\mathbb{E}^0[\tau] = \Omega(n)$.

In fact, from Theorem 1.3.1, we have the following bound for any moment of the consensus time.

Corollary 1.3.3. *For every $x \in \{0,1\}^n$ such that $x \notin C$ and $k \geq 0$,*

$$\mathbb{E}_x^0[\tau^k] \leq \frac{1}{2}\left(\frac{k}{e}\right)^k \frac{1}{\Phi_A^k}\frac{1}{\pi^*}.$$

Proof. The proof follows readily from Theorem 1.3.1 and the the elementary fact that for any non-negative random variable X, for any $k \geq 0$ and $\theta > 0$, $\mathbb{E}[X^k] \leq (k/(e\theta))^k \mathbb{E}[e^{\theta X}]$. □

The bound for the first moment of consensus time in Corollary 1.3.2 is better than that in Corollary 1.3.3 in having a logarithmic dependence on $1/\pi^*$ instead of linear dependence on this parameter.

By using the bound on the exponential moment of consensus time in Theorem 1.3.1, we can obtain a bound on the tail probability of the consensus time. This allows us to derive bounds on the consensus time that hold with high probability. We will next state a more general result that applies to the sum of consensus times of $m \geq 1$ independent voter model processes. We will use this more general result for the parameter estimation in Section 1.4.

Theorem 1.3.2. *For $m \geq 1$ independent voter model processes with parameter A and independent initial states according to distribution μ, for any $a \geq 0$,*

$$\mathbb{P}^0\left[\sum_{i=1}^m \tau_i \geq ma\right] \leq \left(\left(\frac{\mathbb{E}^0[V_\pi(X_0)]}{\min_{z\in\{0,1\}^n\setminus C} V_\pi(z)}\right)(1-\Phi_A)^a\right)^m. \tag{1.3.3}$$

Proof. By Chernoff's bound, for any $\theta \geq 0$,

$$\begin{aligned}\mathbb{P}^0\left[\sum_{i=1}^m \tau_i \geq ma\right] &\leq e^{-ma\theta}\mathbb{E}^0\left[e^{\theta\sum_{i=1}^m \tau_i}\right] \\ &= e^{-ma\theta}\mathbb{E}^0[e^{\theta\tau_1}]^m.\end{aligned}$$

Let $\theta^* = -\log(1-\Phi_A)$. By Theorem 1.3.1, we have

$$\mathbb{E}^0[e^{\theta^*\tau_1}] \leq \frac{\mathbb{E}^0[V_\pi(X_0)]}{\min_{z\in\{0,1\}^n\setminus C} V_\pi(z)}.$$

Hence, (1.3.3) follows. □

From Theorem 1.3.2, for any $\delta \in (0,1]$, with probability at least $1-\delta$,

$$\frac{1}{m}\sum_{i=1}^{m}\tau_i \leq \left(\log\left(\frac{\mathbb{E}^0[V_\pi(X_0)]}{\min_{z\in\{0,1\}^n\setminus C} V_\pi(z)}\right) + \frac{1}{m}\log\left(\frac{1}{\delta}\right)\right)\frac{1}{\log\left(\frac{1}{1-\Phi_A}\right)}.$$

From the last statement, we have the following corollary.

Corollary 1.3.4. *For $m \geq 1$ independent voter model processes with parameter A and independent initial states with distribution μ, for any $\delta \in (0,1]$, with probability at least $1-\delta$,*

$$\frac{1}{m}\sum_{i=1}^{m}\tau_i \leq \frac{1}{\Phi_A}\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{1}{\delta}\right)\right).$$

The corollary implies us a high probability bound for consensus time, $O((1/\Phi_A)(\log(1/\pi^*)+\log(n)))$, which holds with probability at least $1-1/n^c$, for any constant $c>0$.

Asynchronous discrete-time voter model Similar results hold for the asynchronous discrete-time voter model. An analogous lemma to Lemma 1.2.1 holds which is given as follows.

Lemma 1.3.2. *For every $x \in \{0,1\}^n$ and $t \in \mathbb{Z}$, function V_π satisfies the following expected drift equation:*

$$\begin{aligned} & \mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] \\ = \; & -\frac{1}{n}\sum_{u=1}^{n}\pi_u^2(x_u(1-a_u^\top x) + (1-x_u)a_u^\top x) + \mathbb{E}^0[V_\pi(X_0)]\mathbb{I}_{\{x\in C\}}. \end{aligned}$$

The statements in Corollary 1.3.2 and Theorem 1.3.2 hold true for asynchronous discrete-time voter

model by replacing Φ_A with Φ'_A where

$$\Phi'_A = \frac{1}{n} \min \left\{ \frac{\sum_{u=1}^{n} \pi_u^2 V'_u(x)}{\pi^\top x(1-\pi^\top x)} : x \in \{0,1\}^n, x \neq C \right\} \tag{1.3.4}$$

with $V'_u(x) := x_u(1 - a_u^\top x) + (1 - x_u)a_u^\top x$.

It is readily observed that $n\Phi'_A \geq \Phi_A$ and it follows that the statements in Corollary 1.3.2 and Theorem 1.3.2 remain to hold true for asynchronous discrete-time voter model by replacing Φ_A with Φ_A/n. The factor $1/n$ occurs because under asynchronous discrete-time voter model, at each time step, exactly one node updates its state, which is in contrast to the voter model under which all nodes update their states.

1.3.2 Discussion and comparison with previously-known consensus time bounds

We discuss the value of parameter Φ_A and expected consensus time for some node interaction matrices A and compare with the best previously-known consensus time bounds.

We will discuss node interaction probabilities that can be defined by a graph, as common in the literature on voter model processes and random walks on graphs. Let $G = (V, E)$ be a connected graph where V is the set of $|V| = n$ vertices and E is the set of edges. Let d_v denote the *degree* of vertex v, defined as the number edges incident to vertex v. For any set $S \subseteq V$, let $d(S) = \sum_{v \in S} d_v$, and let $d_v(S)$ denote the number edges incident to v and S. Let $d_{\min}$ denote the minimum degree of a vertex in G. Graph G may contain self-loops, i.e. an edge connecting a vertex with itself.

We represent any given vector $x \in \{0,1\}^n$ by the set $S = \{v \in V : x_v = 1\}$, and we will use the notation $S^c := V \setminus S$. Note that a vector $x \in \{0,1\}^n$ defines a graph partition, i.e. partition of the set of vertices into two components, S and S^c.

We will compare Φ_A and bonds on the expected consensus time with some functions of graph conductance. Conductance $\Phi(G)$ of graph G is defined as

$$\Phi(G) = \min_{S \subset V : 0 < |S| < n} \frac{|E(S, S^c)|}{\min\{d(S), d(S^c)\}} \tag{1.3.5}$$

where $E(S, S^c)$ is the set of edges connecting S and S^c. Let L be the normalized Laplacian matrix of graph G, defined as $L = I - D^{-1/2}AD^{-1/2}$ where D is the diagonal matrix with diagonal elements corresponding to vertex degrees. Let λ_2 be the second smallest eigenvalue of L. By Cheeger's

inequality, for any connected graph G,

$$\lambda_2/2 \leq \Phi(G) \leq \sqrt{2\lambda_2}. \tag{1.3.6}$$

We first consider node interactions such that in each time step, every vertex copies the state of a randomly chosen neighbor with probability $1/2$, where graph G has no self-loops. For example, this case was studied in Berenbrink et al. [2016] and Kanade et al. [2019], and is commonly referred to as lazy random walk. The node interaction matrix A has elements

$$a_{u,v} = \frac{1}{2}\mathbb{I}_{\{u=v\}} + \frac{1}{2}\frac{1}{d_u}\mathbb{I}_{\{(u,v)\in E\}}, \text{ for } u, v \in V. \tag{1.3.7}$$

It can be readily checked that $\pi_v = d_v/d(V)$ for $v \in V$, and we also have

$$\Phi_A = \frac{1}{4}\min_{S\subset V: 0<|S|<n} \frac{\sum_{u\in V}(d_u\mathbb{I}_{\{u\in S\}} + d_u(S))(d_u\mathbb{I}_{\{u\in S^c\}} + d_u(S^c))}{d(S)d(S^c)}.$$

Lemma 1.3.3. *Assume that node interaction matrix A is according to (1.3.7). Then, we have*

$$\frac{1}{\Phi_A} \leq 2\frac{d(V)}{d_{\min}}\frac{1}{\Phi(G)}.$$

Together with Corollary 1.3.2, Lemma 1.3.3 implies $\mathbb{E}^0[\tau] = O(((d(V)/d_{\min})/\Phi(G))\log(n))$, which is within a logarithmic factor in n to the expected consensus time bound in Berenbrink et al. [2016].

For comparing with the expected consensus time bound in Cooper and Rivera [2016], $\mathbb{E}^0[\tau] \leq 64/\Psi_A$, we consider Ψ_A which is defined as $\Psi_A = \pi^*\tilde{\Psi}_A$ with

$$\tilde{\Psi}_A = \min_{x\in\{0,1\}^n\setminus C} \frac{\mathbb{E}\left[\left|\sum_{u=1}^n \pi_u\left(x_u - \sum_{v=1}^n Z_{u,v}x_v\right)\right|\right]}{\min\{\pi^\top x, 1-\pi^\top x\}}.$$

Lemma 1.3.4. *For node interaction matrix A according to (1.3.7), we have*

$$\tilde{\Psi}_A \leq \Phi(G).$$

From Lemmas 1.3.3 and 1.3.4, we have

Lemma 1.3.5. *For node interaction matrix A according to (1.3.7), we have*

$$\frac{1}{\Phi_A} \leq 2\frac{1}{\Psi_A}.$$

The last lemma, together with Corollary 1.3.2, implies that our bound on the expected consensus time is at most a logarithmic factor in n to the expected consensus time bound in Cooper and Rivera [2016].

We also considered another type of node interactions, where in each time step, each node copies the state of a node from its neighborhood set which includes the node itself, i.e. we consider A defined as

$$a_{u,v} = \frac{1}{d_u}\mathbb{I}_{\{(u,v)\in E\}}, \text{ for } u, v \in V. \tag{1.3.8}$$

It is readily checked that $\pi_v = d_v/d(V)$, for $v \in V$, and

$$\Phi_A = \min_{S\subset V: 0<|S|<n} \frac{|E_2(S, S^c)|}{d(S)d(S^c)} \tag{1.3.9}$$

where $E_2(S, S^c)$ is the set of paths consisting of two edges connecting S and S^c. Equation (1.3.9) can be shown as follows. Note that

$$\begin{aligned}
\sum_{u\in V} \pi_u^2 a_u^\top x(1 - a_u^\top x) &= \sum_{u\in V} \frac{d_u^2}{d(V)^2}\frac{d_u(S)}{d_u}\left(1 - \frac{d_u(S)}{d_u}\right) \\
&= \frac{1}{d(V)^2}\sum_{u\in V} d_u(S)d_u(S^c) \\
&= \frac{1}{d(V)^2}\sum_{v\in S}\sum_{w\in S^c}\sum_{u\in V} \mathbb{I}_{\{(v,u)\in E\}}\mathbb{I}_{\{(w,u)\in E\}} \\
&= \frac{1}{d(V)^2}|E_2(S, S^c)|
\end{aligned}$$

and $\pi^\top x(1 - \pi^\top x) = d(S)d(S^c)/d(V)^2$. We will assume that every node in G has a self-loop. The node interaction matrix A in this case can be interpreted to correspond to a lazy random walk that has a higher probability of remaining at a vertex for smaller degree vertices. In other words, a higher-degree vertex has a higher probability of adopting the state of a neighbor.

Lemma 1.3.6. *Assume that node interaction matrix A is according to (1.3.8) and every node in G has*

a self-loop. Then, we have

$$\frac{1}{\Phi_A} \leq \frac{1}{2} d(V) \frac{1}{\Phi(G)}.$$

Together with Corollary 1.3.2, Lemma 1.3.6 implies $\mathbb{E}^0[\tau] = O((d(V)/\Phi(G)) \log(n))$. By the same arguments as in the proof of Lemma 1.3.4, we have the following lemma.

Lemma 1.3.7. *Assume that node interaction matrix A is according to (1.3.8). Then, we have*

$$\tilde{\Psi}_A \leq 2\Phi(G).$$

From Lemmas 1.3.6 and 1.3.7, we have the following lemma.

Lemma 1.3.8. *Assume that node interaction matrix A is according to (1.3.8) and every node in G has a self-loop. Then, we have*

$$\frac{1}{\Phi_A} \leq d_{\min} \frac{1}{\Psi_A}.$$

Together with Corollary 1.3.2, Lemma 1.3.8 implies that our bound on the expected consensus time is at most a $O(d_{\min} \log(n))$ factor of the expected consensus time bound in Cooper and Rivera [2016].

For the asynchronous discrete-time voter model with A according to (1.3.8), we have

$$\mathbb{E}^0[\tau] \leq \frac{1}{\Phi'_A} \log\left(\frac{d(V)}{2d_{\min}}\right)$$

where

$$\Phi'_A = \frac{1}{n} \min_{S \subset V : 0 < |S| < n} \left\{ \frac{\sum_{u \in S, v \in S^c} (d_u + d_v) \mathbb{I}_{\{(u,v) \in E\}}}{d(S) d(S^c)} \right\}.$$

From Cooper and Rivera [2016], $\mathbb{E}^0[\tau] \leq 64/\Psi_A$ where

$$\Psi_A = \frac{2}{n} \frac{d_{\min}}{d(V)} \min_{S \subset V : 0 < |S| < n} \frac{|E(S, S^c)|}{\min\{d(S), d(S^c)\}}.$$

From the above relations, we have the following lemma.

Lemma 1.3.9. *For the asynchronous discrete-time voter model with A according to (1.3.8), we have*

$$\frac{1}{\Phi'_A} \leq \frac{1}{\Psi_A}.$$

The last lemma implies that our bound on the expected consensus time is within a logarithmic factor in n to the expected consensus time bound in Cooper and Rivera [2016].

We next provide explicit characterizations of Φ_A and bounds on the expected consensus time, from Corollary 1.3.2, for node interactions such that $a_{u,v} = 1/d_v$ for $(u, v) \in E$, for the case of a complete graph and a cycle. Note that it holds

$$\Phi_A = \min_{S \subset V : 0 < |S| < n} \frac{|E_2(S, S^c)|}{d(S)d(S^c)} \tag{1.3.10}$$

where $E_2(S, S^c)$ is the set of paths with two edges connecting S and S^c.

Complete graph K_n Let G be the complete graph with n vertices. Then, for any $S \subseteq V$,

$$\begin{aligned} |E_2(S, S^c)| &= |S|(n - |S|)(n - 2) \\ d(S) &= |S|(n - 1) \\ d(S^c) &= (n - |S|)(n - 1). \end{aligned}$$

Using this in (1.3.10), we have

$$\Phi_A = \frac{n - 2}{(n - 1)^2} = \frac{1}{n}(1 + o(1)).$$

Combining with $\pi^* = 1/n$, and Corollary 1.3.2, we have

$$\mathbb{E}^0[\tau] \leq n \log(n)(1 + o(1)).$$

Cycle C_n Let G be the cycle with $n \geq 3$ nodes. Then, for any $S \subseteq V$, we have

$$d(S) = 2|S| \text{ and } d(S^c) = 2(n - |S|).$$

Conditional on $|S|$, the smallest value of $|E_2(S, S^c)|$ is achieved when vertices in S are adjacent. In this case, we distinguish two cases. First, if $|S| = n - 1$, then, $|E_2(S, S^c)| = 2$, and $|E_2(S, S^c)|/(d(S)d(S^c)) = 1/(2(n-1))$. Second, if $|S| < n - 1$, then $|E_2(S, S^c)| = 4$, and thus

$$\frac{|E_2(S, S^c)|}{d(S)d(S^c)} = \frac{1}{|S|(n - |S|)}.$$

The minimum over $|S|$ is achieved for $|S| = n/2$ if n is even, otherwise, it is achieved for $(n-1)/2$. For n even, $\Phi_A = 4/n^2$ and, otherwise, $\Phi_A = 4/(n^2 - 1)$. Therefore, we have

$$\Phi_A = 4\frac{1}{n^2}(1 + o(1)).$$

Combining with $\pi^* = 1/n$, and Corollary 1.3.2, we have

$$\mathbb{E}^0[\tau] \leq \frac{1}{4}n^2 \log(n)(1 + o(1)).$$

1.4 Parameter estimation

In this section we first show an upper bound for the parameter estimation error, using a maximum likelihood estimator with a regulaizer, for the voter model parameter A^* from observed node states over time for $m \geq 1$ independent voter model process realizations with parameter A^* and independent initial states according to distribution μ. We then show a lower bound for the parameter estimation error for the class of locally stable estimators.

1.4.1 Parameter estimation error upper bound

Let $A \mapsto \mathcal{L}(A; X)$ denote a loss function of the voter model for given observation data X, where X are observed node states for m independent voter model process realizations with parameter A^* and initial state distribution μ,

$$X = (X_0^{(1)}, \ldots, X_{\tau_1}^{(1)}, \ldots, X_0^{(m)}, \ldots, X_{\tau_m}^{(m)})^\top.$$

We define estimator $\hat{A}$ as a minimizer of the loss function $\mathcal{L}(A; X)$, i.e.,

$$\hat{A} \in \arg\min_{A \in \Theta} \{\mathcal{L}(A; X)\} \tag{1.4.1}$$

where Θ is some given set of parameters.

Specifically, we will consider the loss function $\mathcal{L}(A; X)$ defined as the sum of the negative log-likelihood function and a regularizer defined as

$$\mathcal{L}(A; X) = -\ell(A; X) + \lambda_m ||A||_{1,1} \tag{1.4.2}$$

where $\ell(A)$ is the log-likelihood function and $\lambda_m \geq 0$ is the regularization parameter. Let $\Delta = \hat{A} - A^*$ denote the parameter estimation error. As common in statistical inference theory, we will measure the parameter estimation error by the Frobenious norm $||\Delta||_F$.

The log-likelihood function can be expressed as

$$\ell(A; X) = \sum_{i=1}^{m} \left(\log(\mu(X_0^{(i)})) - \sum_{u=1}^{n} \sum_{t=0}^{\tau_i - 1} H(X_{t+1,u}^{(i)}, a_u^\top X_t^{(i)}) \right) \tag{1.4.3}$$

where $H(p, q)$ is the cross-entropy between two Bernoulli distributions with mean values p and q, i.e.

$$H(p, q) = -(p \log(q) + (1 - p) \log(1 - q)).$$

Let us define

$$T_{u,i} = \{t \in \{0, \ldots, \tau_i\} : \ 0 < a_u^\top X_t^{(i)} < 1\}.$$

Intuitively, $T_{u,i}$ is the set of time steps at which the state of the i-th voter model process is such that vertex u has a pair of neighbors in different states. Note that having such mixed neighborhood sets is necessary for the parameter estimation, as otherwise, no useful information can be gained from observed vertex states for the parameter estimation.

For our analysis we will consider the gradient vector $\nabla \ell(A; X)$ and the Hessian matrix $\nabla^2 \ell(A; X)$ of the log-likelihood function. The gradient vector $\nabla \ell(A; X)$ has elements given as follows

$$\frac{\partial}{\partial a_{u,v}} \ell(A; X) = \sum_{i=1}^{m} \sum_{t \in T_{u,i}} \left(\frac{X_{t+1,u}^{(i)}}{a_u^\top X_t^{(i)}} - \frac{1 - X_{t+1,u}^{(i)}}{1 - a_u^\top X_t^{(i)}} \right) X_{t,v}^{(i)}.$$

The Hessian matrix $\nabla^2 \ell(A; X)$ has elements given as follows

$$\frac{\partial^2}{\partial a_{u,v} \partial a_{u,w}} \ell(A; X) = -\sum_{i=1}^{m} \sum_{t \in T_{u,i}} \varphi_u(X_t^{(i)}, X_{t+1}^{(i)}) X_{t,v}^{(i)} X_{t,w}^{(i)} \tag{1.4.4}$$

where

$$\varphi_u(X, Y) = \frac{Y_u}{(a_u^\top X)^2} + \frac{1 - Y_u}{(1 - a_u^\top X)^2}$$

and

$$\frac{\partial^2}{\partial a_{u,v} \partial a_{u',w}} \ell(A; X) = 0, \text{ if } u \neq u'. \tag{1.4.5}$$

For bounding the parameter estimation error, we use the framework for analysis of M-estimators with decomposable regularizers from high-dimensional statistics, e.g. Negahban et al. [2012c], Wainwright [2019]. The parameter estimator defined by (1.4.1) with the loss function (1.4.2) is an instance of an M-estimator with a decomposable regularizer.

For any set $S \subseteq V^2$ and $A \in \mathbb{R}^{n \times n}$, let A_S be the $n \times n$ matrix with support restricted to S, i.e. A_S is such that $(A_S)_{u,v} = a_{u,v}$ if $(u, v) \in S$ and $(A_S)_{u,v} = 0$ if $(u, v) \in S^c = V^2 \setminus S$. For any set $S \subseteq V^2$, let

$$\mathbb{C}(S; A^*) := \{\Delta : ||\Delta_{S^c}||_{1,1} \leq 3||\Delta_S||_{1,1} + 4||A^*_{S^c}||_{1,1}\}.$$

When the support of A^* is contained in S, we have $||A^*_{S^c}||_{1,1} = 0$.

For a given positive integer s, let S^* be a minimizer of $||A^*_{S^c}||_{1,1}$ over $S \subseteq V$ such that $|S| \leq s$. Let $\mathbb{C}^* := \mathbb{C}(S^*; A^*)$.

For any differentiable loss function $\mathcal{L} : \mathbb{R}^{n \times n} \mapsto \mathbb{R}$, we define the *first-order Taylor error* as

$$\mathcal{E}(\Delta) = \mathcal{L}(A^* + \Delta) - \mathcal{L}(A^*) - \nabla \mathcal{L}(A^*)^\top vec(\Delta)$$

where $vec(\Delta)$ denotes the vector defined by stacking the rows of matrix Δ.

A key concept in the framework of M-estimators is that of restricted strongly convex functions which is defined as follows.

Definition 1.4.1. *(Restricted Strong Convexity (RSC)) A loss function $\mathcal{L}$ satisfies restricted strong*

convexity relative to A^* *and* $S \subseteq V^2$ *with curvature* $\kappa > 0$ *and tolerance* γ^2 *if, for all* $\Delta \in \mathbb{C}(S; A^*)$,

$$\mathcal{E}(\Delta) \geq \kappa ||\Delta||_F^2 - \gamma^2.$$

The following bound on the parameter estimation error follows from the framework of M-estimators with decomposable regularizers (e.g. Theorem 1 in Negahban et al. [2012c]).

Theorem 1.4.1. *Assume that the loss function* $\mathcal{L}(A; X)$ *in (1.4.2) has the regularization parameter* λ_m *such that*

$$\lambda_m \geq 2||\nabla \ell(A^*)||_\infty \tag{1.4.6}$$

and, for some $S \subseteq V^2$, *the negative log-likelihood function* $-\ell(A; X)$ *satisfies the RSC condition relative to* A^* *and* S *with curvature* $\kappa > 0$ *and tolerance* γ^2. *Then, we have*

$$||\hat{A} - A^*||_F^2 \leq 9|S| \left(\frac{\lambda_m}{\kappa}\right)^2 + \left(2\gamma^2 \frac{1}{m} + 4||A^*_{S^c}||_{1,1}\right) \frac{\lambda_m}{\kappa}.$$

To bound the parameter estimation error for the voter model, we need to show (1) that condition (1.4.6) holds for the extended voter model with a given probability and (2) that the negative log-likelihood function satisfies the RSC condition with a given probability.

We first show a lemma that allows us to set the regularization parameter λ_m such that condition in (1.4.6) holds with high probability.

Lemma 1.4.1. *For any* $\delta \in (0, 1]$, *and any* $m \geq 1$ *independent realizations of the voter model process with parameter* A^* *and initial distribution* μ, *with probability at least* $1 - \delta$,

$$||\nabla \ell(A^*)||_\infty \leq \sqrt{2} \frac{1}{\alpha} \frac{1}{\sqrt{\Phi_{A^*}}} \sqrt{m} c_{n,\delta,\pi^*}(m) \tag{1.4.7}$$

where

$$c_{n,\delta,\pi^*}(m)^2 := \left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m} \log\left(\frac{2n^2}{\delta}\right)\right) \log\left(\frac{4n^2}{\delta}\right).$$

The proof of the lemma bounds the probability that $||\nabla \ell(A^*)||_\infty$ exceeds a fixed value with the sum

of probabilities of two events. One of these events is a deviation event for the sum of a bounded-difference martingale sequence defined by the sequence of gradients of the log-likelihood function over a fixed horizon time; which we bound by using Azuma-Hoeffding's inequality. The other event is the probability that the sum of consensus times of m independent voter model processes exceeds the value of the fixed horizon time; which we bound by using the probability tail bound for the sum of consensus times in Theorem 1.3.2.

Note that the bound on $||\nabla \ell(A^*)||_\infty$ in Lemma 1.4.1 involves the term $\sqrt{m/\Phi_{A^*}}$. In view of the bound on the expected consensus time in Corollary 1.3.2, we may intuitively think of the term $\sqrt{m/\Phi_{A^*}}$ as an upper bound on the square-root of the expected number of observed time steps of the extended voter model process, i.e. $\sqrt{m\mathbb{E}^0[\tau]}$. Note also that the bound (1.4.7) in Lemma 1.4.1 remains to hold by replacing $c_{n,\delta,\pi^*}(m)$ with $c_{n,\delta,\pi^*}(1)$, in which case the right-hand side in (1.4.7) scales with m as $\sqrt{m}$.

We can lower bound the first-order Taylor error function as follows.

Lemma 1.4.2. *Assume that A^* and $A^* + \Delta$ with $\Delta = (\Delta_1, \ldots, \Delta_n)^\top$ have a common support. Then, we have*

$$\mathcal{E}(\Delta) \geq h(\Delta; X)$$

where

$$h(\Delta; X) := \sum_{i=1}^{m} \sum_{t=0}^{\tau_i - 1} \sum_{u=1}^{n} \left(\Delta_u^\top X_t^{(i)} \right)^2 .$$

By Lemma 1.4.2, in order to show that the first-order Taylor error function $\mathcal{E}(\Delta)$ satisfies the RSC condition in Definition 1.4.1, it suffices to show that the RSC condition holds for function $h(\Delta; X)$. We first show that the RSC condition holds for the expected value of $h(\Delta; X)$ for any fixed value of Δ.

Lemma 1.4.3. *For any $\Delta = (\Delta_1, \ldots, \Delta_n)^\top$, we have*

$$\mathbb{E}^0[h(\Delta; X)] \geq \kappa_1 ||\Delta||_F^2$$

for any $\kappa_1 > 0$ such that

$$\kappa_1 \leq m\mathbb{E}^0[\tau] \lambda_{\min}(\mathbb{E}[X_0 X_0^\top]).$$

The correlation matrix $\mathbb{E}[X_0 X_0^\top]$ is with respect to the stationary distribution of the extended voter

process that does not include final consensus states of individual voter processes. The smallest eigenvalue $\lambda_{\min}(\mathbb{E}[X_0 X_0^\top])$ plays an important role in Lemma 1.4.3 and the results that follow. Note that by the Palm inversion formula (1.2.2),

$$\lambda_{\min}(\mathbb{E}[X_0 X_0^\top]) = \frac{1}{\mathbb{E}^0[\tau]} \lambda_{\min}\left(\mathbb{E}^0\left[\sum_{t=0}^{\tau-1} X_t X_t^\top\right]\right). \tag{1.4.8}$$

From (1.4.8), it can be readily observed that

$$\lambda_{\min}(\mathbb{E}[X_0 X_0^\top]) \geq \frac{1}{\mathbb{E}^0[\tau]} \lambda_{\min}(\mathbb{E}^0[X_0 X_0^\top]). \tag{1.4.9}$$

If the initial state distribution μ is of product-form with Bernoulli (p) marginal distributions, with $0 < p < 1$, then $\lambda_{\min}(\mathbb{E}^0[X_0 X_0^\top]) = p(1-p)$, and we have

$$\lambda_{\min}(\mathbb{E}[X_0 X_0^\top]) \geq p(1-p)\frac{1}{\mathbb{E}^0[\tau]}.$$

This bound is not tight. Tighter bounds can be obtained by analysis of the spectrum of the stationary correlation matrix $\mathbb{E}[X_0 X_0^\top]$ by using the Lyapunov matrix equation, which we discuss in Section 1.5.14. For example, for the complete graph case, when $a_{u,u} = 0$ and $a_{u,v} = 1/(n-1)$ for all $u \neq v$, we have

$$\lambda_{\min}(\mathbb{E}[X_0 X_0^\top]) = p(1-p)\frac{n}{\mathbb{E}^0[\tau]}(1 + o(1)).$$

We next show that for any fixed value Δ, $h(\Delta; X)$ satisfies the RSC condition in Definition 1.4.1 with a prescribed probability, provided that the number of observations m is sufficiently large.

Lemma 1.4.4. *For a voter model process with parameter A^* and initial distribution μ, for any $\delta \in (0, 1/2]$, any $S \subseteq V^2$ such that $|S| \leq s$ for some positive integer s, and any $\Delta \in \mathbb{C}(S, A^*)$, $h(\Delta; X)$ satisfies the RSC condition relative to A^* and S, with curvature $\kappa = \kappa_1/2$ and tolerance $\gamma = 0$, with probability at least $1 - \delta$, under condition*

$$m \geq \frac{s^2}{\Phi_{A^*}} \frac{1}{\mathbb{E}^0[\tau]^2 \lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^2} c_{\delta,\pi^*}(m),$$

where

$$c_{\delta,\pi^*}(m) = 8\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{2}{\delta}\right)\right)\log\left(\frac{2}{\delta}\right).$$

We next show that $h(\Delta; X)$ satisfies the RSC condition in Definition 1.4.1 for every $\Delta \in C^*$ with certain probability.

Lemma 1.4.5. *For a voter model process with parameter A^* and initial distribution μ, function $h(\Delta; X)$ satisfies the RSC condition relative to A^* and S^*, for every $\Delta \in \mathbb{C}^*$ with probability at least $1 - 4/n$,*

$$h(\Delta; X) \geq \kappa' ||\Delta||_F^2 - {\gamma'}^2 \text{ for all } \Delta \in \mathbb{C}^*$$

where $\kappa' = \kappa_1/8$ and $\gamma' = \sqrt{\kappa_1/8}||A^*_{S^*}||_{1,1}/\sqrt{s}$, provided that

$$m \geq m_1 := c_1 s^2 \frac{\log(1/(2\pi^*))(a+1)\log(n) + (a+1)^2 \log(n)^2}{\Phi_{A^*}\mathbb{E}^0[\tau]^2 \lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^2} \tag{1.4.10}$$

and

$$m \geq m_2 := c_2 n^3 (1/\pi^*) \frac{1}{(\Phi_{A^*}\mathbb{E}^0[\tau])^2 \lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^2} \tag{1.4.11}$$

where

$$a = sn \frac{\log(1/(2\pi^*)) + \log(n)}{(\Phi_{A^*}\mathbb{E}^0[\tau])\lambda_{\min}(\mathbb{E}[X_0 X_0^\top])}$$

for some constants $c_1, c_2 > 0$.

The proof of Lemma 1.4.5 relies on some set covering arguments to bound the probability of events indexed with Δ, which takes values in the infinite set $\mathbb{C}^*$. These covering arguments require stronger conditions on the number of voter model realizations m than in Lemma 1.4.4, which shows that the RSC condition holds in probability, for any fixed value Δ.

Consider the case when the voter model process has the stationary distribution π of A^* such that π^* is lower bounded by a polynomial in $1/n$. Then, a sufficient condition for (1.4.10) is that for some constant $c > 0$,

$$m \geq c \frac{1}{\mathbb{E}^0[\tau]} \frac{s^4 n^2}{\lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^4} \frac{\log(n)^2}{(\Phi_{A^*}\mathbb{E}^0[\tau])^3}.$$

We next present our main theorem that provides a bound on the parameter estimation error.

Theorem 1.4.2. *Consider the voter model process with parameter A^* with support size s. Assume that*

$\hat{A}$ is a minimizer of the loss function $\mathcal{L}(A; X)$ defined by (1.4.2) with the regularization parameter

$$\lambda_m = 2\sqrt{2}\frac{c_{n,\pi^*}}{\alpha\sqrt{\Phi_{A^*}}}\sqrt{m},$$

and conditions (1.4.10) and (1.4.11) hold. Then, for some constant $c > 0$, with probability at least $1 - 5/n$,

$$||\hat{A} - A^*||_F^2 \leq c\frac{sc_{n,\pi^*}^2}{\alpha^2(\Phi_{A^*}\mathbb{E}^0[\tau])^2\lambda_{\min}(\mathbb{E}[X_0X_0^\top])^2}\Phi_{A^*}\frac{1}{m} \tag{1.4.12}$$

where

$$c_{n,\pi^*}^2 = \left(\log\left(\frac{1}{2\pi^*}\right) + \log\left(2n^3\right)\right)\log\left(4n^3\right).$$

Theorem 1.4.2 gives us a bound on the parameter estimation error in (1.4.12) that holds with high probability under sufficient conditions (1.4.10) and (1.4.11) for consistency of the estimator. For any initial distribution μ and parameter A^* such that $\mathbb{E}^0[\tau]$ is equal to $1/\Phi_{A^*}$ up to a poly-logarithmic factor in n, the term $\Phi_{A^*}\mathbb{E}^0[\tau]$ contributes only poly-logarithmic factors in (1.4.12). In the case when the product $\Phi_{A^*}\mathbb{E}^0[\tau]$ is poly-logarithmic in n, for asymptotically large m,

$$m||\hat{A} - A^*||_F^2 = \tilde{O}\left(\frac{s\Phi_{A^*}}{\alpha^2\lambda_{\min}(\mathbb{E}[X_0X_0^\top])^2}\right).$$

1.4.2 Sampling complexity lower bound

In this section, we show a lower bound on the sampling complexity for the parameter estimation of the voter model. This lower bound is derived using the framework of locally stable estimators Jedra and Proutiere [2019]. Intuitively, a locally stable estimator of a parameter is robust to small perturbations of the true parameter value.

To make a precise definition, let $\mathbb{B}(A^*, r)$ be the ball with centre point A^* and radius r, i.e. $\mathbb{B}(A^*, r) = \{A \in \Theta : ||A^* - A||_F \leq r\}$, where Θ is some set of parameter values. Let $\mathbb{P}_A[\cdot]$ denote the probability distribution under a statistical model with parameter A. The notion of locally stable estimators is defined as follows.

Definition 1.4.2. *An estimator $\hat{A}$ is said to be (ϵ, δ)-locally stable in A^* with parameters $\epsilon > 0$ and*

$\delta \in (0,1)$, *if there exists a finite* m_0 *such that for all* $m \geq m_0$ *and* $A \in \mathbb{B}(A^*, 3\epsilon)$,

$$\mathbb{P}_A[||\hat{A} - A||_F \leq \epsilon] \geq 1 - \delta.$$

Roughly speaking, for any locally stable estimator in A^*, a given bound on the parameter estimation error holds in probability with respect to any parameter A that is in a neighborhood of A^*. In our setting, we let Θ be the set of $n \times n$ substochastic matrices.

The following theorem gives a lower bound on the sampling complexity for the class of locally stable estimators.

Theorem 1.4.3. *Assume that* A^* *is a stochastic matrix such that each element in its support has value at least* $\alpha > 0$. *Let* q_1 *be the eigenvector corresponding to the smallest eigenvalue of the correlation matrix* $\mathbb{E}[X_0 X_0^\top]$ *of the extended voter process with parameter* A^* *and initial state distribution* μ.

Then, for any (ϵ, δ)*-locally stable estimator in* A^*, *such that* $\delta \in (0,1)$ *and* $\epsilon \in (0, \min\{1/(2|q_1^\top \mathbf{1}|), \alpha/4\})$, *it holds*

$$m\mathbb{E}^0[\tau]\lambda_{\min}(\mathbb{E}[X_0 X_0^\top]) \geq \frac{\alpha}{16}\frac{1}{\epsilon^2}\log\left(\frac{1}{2.4\delta}\right). \tag{1.4.13}$$

Moreover, (1.4.13) holds under stronger condition $\epsilon \in (0, \min\{1/(2\sqrt{n}), \alpha/4\})$.

The proof of the theorem follows similar arguments as in the proof of a sampling complexity lower bound for linear discrete-time dynamical systems with additive Gaussian noise Jedra and Proutiere [2019]. The extended voter process requires us to study a different discrete-time random dynamical system, and addressing certain technical points that arise due to Bernoulli random variables and constraints on the node interaction parameters.

The result in Theorem 1.4.3 shows us that the upper bound in Theorem 1.4.2 is tight with respect to the relation between $||\hat{A} - A^*||_F$ and m for small estimation error case. If δ is polynomial in $1/n$ and $\epsilon^2 \leq \min\{1/n, \alpha^2\}/16$, then from (1.4.13) we have

$$m\epsilon^2 = \Omega\left(\frac{\alpha}{\mathbb{E}^0[\tau]\lambda_{\min}(\mathbb{E}[X_0 X_0^\top])}\log(n)\right).$$

1.5 Proofs and additional results

1.5.1 Mathematical background

1.5.1.1 KL divergence bounds

Let p and q be two distributions on $\mathcal{X}$ such that $p(x) = 0$ for all $x \in \mathcal{X}$ such that $q(x) = 0$. The KL divergence between p and q is defined by

$$\mathrm{KL}(p \mid\mid q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

The total variation distance between p and q is defined by

$$\delta(p, q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)|.$$

The Pinsker's inequality is

$$\mathrm{KL}(p \mid\mid q) \geq 2\delta(p, q)^2. \tag{1.5.1}$$

Because $(\sum_{x \in \mathcal{X}} |p(x) - q(x)|)^2 \geq ||p - q||^2$, it follows

$$\mathrm{KL}(p \mid\mid q) \geq \frac{1}{2} ||p - q||^2. \tag{1.5.2}$$

Let $\alpha > 0$ be a constant such that $q(x) \geq \alpha$ for all $x \in \mathcal{X}$ such that $q(x) > 0$. Then, we have the following upper bound for the KL divergence:

$$\mathrm{KL}(p \mid\mid q) \leq \frac{1}{\alpha} ||p - q||^2. \tag{1.5.3}$$

The proof is easy and is provided here for completeness:

$$\begin{aligned}
\mathrm{KL}(p \,||\, q) &= \sum_{x \in \mathcal{X}: q(x)>0} p(x) \log\left(\frac{p(x)}{q(x)}\right) \\
&= \sum_{x \in \mathcal{X}: q(x)>0} q(x) \left(\frac{p(x)-q(x)}{q(x)} + 1\right) \log\left(\frac{p(x)-q(x)}{q(x)} + 1\right) \\
&\leq \sum_{x \in \mathcal{X}: q(x)>0} q(x) \left(\frac{p(x)-q(x)}{q(x)} + 1\right) \frac{p(x)-q(x)}{q(x)} \\
&= \sum_{x \in \mathcal{X}: q(x)>0} \frac{(p(x)-q(x))^2}{q(x)} \\
&\leq \frac{1}{\alpha}||p-q||^2
\end{aligned}$$

where the first inequality follows by the fact that $\log(x+1) \leq x$ for all $x > -1$ and the last inequality follows by the definition of α.

Suppose that p and q are two Bernoulli distributions with parameters $\bar{p}$ and $\bar{q}$, respectively. With a slight abuse of notation, let $\mathrm{KL}(\bar{p} \,||\, \bar{q})$ denote the KL divergence between p and q. Using Pinsker's inequality (1.5.1) and $\delta(p,q) = \frac{1}{2}(|\bar{p}-\bar{q}| + |\bar{p}-\bar{q}|) = |\bar{p}-\bar{q}|$, we have the lower bound

$$\mathrm{KL}(\bar{p} \,||\, \bar{q}) \geq 2(\bar{p}-\bar{q})^2. \tag{1.5.4}$$

Using (1.5.3), under $\alpha \leq \bar{q} \leq 1-\alpha$, we have the upper bound

$$\mathrm{KL}(\bar{p} \,||\, \bar{q}) \leq \frac{2}{\alpha}(\bar{p}-\bar{q})^2. \tag{1.5.5}$$

1.5.1.2 Concentration of measure inequalities

Theorem 1.5.1 (Azuma-Hoeffding)**.** *Assume $X_0, X_1, \ldots$ is a martingale sequence such that $|X_i - X_{i-1}| \leq c_i$ almost surely. Then, for all positive integers N and all $\epsilon > 0$,*

$$\mathbb{P}[X_N - X_0 \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^N c_i^2}\right)$$

with an identical bound for the other tail.

1.5.1.3 Covering and metric entropy

Let (M, ρ) be a metric space, where M is a set and $\rho : M \times M \to \mathbb{R}_+$ is a metric.

An *ϵ-covering* of M in metric ρ is a collection of points $\{x_1, \ldots, x_k\} \subset M$ such that for every $x \in M$, $\rho(x, x_i) \leq \epsilon$, for some $i \in \{1, \ldots, k\}$.

The *ϵ-covering number*, denoted as $N(\epsilon, M, \rho)$, is the cardinality of the smallest ϵ-covering of M in metric ρ. In other words, $N(\epsilon, M, \rho)$ is the minimum number of balls with radius ϵ under metric ρ required to cover M.

Let $B_M(x, \epsilon, \rho)$ be the closed ball with center x and radius ϵ under metric ρ, i.e.

$$B_M(x, \epsilon, \rho) = \{y \in M : \rho(x, y) \leq \epsilon\}.$$

Then,

$$N(\epsilon, M, \rho) = \min\{k \in \mathbb{N} : \exists x_1, \ldots, x_k : M \subset \cup_{i=1}^{k} B_M(x_i, \epsilon, \rho)\}.$$

The *metric entropy* is defined as the logarithm of the covering number, i.e. $\log(N(\epsilon, M, \rho))$.

The *dyadic entropy number* $\epsilon_k(M, \rho)$ is defined as

$$\epsilon_k(M, \rho) = \inf\{\epsilon > 0 : N(\epsilon, M, \rho) \leq 2^{k-1}\}.$$

Note that $\epsilon_k(M, \rho) \leq \epsilon$ if and only if $\log(N(\epsilon, M, \rho)) \leq k$.

Let $M = \mathbb{R}^d$ and let $B_q(r)$ be the closed ball with center 0 and radius r under metric ℓ_q. By Raskutti et al. [2011], for every $q \in (0, 1]$ and $p \in [1, \infty]$ such that $p > q$, there exists a constant $c_{q,p}$ such that

$$\log(N(\epsilon, B_q(r), \ell_p)) \leq c_{q,p} r^{p/(p-q)} \left(\frac{1}{\epsilon}\right)^{1/(1/q-1/p)} \log(d), \text{ for all } \epsilon \in (0, r^{1/q}).$$

In particular, for $q = 1$ and $p = 2$, for some constant $c > 0$,

$$\log(N(\epsilon, B_1(r), \ell_2)) \leq c \left(\frac{r}{\epsilon}\right)^2 \log(d), \text{ for all } \epsilon \in (0, r). \tag{1.5.6}$$

In the following lemma we provide a bound on the metric entropy for a certain metric ρ that is of interest for our parameter estimation problem. A similar lemma was stated in Pandit et al. [2019]

(Lemma A.8) without a proof. Our lemma shows that parameters c_1 and c_2 in the lemma are not constants but depend on matrix X. This has significant implications on required conditions when applying the lemma to the parameter estimation problem.

Lemma 1.5.1. *Let X be a real $T \times n$ matrix with each column having ℓ_2 norm bounded by $\sqrt{T}$. Let $B_1(r) = \{\Delta \in \mathbb{R}^{n \times n} : ||\Delta||_1 \leq r\}$ and $\rho(\Delta, \Delta') = 1/\sqrt{T}||X(\Delta - \Delta')^\top||_F$. Then, there exist constant $c > 0$ such that the metric entropy of $B_1(r)$ in ρ is bounded as*

$$\log(N(\epsilon, B_1(r), \rho)) \leq c_1 \left(\frac{r}{\epsilon}\right)^2 \log(n), \text{ for all } \epsilon \in (0, c_2 r]$$

where $c_1 = c\sigma_{\max}(X)^2/T$ and $c_2 = \sigma_{\max}(X)/\sqrt{T}$.

Proof. For any real $k \times n$ matrix M and real $n \times n$ matrix D with $k \geq n$, we have

$$||MD||_F \leq ||M||||D||_F = \sigma_{\max}(M)||D||_F$$

where $\sigma_{\max}(M)$ denotes the largest singular value of M.

Hence, we have

$$1/\sqrt{T}||X(\Delta - \Delta')^\top||_F \leq (\sigma_{\max}(X)/\sqrt{T})||\Delta - \Delta'||_F.$$

Let $\rho'(\Delta, \Delta') = (\sigma_{\max}(X)/\sqrt{T})||\Delta - \Delta'||_F$, and $N(\epsilon, B_1(r), \rho')$ be the minimum number of balls with radius ϵ under metric ρ' to cover $B_1(r)$, and $\epsilon' = \sqrt{T}\epsilon/\sigma_{\max}(X)$. Then, we have

$$\begin{aligned} N(\epsilon, B_1(r), \rho) &\leq N(\epsilon, B_1(r), \rho') \\ &= N(\epsilon', B_1(r), \ell_2) \\ &\leq e^{c_1\left(\frac{r}{\epsilon}\right)^2 \log(n)} \end{aligned}$$

for all $\epsilon \in (0, c_2 r)$. The first inequality comes from the fact that ρ' balls take up less space than ρ balls, so it would take more of them to do the covering, and the second inequality comes from (1.5.6). □

1.5.2 Proof of Lemma 1.2.1

Lemma 1.2.1 is a known result. We provide a proof for the sake of completeness.

We claim that for every $t \geq 0$ and $x \notin C$,

$$V(x) \geq \lambda^{-t}\mathbb{E}_x[V(X_t)\mathbb{I}_{\{\tau_C > t\}}] + \sum_{s=1}^{t} \lambda^{-s}\mathbb{P}_x[\tau_C = s]. \tag{1.5.7}$$

We will prove this by induction on t. Base case $t = 0$ trivially holds. For the induction step, assume that (1.5.7) holds for t, and we are going to prove that it then holds for $t+1$. For any $x \notin C$,

$$\begin{aligned}
& \mathbb{E}_x[V(X_t)\mathbb{I}_{\{\tau_C > t\}}] \\
\geq\ & \lambda^{-1}\mathbb{E}_x[PV(X_{t+1})\mathbb{I}_{\{\tau_C > t\}}] \\
=\ & \lambda^{-1}\mathbb{E}_x[V(X_{t+1}) \mid \mathbb{I}_{\{\tau_C > t\}}] \\
=\ & \lambda^{-1}\left(\mathbb{E}_x[V(X_{t+1})\mathbb{I}_{\{\tau_C > t+1\}}] + \mathbb{E}_x[V(X_{t+1})\mathbb{I}_{\{\tau_C = t+1\}}]\right) \\
\geq\ & \lambda^{-1}\mathbb{E}_x[V(X_{t+1})\mathbb{I}_{\{\tau_C > t+1\}}] + \lambda^{-1}\mathbb{P}_x[\tau_C = t+1]
\end{aligned}$$

where the first inequality is by the drift condition and the last inequality is because $V(x) \geq 1$ for all $x \in \mathcal{X}$.

Now, use the last inequality to obtain that (1.5.7) holds for $t+1$. It follows that

$$V(x) \geq \lambda^{-t}\mathbb{P}_x[\tau_C > t] + \sum_{s=1}^{t} \lambda^{-s}\mathbb{P}_x[\tau_C = s].$$

By letting t goes to infinity, we have

$$V(x) \geq \sum_{s=1}^{\infty} \lambda^{-s}\mathbb{P}[\tau_C = s] = \mathbb{E}_x[\lambda^{-\tau_C}].$$

1.5.3 Proof of Lemma 1.3.1

We first consider the case when $x \in \{0,1\}^n$ such that $x \notin C$. Note that

$$\begin{aligned}
& \mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] \\
=\ & \pi^\top \mathbb{E}[X_{t+1} \mid X_t = x] \\
& -\pi^\top x - (\mathbb{E}[(\pi^\top X_{t+1})^2 \mid X_t = x] - (\pi^\top x)^2) \\
& -(\mathbb{E}[(\pi^\top X_{t+1})^2 \mid X_t = x] - (\pi^\top x)^2)
\end{aligned} \tag{1.5.8}$$

where the last equation is by the fact that $\pi^\top X_t$ is a martingale.

Let

$$D_A(x) = \text{diag}(a_1^\top x(1 - a_1^\top x), \ldots, a_n^\top x(1 - a_n^\top x)).$$

Note that

$$\begin{aligned}
& \mathbb{E}[(\pi^\top X_{t+1})^2 \mid X_t = x] \\
= \; & \pi^\top \mathbb{E}[X_{t+1} X_{t+1}^\top \mid X_t = x]\pi \\
= \; & \pi^\top (A\mathbb{E}[X_t X_t^\top \mid X_t = x]A^\top + \mathbb{E}[D_A(X_t) \mid X_t = x])\pi \\
= \; & \pi^\top x x^\top \pi + \pi^\top D_A(x)\pi \\
= \; & (\pi^\top x)^2 + \pi^\top D_A(x)\pi.
\end{aligned}$$

Plugging the derived identity in (1.5.8), we obtain

$$\mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] = -\sum_{u=1}^{n} \pi_u^2 V_{a_u}(x).$$

Now, consider the case when $x \in C$. Then, $V_\pi(x) = 0$, and we have

$$\mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] = \mathbb{E}^0[V_\pi(X_0)].$$

Hence, we have shown that

$$\begin{aligned}
& \mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] \\
= \; & -\sum_{u=1}^{n} \pi_u^2 V_{a_u}(x) + \mathbb{E}^0[V_\pi(X_0)]\mathbb{I}_{\{x \in C\}}.
\end{aligned}$$

1.5.4 Proof of Lemma 1.3.2

We first consider the case when $x \in \{0, 1\}^n$ such that $x \notin C$. In this case, we have

$$\begin{aligned}
& \mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] \\
= \; & \pi^\top \mathbb{E}[X_{t+1} \mid X_t = x] - \pi^\top \mathbb{E}[X_{t+1} X_{t+1}^\top \mid X_t = x]\pi - \pi^\top x + \pi^\top x x^\top \pi.
\end{aligned}$$

Now, note

$$\mathbb{E}[X_{t+1} \mid X_t = x] = \frac{1}{n}Ax + \left(1 - \frac{1}{n}\right)x.$$

Hence, $\pi^\top \mathbb{E}[X_{t+1} \mid X_t = x] = \pi^\top x$. It follows

$$\mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] = -\pi^\top \mathbb{E}[X_{t+1}X_{t+1}^\top \mid X_t = x]\pi + \pi^\top x x^\top \pi.$$

For $u \neq v$, we have

$$\mathbb{E}[X_{t+1,u}X_{t+1,v} \mid X_t = x] = \frac{1}{n}a_u^\top x x_v + \frac{1}{n}x_u a_v^\top x + \left(1 - \frac{2}{n}\right)x_u x_v$$

and

$$\mathbb{E}[X_{t+1,u}X_{t+1,u} \mid X_t = x] = \frac{1}{n}a_u^\top x + \left(1 - \frac{1}{n}\right)x_u.$$

In a matrix notation, we have

$$\mathbb{E}[X_{t+1}X_{t+1}^\top \mid X_t = x] = \frac{1}{n}(Ax)x^\top + \frac{1}{n}x(Ax)^\top + \left(1 - \frac{2}{n}\right)xx^\top + \frac{1}{n}D_A(x)$$

where $D_A(x)$ is the diagonal matrix with diagonal elements

$$(D_A(x))_{u,u} = x_u(1 - a_u^\top x) + (1 - x_u)a_u^\top x.$$

It follows that

$$\pi^\top \mathbb{E}[X_{t+1}X_{t+1}^\top \mid X_t = x]\pi = \pi^\top x x^\top \pi + \frac{1}{n}\sum_{u=1}^{n} \pi_u^2(x_u(1 - a_u^\top x) + (1 - x_u)a_u^\top x)$$

and, hence,

$$\mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] = -\frac{1}{n}\sum_{u=1}^{n} \pi_u^2(x_u(1 - a_u^\top x) + (1 - x_u)a_u^\top x).$$

For the case when $x \in C$, we have

$$\mathbb{E}[V_\pi(X_{t+1}) - V_\pi(X_t) \mid X_t = x] = \mathbb{E}^0[V_\pi(X_0)].$$

1.5.5 Proof of Lemma 1.3.3

It is can be readily checked that $\pi_u = d_u/d(V)$, for $u \in V$.

Next, note

$$\begin{aligned}\sum_{u\in V} \pi_u^2 a_u^\top x(1-a_u^\top x) &= \frac{1}{4d(V)^2}\sum_{u\in V}(d_u\mathbb{I}_{\{u\in S\}} + d_u(S))(d_u\mathbb{I}_{\{u\in S^c\}} + d_u(S^c)) \\ &\geq \frac{d_{\min}}{4d(V)^2}\left(\sum_{u\in S} d_u(S^c) + \sum_{u\in S^c} d_u(S)\right) \\ &= \frac{d_{\min}}{2d(V)^2}|E(S,S^c)|.\end{aligned}$$

It follows

$$\begin{aligned}\frac{\sum_{u\in V}\pi_u^2 a_u^\top x(1-a_u^\top x)}{\pi^\top x(1-\pi^\top x)} &\geq \frac{d_{\min}}{2}\frac{|E(S,S^c)|}{d(S)d(S^c)} \\ &\geq \frac{d_{\min}}{2d(V)}\frac{|E(S,S^c)|}{\min\{d(S),d(S^c)\}}.\end{aligned}$$

Hence, we have

$$\frac{1}{\Phi_A} \leq 2\frac{d(V)}{d_{\min}}\frac{1}{\Phi(G)}.$$

1.5.6 Proof of Lemma 1.3.4

Note that

$$\begin{aligned}\mathbb{E}[|\pi^\top(I-Z)x|] &\leq \sum_{u=1}^n \pi_u \mathbb{E}\left[\left|x_u - \sum_{v=1}^n Z_{u,v}x_v\right|\right] \\ &= \sum_{u=1}^n \pi_u[x_u(1-a_u^\top x) + (1-x_u)(a_u^\top x)] \\ &= 2\sum_{u=1}^n\sum_{v=1}^n \pi_u a_{u,v}x_u(1-x_v)\end{aligned}$$

where the inequality is by Jensen's inequality. For A according to (1.3.8), we have

$$\sum_{u=1}^n\sum_{v=1}^n \pi_u a_{u,v}x_u(1-x_v) = \frac{1}{2d(V)}|E(S,S^c)|$$

and

$$\min\{\pi^\top x, 1-\pi^\top x\} = \frac{1}{d(V)}\min\{d(S), d(S^c)\}.$$

Hence, it follows

$$\tilde{\Psi}_A \leq \Phi(G).$$

1.5.7 Proof of Lemma 1.3.6

For every $S \subseteq V$, we have

$$\begin{aligned} |E_2(S,S^c)| &= \sum_{u\in V} d_u(S)d_u(S^c) \\ &= \sum_{u\in S^c} d_u(S)d_u(S^c) + \sum_{u\in S} d_u(S)d_u(S^c) \\ &\geq \sum_{u\in S^c} d_u(S) + \sum_{u\in S} d_u(S^c) \\ &= 2|E(S,S^c)| \end{aligned}$$

where the inequality holds by the fact that $d_u(S) \geq 1$ when $u \in S$, , for any set $S \subset V$, because each vertex in G has a self-loop.

Thus, we have

$$\sum_{u\in V} \pi_u^2 a_u^\top x(1-a_u^\top x) = \frac{1}{d(V)^2}|E_2(S,S^c)| \geq \frac{2}{d(V)^2}|E(S,S^c)|.$$

Combining with

$$\pi^\top x(1-\pi^\top x) = \frac{d(S)d(S^c)}{d(V)^2} \leq \frac{1}{d(V)}\min\{d(S), d(S^c)\},$$

and (1.3.5) and (1.3.9), we have

$$\frac{1}{\Phi_A} \leq \frac{1}{2}d(V)\frac{1}{\Phi(G)}.$$

1.5.8 Proof of Lemma 1.4.1

By the union bound, for any $\Lambda \geq 0$,

$$\mathbb{P}^0[||\nabla \ell(A^*;X)||_\infty \geq \Lambda] \leq n^2 \max_{(u,v)\in V^2} \mathbb{P}^0\left[\left|\frac{\partial}{\partial a_{u,v}}\ell(A^*;X)\right| \geq \Lambda\right].$$

Fix $u, v \in V$. Let us define

$$Y_t^{(i)} = \left(\frac{X_{t,u}^{(i)}}{a_u^{*\top} X_{t-1}^{(i)}} - \frac{1 - X_{t,u}^{(i)}}{1 - a_u^{*\top} X_{t-1}^{(i)}} \right) X_{t-1,v}^{(i)} \mathbb{I}_{\{0 < a_u^{*\top} X_{t-1}^{(i)} < 1\}}.$$

Note that

$$\mathbb{P}^0 \left[\left| \frac{\partial}{\partial a_{u,v}} \ell(A^*; X) \right| \geq \Lambda \right] = \mathbb{P}^0 \left[\left| \sum_{i=1}^{m} \sum_{t=1}^{\tau_i} Y_t^{(i)} \right| \geq \Lambda \right]. \tag{1.5.9}$$

For $0 < s \leq \sum_{i=1}^{m} \tau_i$, let us define

$$Y_s = Y^{(k)}_{s - \sum_{i=1}^{k-1} \tau_i}, \text{ for } \sum_{i=1}^{k-1} \tau_i < s \leq \sum_{i=1}^{k} \tau_i \text{ and } k \in [m]$$

where $\sum_{i=1}^{0} \tau_i \equiv 0$, and

$$Y_s = Y_{\tau_m}^{(m)} = 0, \text{ for } s > \sum_{i=1}^{m} \tau_i.$$

From (1.5.9), for any $T > 0$,

$$\mathbb{P}^0 \left[\left| \frac{\partial}{\partial a_{u,v}} \ell(A^*; X) \right| \geq \Lambda \right] \leq \mathbb{P}^0 \left[\left| \sum_{t=1}^{T} Y_t \right| \geq \Lambda \right] + \mathbb{P}^0 \left[\sum_{i=1}^{m} \tau_i > T \right].$$

For every $t \geq 0$, we have

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = 0.$$

Hence, $Y_1, Y_2, \ldots$ is a martingale difference sequence. For all $t \geq 1$, $|Y_t| \leq 1/\alpha$, with probability 1. Hence, we can apply Azuma-Hoeffding's inequality (Theorem 1.5.1) to obtain

$$\mathbb{P} \left[\left| \sum_{t=1}^{T} Y_t \right| \geq \Lambda \right] \leq 2e^{-\frac{\alpha^2 \Lambda^2}{2T}}. \tag{1.5.10}$$

By Theorem 1.3.2, we have

$$\mathbb{P}^0 \left[\sum_{i=1}^{m} \tau_i > T \right] \leq e^{-cm} \tag{1.5.11}$$

for any

$$T \geq m \left(\log \left(\frac{1}{2\pi^*} \right) + c \right) \frac{1}{\Phi_{A^*}}.$$

Let $\delta \in (0, 1]$. We require that the right-hand sides of the inequalities in (1.5.10) and (1.5.11) are less

than or equal to $\delta/(2n^2)$. This yields

$$\Lambda \geq \sqrt{2}\sqrt{T}\frac{1}{\alpha}\sqrt{\log\left(\frac{4n^2}{\delta}\right)}$$

and

$$T \geq m\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{2n^2}{\delta}\right)\right)\frac{1}{\Phi_{A^*}}.$$

Hence, with probability at least $1-\delta$,

$$||\nabla\ell(A^*)||_\infty \leq \frac{\sqrt{2}}{\alpha}\sqrt{m}\frac{1}{\sqrt{\Phi_{A^*}}}\sqrt{\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{2n^2}{\delta}\right)\right)\log\left(\frac{4n^2}{\delta}\right)}.$$

1.5.9 Proof of Lemma 1.4.2

By a limited Taylor expansion, for some $\lambda \in [0,1]$,

$$\mathcal{E}(\Delta) = vec(\Delta)^\top \nabla^2(-\ell(A^* + \lambda\Delta))vec(\Delta).$$

By the properties of the Hessian matrix of the negative log-likelihood function, (1.4.4) and (1.4.5), we have

$$\frac{\partial^2}{\partial a_{u,v}\partial a_{u,w}}(-\ell(A;X)) \geq \sum_{i=1}^{m}\sum_{t=0}^{\tau_i-1} X_{t,v}^{(i)}X_{t,w}^{(i)}\mathbb{I}_{\{0<a_u^\top X_t^{(i)}<1\}}.$$

Note that

$$\begin{aligned}
& vec(\Delta)^\top\nabla^2(-\ell(A;X))vec(\Delta) \\
\geq & \sum_{u=1}^{n}\sum_{v=1}^{n}\sum_{w=1}^{n}\Delta_{u,v}\left(\sum_{i=1}^{m}\sum_{t=0}^{\tau_i-1}X_{t,v}^{(i)}X_{t,w}^{(i)}\mathbb{I}_{\{0<a_u^\top X_t^{(i)}<1\}}\right)\Delta_{u,w} \\
= & \sum_{i=1}^{m}\sum_{t=0}^{\tau_i-1}\sum_{u=1}^{n}\left(\Delta_u^\top X_t^{(i)}\right)^2\mathbb{I}_{\{0<a_u^\top X_t^{(i)}<1\}}.
\end{aligned}$$

Hence,

$$vec(\Delta)^\top\nabla^2(-\ell(A;X))vec(\Delta) \geq h(A,\Delta;X)$$

where

$$h(A,\Delta;X) = \sum_{i=1}^{m}\sum_{t=0}^{\tau_i-1}\sum_{u=1}^{n}\left(\Delta_u^\top X_t^{(i)}\right)^2\mathbb{I}_{\{0<a_u^\top X_t^{(i)}<1\}}.$$

For $A = A^* + \lambda\Delta = (1-\lambda)A^* + \lambda(A^* + \Delta)$, since A^* and $A^* + \Delta$ have the same support, we have the following properties. If $a_u^\top X_t^{(i)} = 0$, then $\Delta_u^\top X_t^{(i)} = 0$ because ${a_u^*}^\top X_t^{(i)} = 0$. If $a_u^\top X_t^{(i)} = 1$, then again $\Delta_u^\top X_t^{(i)} = 0$ because ${a_u^*}^\top X_t^{(i)} = 1$. It thus follows that for any A with the same support as A^*, we have

$$h(A, \Delta; X) = h(\Delta; X) := \sum_{i=1}^{m} \sum_{t=0}^{\tau_i - 1} \sum_{u=1}^{n} \left(\Delta_u^\top X_t^{(i)}\right)^2.$$

1.5.10 Proof of Lemma 1.4.3

We consider

$$\mathbb{E}^0[h(\Delta; X)] = m\mathbb{E}^0[\tau] \sum_{u=1}^{n} \mathbb{E}\left[\left(\Delta_u^\top X_0\right)^2\right].$$

The following relations hold

$$\begin{aligned}
\sum_{u=1}^{n} (\Delta_u^\top x)^2 &= \sum_{u=1}^{n} x^\top \Delta_u \Delta_u^\top x \\
&= x^\top \left(\sum_{u=1}^{n} \Delta_u \Delta_u^\top\right) x \\
&= x^\top \Delta^\top \Delta x \\
&= \langle \Delta x, \Delta x \rangle \\
&= \mathrm{tr}((\Delta x)^\top \Delta x) \\
&= \mathrm{tr}(x^\top \Delta^\top \Delta x) \\
&= \mathrm{tr}(x x^\top \Delta^\top \Delta).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\sum_{u=1}^{n} \mathbb{E}\left[\left(\Delta_u^\top X_0\right)^2\right] &= \mathrm{tr}(\mathbb{E}[X_0 X_0^\top] \Delta^\top \Delta) \\
&\geq \lambda_{\min}(\mathbb{E}[X_0 X_0^\top]) ||\Delta||_F^2.
\end{aligned}$$

It follows that $\mathbb{E}^0[h(\Delta; X)] \geq \kappa_1 ||\Delta||_2^2$, for every $\kappa_1 > 0$ such that

$$\kappa_1 \leq m\mathbb{E}^0[\tau] \lambda_{\min}(\mathbb{E}[X_0 X_0^\top]).$$

1.5.11 Proof of Lemma 1.4.4

We first show a concentration bound for random variable $h(\Delta; X)$ in the following lemma.

Lemma 1.5.2. *For a voter model process with parameter A^*, for any $\delta \in (0, 1/2]$ and $\Delta \geq 0$, with probability at least $1 - \delta$,*

$$|h(\Delta; X) - \mathbb{E}[h(\Delta; X)]| \leq \epsilon \|\Delta\|_F^2$$

where

$$\epsilon = 2\sqrt{2}s\sqrt{m\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{2}{\delta}\right)\right)\frac{1}{\Phi_{A^*}}}\sqrt{\log\left(\frac{2}{\delta}\right)}.$$

Proof. Recall that

$$h(\Delta; X) = \sum_{i=1}^{m}\sum_{t=0}^{\tau_i - 1}\sum_{u=1}^{n}(\Delta_u^\top X_t^{(i)})^2.$$

Let us define, if $0 \leq s < \sum_{i=1}^{m}\tau_i$,

$$X_s = X^{(k)}_{s - \sum_{i=1}^{k-1}\tau_i} \text{ for } \sum_{i=1}^{k-1}\tau_i \leq s < \sum_{i=1}^{k}\tau_i \text{ and } k \in [m]$$

where $\sum_{i=1}^{0}\tau_i = 0$ and, otherwise, if $s > \sum_{i=1}^{m}\tau_i$,

$$X_s = 0.$$

Now, we can write

$$h(\Delta; X) = \sum_{t=0}^{\sum_{i=1}^{m}\tau_i - 1}\sum_{u=1}^{n}(\Delta_u^\top X_t)^2.$$

Let $Y_s = \sum_{u=1}^{n} (\Delta_u^\top X_s)^2$. For any $T > 0$, we have

$$\begin{aligned}
& \mathbb{P}^0\left[|h(\Delta;X) - \mathbb{E}[h(\Delta;X)]| \geq \epsilon\|\Delta\|_F^2\right] \\
\leq\ & \mathbb{P}^0\left[\left|\sum_{t=0}^{T-1}\sum_{u=1}^{n}(\Delta_u^\top X_t)^2 - \mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{u=1}^{n}(\Delta_u^\top X_t)^2\right]\right| \geq \epsilon\|\Delta\|_F^2\right] + \mathbb{P}^0\left[\sum_{i=1}^{m}\tau_i > T\right] \\
=\ & \mathbb{P}^0\left[\left|\sum_{s=0}^{T-1} Y_s - \mathbb{E}^0\left[\sum_{s=0}^{T-1} Y_s\right]\right| \geq \epsilon\|\Delta\|_F^2\right] + \mathbb{P}^0\left[\sum_{i=1}^{m}\tau_i > T\right] \\
\leq\ & \mathbb{P}^0\left[\sum_{s=0}^{T-1}|Y_s - \mathbb{E}\left[Y_s\right]| \geq \epsilon\|\Delta\|_F^2\right] + \mathbb{P}^0\left[\sum_{i=1}^{m}\tau_i > T\right]
\end{aligned}$$

where the last inequality follows from the following basic relations

$$\left|\sum_{s=1}^{T} Y_s - \mathbb{E}\left[\sum_{s=1}^{T} Y_s\right]\right| = \left|\sum_{s=1}^{T}(Y_s - \mathbb{E}\left[Y_s\right])\right| \leq \sum_{s=1}^{T}|Y_s - \mathbb{E}\left[Y_s\right]|.$$

Let us define

$$G_t = \sum_{s=1}^{t}|Y_s - \mathbb{E}\left[Y_s\right]| \text{ and } \delta G_t = G_t - G_{t-1}.$$

We have $\mathbb{E}[G_t \mid \mathcal{F}_{t-1}] = G_{t-1} + \mathbb{E}\left[|Y_t - \mathbb{E}[Y_t]| \mid \mathcal{F}_{t-1}\right] \geq G_{t-1}$ for all $t \in \{1,\ldots,T\}$. Hence, $G_0, G_1, \ldots, G_T$ is a super-martingale sequence. Moreover, we have

$$\begin{aligned}
Y_s &= \sum_{u=1}^{n}(\Delta_u^\top X_s)^2 \\
&\leq \sum_{u=1}^{n}\|\Delta_u\|_1^2 \\
&\leq \|\Delta\|_{1,1}^2 \\
&\leq s\|\Delta\|_F^2
\end{aligned}$$

where the first inequality follows from $\Delta_{u,i}X_{s,i} \leq |\Delta_{u,i}|$ for all $i = 1,\ldots,n$, the second inequality follows from the fact $\sum_{u=1}^{n}\|\Delta_u\|_1^2 \leq \left(\sum_{u=1}^{n}\|\Delta_u\|_1\right)^2 = \|\Delta\|_{1,1}^2$ and the last inequality follows from the assumption that Δ has sparsity s, and, hence, $\|\Delta\|_{1,1} \leq \sqrt{s}\|\Delta\|_{2,2}$.

It follows that $|\delta G_t| = |Y_t - \mathbb{E}[Y_t]| \leq 2s\|\Delta\|_F^2$ for all $t \in \{1,\ldots,T\}$, almost surely.

By applying Azuma-Hoeffding's inequality, we obtain

$$\mathbb{P}\left[G_T - G_0 \geq \epsilon\right] = \mathbb{P}\left[\sum_{s=1}^{T} |Y_s - \mathbb{E}\left[Y_s\right]| \geq \epsilon \|\Delta\|_F^2\right] \leq e^{-\frac{\epsilon^2}{8Ts^2}}.$$

Hence, $\mathbb{P}\left[G_T - G_0 \geq \epsilon\right] \leq \delta/2$, for

$$\epsilon \geq 2\sqrt{2}\sqrt{T}s\sqrt{\log\left(\frac{2}{\delta}\right)}.$$

From Theorem 1.3.2, we also have $\mathbb{P}\left[\sum_{i=1}^{m} \tau_i > T\right] \leq \delta/2$, for any

$$T \geq m\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{2}{\delta}\right)\right)\frac{1}{\Phi_{A^*}}.$$

Hence, it follows that with probability at least $1-\delta$,

$$|h(\Delta; X) - \mathbb{E}[h(\Delta; X)]| \leq \epsilon\|\Delta\|_F^2$$

where

$$\epsilon = 2\sqrt{2}s\sqrt{m\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{2}{\delta}\right)\right)\frac{1}{\Phi_{A^*}}}\sqrt{\log\left(\frac{2}{\delta}\right)}.$$

□

From Lemma 1.5.2, it follows that with probability at least $1-\delta$, $h(\Delta; X) \geq \mathbb{E}[h(\Delta; X)] - \epsilon||\Delta||_F^2$. Combining with Lemma 1.4.3, with probability at least $1-\delta$, $h(\Delta; X) \geq (\kappa_1 - \epsilon)||\Delta||_F^2$. Hence, with probability at least $1-\delta$, $h(\Delta; X) \geq (\kappa_1/2)||\Delta||_F^2$, provided that $\kappa_1/2 \geq \epsilon$, which is equivalent to

$$m \geq \frac{s^2}{\Phi_{A^*}}\frac{1}{\mathbb{E}^0[\tau]^2\lambda_{\min}(\mathbb{E}[X_0X_0^\top])^2}c_{\delta,\pi^*}(m).$$

where

$$c_{\delta,\pi^*}(m) = 8\left(\log\left(\frac{1}{2\pi^*}\right) + \frac{1}{m}\log\left(\frac{2}{\delta}\right)\right)\log\left(\frac{2}{\delta}\right).$$

1.5.12 Proof of Lemma 1.4.5

The proof of the lemma is based on using the concepts of covering and metric entropy which we discussed in Appendix 1.5.1.3. With a slight abuse of notation, in the proof, we assume

$$X = (X_0^{(1)}, \ldots, X_{\tau_1-1}^{(1)}, \ldots, X_0^{(m)}, \ldots, X_{\tau_m-1}^{(m)})^\top.$$

The proof follows similar steps as that of Lemma A.6 in Pandit et al. [2019]. The main difference in our case is that X does not have fixed dimensions, which requires additional technical steps.

We separately consider three different cases: (a) $||\Delta||_F = r$, (b) $||\Delta||_F > r$, and (c) $||\Delta||_F < r$, for value of r defined as

$$r = \frac{1}{\sqrt{s}}||A^*_{S^{*c}}||_{1,1} + \mathbb{I}_{\{||A^*_{S^{*c}}||_{1,1}=0\}}. \tag{1.5.12}$$

Case 1: $||\Delta||_F = r$ For every $\Delta \in \mathbb{C}^*$, we have

$$\begin{aligned} ||\Delta||_{1,1} &= ||\Delta_{S^*}||_{1,1} + ||\Delta_{S^{*c}}||_{1,1} \\ &\leq 4||\Delta_{S^*}||_{1,1} + 4||A^*_{S^{*c}}||_{1,1} \\ &\leq 4(\sqrt{s}||\Delta||_F + ||A^*_{S^{*c}}||_{1,1}). \end{aligned}$$

Hence, for any $r > 0$, $\mathbb{C}^* \cap \partial B_2(r) \subseteq B_1(r')$ where

$$r' := 4(r\sqrt{s} + ||A^*_{S^{*c}}||_{1,1}). \tag{1.5.13}$$

Note that under (1.5.12), $4\sqrt{s} \leq r'/r \leq 8\sqrt{s}$.

By the triangle inequality, for all $\Delta, \Delta' \in \mathbb{R}^{n\times n}$,

$$\left|\sqrt{h(\Delta; X)} - \sqrt{h(\Delta'; X)}\right| \leq \sqrt{T}\rho(\Delta, \Delta')$$

where

$$\rho(\Delta, \Delta') = \frac{1}{\sqrt{T}}||X(\Delta - \Delta')^\top||_F.$$

Because it holds $(a-b)^2 \geq a^2/2 - b^2$ for all $a, b \geq 0$, it follows that for all $\Delta, \Delta' \in \mathbb{R}^{n\times n}$,

$$h(\Delta; X) \geq \frac{1}{2} h(\Delta'; X) - T\rho(\Delta, \Delta')^2.$$

Let N be an $r\epsilon$-cover of $\mathbb{C}^* \cap \partial B_2(r)$. Fix an arbitrary $\Delta \in \mathbb{C}^* \cap \partial B_2(r)$, and let Δ'' be such that $\Delta'' \in N$ and $\rho(\Delta, \Delta'') \leq r\epsilon$. Then, note

$$h(\Delta; X) \geq \frac{1}{2} h(\Delta''; X) - T(r\epsilon)^2 \geq \frac{1}{2} \min_{\Delta' \in N} h(\Delta'; X) - T(r\epsilon)^2.$$

It follows that

$$\inf_{\Delta \in \mathbb{C}^* \cap \partial B_2(r)} h(\Delta; X) \geq \frac{1}{2} \min_{\Delta \in N} h(\Delta; X) - T(r\epsilon)^2. \tag{1.5.14}$$

Let $\epsilon^2 = (\kappa_1/8)/T$. It then follows

$$\mathbb{P}^0 \left[\inf_{\Delta \in \mathbb{C}^* \cap \partial B_2(r)} h(\Delta; X) \leq \frac{1}{8}\kappa_1 r^2 \right] \leq \mathbb{P}^0 \left[\min_{\Delta \in N} h(\Delta; X) \leq \frac{1}{2}\kappa_1 r^2 \right]. \tag{1.5.15}$$

For any matrix X with the number of rows T such that $T \leq T^*$, for some fixed T^*, we can bound $|N|$ by a fixed value N^*, which we show next. Recall that for any $r > 0$, $\mathbb{C}^* \cap \partial B_2(r) \subseteq B_1(r')$, where r' is defined in (1.5.13). Hence, we can bound $|N|$, the covering number of $\mathbb{C}^* \cap \partial B_2(r)$, by the covering number of $B_1(r')$. From Lemma 1.5.1, $r'/r \leq 8\sqrt{s}$ and $\epsilon^2 = (\kappa_1/8)/T$, we have

$$|N| \leq e^{c_1 \left(\frac{r'}{r\epsilon}\right)^2 \log(n)} \leq e^{8^3 s c_1 T \frac{1}{\kappa_1} \log(n)},$$

where $c_1 = c\sigma_{\max}(X)^2/T$ for some constant $c > 0$, under condition $r\epsilon \leq c_2 r'$ with $c_2 = \sigma_{\max}(X)/\sqrt{T}$. Since $\sigma_{\max}(X)^2 \leq nT$ and $r'/r \geq 4\sqrt{s}$, we have

$$|N| \leq N^* = e^{c8^3 \frac{T^*}{\kappa_1} sn \log(n)} \tag{1.5.16}$$

under conditions

$$C_1 = \{T \leq T^*\} \text{ and } C_2 = \left\{ \frac{\sigma_{\max}(X)^2}{\kappa_1} \geq \frac{1}{128s} \right\}.$$

From (1.5.15), we have

$$\begin{aligned}
& \mathbb{P}^0\left[\inf_{\Delta\in\mathbb{C}^*\cap\partial B_2(r)} h(\Delta;X) \le \frac{1}{8}\kappa_1 r^2\right] \\
\le\ & \mathbb{E}^0\left[\mathbb{I}_{\cup_{\Delta\in N}\{h(\Delta;X)\le(\kappa_1/2)r^2\}}\mathbb{I}_{C_1}\mathbb{I}_{C_2}\right] + \mathbb{E}^0[\mathbb{I}_{C_1^c}] + \mathbb{E}^0[\mathbb{I}_{C_2^c}] \\
\le\ & N^* \max_{\Delta:||\Delta||_F=r} \mathbb{P}^0\left[h(\Delta;X) \le \frac{1}{2}\kappa_1 r^2\right] + \mathbb{P}^0[C_1^c] + \mathbb{P}^0[C_2^c].
\end{aligned}$$

The probability of the event C_1^c can be bounded as follows. By Corollary 1.3.4, for any $\delta' \in (0,1]$, $\Pr[C_1^c] = \Pr[T > T^*] \le \delta'$, when

$$T^* \ge m\left(\log\left(\frac{1}{2\pi^*}\right) + \log\left(\frac{1}{\delta'}\right)\right)\frac{1}{\Phi_{A^*}}. \tag{1.5.17}$$

We next upper bound the probability of the event C_2^c. First, note

$$\sigma_{\max}(X)^2 = \lambda_{\max}(X^\top X) \ge \frac{||X^\top X||_{1,1}}{n}$$

where the last inequality holds by the basic fact that for any real $n \times n$ matrix M,

$$\lambda_{\max}(M) \ge \frac{\mathbf{1}^\top M \mathbf{1}}{\mathbf{1}^\top\mathbf{1}} = \frac{\sum_{i=1}^n\sum_{j=1}^n M_{i,j}}{n}.$$

Hence, we have

$$\mathbb{P}^0[C_2^c] \le \mathbb{P}^0\left[||X^\top X||_{1,1} < \frac{n}{128s}\kappa_1\right]. \tag{1.5.18}$$

Note that

$$||X^\top X||_{1,1} = \sum_{i=1}^m\sum_{t=0}^{\tau_i-1} \mathbf{1}^\top X_t^{(i)} {X_t^{(i)}}^\top \mathbf{1}.$$

By the Palm inversion formula and definition of κ_1, we have

$$\begin{aligned}
\mathbb{E}^0[||X^\top X||_{1,1}] &= m\mathbb{E}^0\left[\sum_{t=0}^{\tau-1} \mathbf{1}^\top X_t {X_t}^\top \mathbf{1}\right] \\
&= m\mathbb{E}^0[\tau]\mathbf{1}^\top\mathbb{E}[X_0X_0^\top]\mathbf{1} \\
&\ge m\mathbb{E}^0[\tau]\lambda_{\min}(\mathbb{E}[X_0X_0^\top])n \\
&= n\kappa_1.
\end{aligned}$$

To bound (1.5.18), we use Chebyshev's inequality: for any random variable X with expected value μ and variance σ^2, $\Pr[|X-\mu| \geq k\sigma] \leq 1/k^2$, for any $k > 0$. By Chebyshev's inequality, we have

$$\mathbb{P}^0[||X^\top X||_{1,1} < x] \leq \frac{\sigma^2}{(\mu - x)^2}$$

for every $0 \leq x < \mu$, where μ and σ^2 are the expected value and variance of $||X^\top X||_{1,1}$, respectively. Let σ_1^2 be the variance of random variable $Y = \sum_{t=0}^{\tau-1} \mathbf{1}^\top X_t {X_t}^\top \mathbf{1}$ and $c_s = 1/(1-1/(128s))^2$. It follows

$$\begin{aligned}\mathbb{P}^0\left[||X^\top X||_{1,1} < \frac{n}{128s}\kappa_1\right] &\leq c_s \frac{m\sigma_1^2}{n^2\kappa_1^2} \\ &= c_s \frac{\sigma_1^2}{n^2 m \mathbb{E}^0[\tau]^2 \lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^2}.\end{aligned}$$

Next, we bound σ_1^2 as follows

$$\sigma_1^2 \leq \mathbb{E}^0[Y^2] \leq n^4 \mathbb{E}^0[\tau^2] \leq \frac{2}{e^2} n^4 \frac{1}{\Phi_{A^*}^2} \frac{1}{\pi^*}$$

where the last inequality is by Corollary 1.3.3.

Putting the pieces together, we have $\Pr[C_2^c] \leq 1/n$, under condition

$$m \geq c_s n^3 (1/\pi^*) \frac{1}{(\Phi_{A^*}\mathbb{E}^0[\tau])^2 \lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^2}. \tag{1.5.19}$$

By Lemma 1.4.4, for any $\delta \in (0,1]$ and Δ, $\Pr[h(\Delta; X) \leq (\kappa_1/2)||\Delta||_F^2] \leq \delta$ provided that

$$m \geq 8s^2 \frac{\log(1/(2\pi^*)) + \log(2/\delta)}{\Phi_{A^*}\mathbb{E}^0[\tau]^2 \lambda_{\min}(\mathbb{E}[X_0 X_0^\top])^2}.$$

Let T^* be defined by equality in (1.5.17) and $\delta' = 1/n$. From (1.5.16), we have $N^* = n^a$ where

$$a = snc8^3 \frac{\log(1/(2\pi^*)) + \log(n)}{(\Phi_{A^*}\mathbb{E}^0[\tau])\lambda_{\min}(\mathbb{E}[X_0 X_0^\top])}.$$

Take $\delta = 2/n^{a+1}$. Then, we have

$$N^* \max_{\Delta:||\Delta||_F = r} \mathbb{P}^0\left[h(\Delta; X) \leq \frac{1}{2}\kappa_1 r^2\right] \leq \frac{2}{n}$$

provided that

$$m \geq 8s^2 \frac{\log(1/(2\pi^*))(a+1)\log(n) + (a+1)^2\log(n)^2}{\Phi_{A^*}\mathbb{E}^0[\tau]^2\lambda_{\min}(\mathbb{E}[X_0X_0^\top])^2}. \tag{1.5.20}$$

We have shown that the RSC condition holds with at least probability $1-4/n$, for all $\Delta \in \mathbb{C}^* \cap \partial B_2(r)$, with curvature $\kappa_1/8$ and tolerance 0, under conditions (1.5.19) and (1.5.20). Condition (1.5.20) is stronger than condition (1.5.19) when

$$\mathbb{E}^0[\tau] = \tilde{O}\left(\frac{s^4\pi^*}{n}\frac{1}{(\Phi_{A^*}\mathbb{E}^0[\tau])\lambda_{\min}(\mathbb{E}[X_0X_0^\top])}\right).$$

Case 2: $||\Delta||_F > r$ Let $t = ||\Delta||_F/r$. Assume that the RSC condition holds for every $\Delta' \in \mathbb{C}^* \cap \partial B_2(r)$ with curvature $\kappa_1/8$ and tolerance 0. Note that, for $\Delta \in \mathbb{C}^* \cap B_2(r)$,

$$h(\Delta; X) = t^2 h(\Delta/t; X) \geq t^2 \frac{1}{8}\kappa_1 r^2 = \frac{1}{8}\kappa_1 ||\Delta||_F^2.$$

Hence, the RSC condition holds on $\mathbb{C}^* \cap B_2(r)$ with curvature $\kappa_1/8$ and tolerance 0.

Case 3: $||\Delta||_F < r$ Let $\gamma'^2 = (\kappa_1/8)r^2 = (\kappa_1/8)||A^*_{S^*}||^2_{1,1}/s$. In this case, we have

$$h(\Delta; X) \geq 0 \geq (\kappa_1/8)||\Delta||_F^2 - \gamma'^2.$$

Hence, the RSC condition holds with curvature $\kappa_1/8$ and tolerance $(\kappa_1/8)||A^*_{S^*}||^2_{1,1}/s$.

1.5.13 Proof of Theorem 1.4.2

The proof follows from Theorem 1.4.1 and Lemmas 1.4.1, 1.4.2 and 1.4.5, which we show as follows.

From Lemma 1.4.1, with probability at least $1 - 1/n$,

$$2||\nabla\ell(A^*)||_\infty \leq \lambda_m = 2\sqrt{2}\frac{1}{\alpha}\frac{1}{\sqrt{\Phi_{A^*}}}c_{n,\pi^*}\sqrt{m}.$$

From Lemma 1.4.2 and Lemma 1.4.5, the negative log-likelihood function satisfies the RSC condition relative to A^* and S that is the support of A^* with curvature $\kappa' = \kappa_1/8$, and tolerance $\gamma'^2 = 0$ with probability at least $1 - 4/n$, under conditions (1.4.10) and (1.4.11).

Recall that $\kappa_1 = m\mathbb{E}^0[\tau]\lambda_{\min}(\mathbb{E}[X_0X_0^\top])$. It follows that

$$\frac{\lambda_m}{\kappa'} = \max\left\{\frac{8}{\mathbb{E}^0[\tau]\lambda_{\min}(\mathbb{E}[X_0X_0^\top])}, 1\right\} 2\sqrt{2}\frac{1}{\alpha}\frac{1}{\sqrt{\Phi_{A^*}}}c_{n,\pi^*}\frac{1}{\sqrt{m}}.$$

Combining the above facts with Theorem 1.4.1, with probability at least $1 - 5/n$,

$$||\hat{A} - A^*||_F^2 \leq 4608\frac{sc_{n,\pi^*}^2}{\alpha^2(\Phi_{A^*}\mathbb{E}^0[\tau])^2\lambda_{\min}(\mathbb{E}[X_0X_0^\top])^2}\Phi_{A^*}\frac{1}{m}$$

which completes the proof.

1.5.14 Stationary correlation matrices

We consider the stationary correlation matrix M of the extended voter model (1.1.3) defined by

$$M = \mathbb{E}[X_0X_0^\top].$$

The stationary correlation matrix exists which follows from the Palm inversion formula (1.2.2) as $\mathbb{E}^0[\tau] < \infty$.

In this section we present analysis for the extended voter model that includes final consensus states of individual voter model processes. The stationary correlation matrix of this process, denoted as M, is related to the stationary correlation matrix, M', for the extended voter model that does not include final consensus states of individual voter model processes as follows

$$M = \frac{\mathbb{E}^0[\tau]}{\mathbb{E}^0[\tau]+1}M' + \frac{\pi^\top\mathbb{E}^0[X_0]}{\mathbb{E}^0[\tau]}\mathbf{1}\mathbf{1}^\top. \tag{1.5.21}$$

This follows by the Palm inversion formula and the fact $\mathbb{P}^0[X_\tau = \mathbf{1}] = \pi^\top\mathbb{E}^0[X_0]$. Indeed,

$$\begin{aligned}
(\mathbb{E}^0[\tau]+1)M &= \mathbb{E}^0\left[\sum_{t=0}^{\tau} X_tX_t^\top\right] \\
&= \mathbb{E}^0\left[\sum_{t=0}^{\tau-1} X_tX_t^\top\right] + \mathbb{E}^0[X_\tau X_\tau^\top] \\
&= \mathbb{E}^0[\tau]M' + \mathbb{P}^0[X_\tau = \mathbf{1}]\mathbf{1}\mathbf{1}^\top \\
&= \mathbb{E}^0[\tau]M' + \pi^\top\mathbb{E}^0[X_0]\mathbf{1}\mathbf{1}^\top.
\end{aligned}$$

By Weyl's inequalities and the fact that eigenvalues of $\mathbf{1}\mathbf{1}^\top$ are either of value 0 or n, the eigenvalues of M and M' are related as follows

$$\frac{\mathbb{E}^0[\tau]}{\mathbb{E}^0[\tau]+1}\lambda_i(M') \leq \lambda_i(M) \leq \frac{\mathbb{E}^0[\tau]}{\mathbb{E}^0[\tau]+1}\lambda_i(M') + \frac{\pi^\top \mathbb{E}^0[X_0]}{\mathbb{E}^0[\tau]}n.$$

1.5.14.1 Lyapunov matrix equation

The stationary correlation matrix satisfies the Lyapunov matrix equation stated in the following lemma. The Lyapunov matrix equation plays an important role for stability of linear dynamical systems Gajic and Qureshi [1995], Barnett and Storey [1970].

Lemma 1.5.3. *For the extended voter model with parameter A and initial state distribution μ such that the process is ergodic, the following Lyapunov matrix equation holds*

$$M = AMA^\top + Q \tag{1.5.22}$$

where

$$Q = \mathbb{E}[D(V(X_0))] + \frac{\mathbb{E}^0[X_0X_0^\top] - (\mu(C_1) + (1-\mu(C))\pi^\top \mathbb{E}^0[X_0 \mid X_0 \notin C])\mathbf{1}\mathbf{1}^\top}{(1-\mu(C))\mathbb{E}^0[\tau]+1}$$

and $D(V(x))$ is the diagonal matrix with diagonal elements $V_{a_u}(x)$.

Proof. The following equations hold

$$\begin{aligned}\mathbb{E}[X_{t+1}X_{t+1}^\top] &= \mathbb{E}[X_{t+1}X_{t+1}^\top \mathbb{I}_{\{X_t \notin C\}}] \\ &\quad + \mathbb{E}^0[X_0X_0^\top]\mathbb{P}[X_t \in C]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X_{t+1}X_{t+1}^\top \mathbb{I}_{\{X_t \notin C\}}] &= \mathbb{E}[Z_{t+1}X_tX_t^\top Z_{t+1}^\top] \\ &\quad - \mathbb{E}[Z_{t+1}X_tX_t^\top Z_{t+1}^\top \mathbb{I}_{\{X_t \in C\}}]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Z_{t+1}X_tX_t^\top Z_{t+1}^\top] &= \mathbb{E}[\mathbb{E}[Z_{t+1}X_t(Z_{t+1}X_t)^\top \mid X_t]] \\ &= A\mathbb{E}[X_tX_t^\top]A^\top + \mathbb{E}[D(V(X_t))]\end{aligned}$$

and

$$\mathbb{E}[Z_{t+1}X_tX_t^\top Z_{t+1}^\top \mathbb{I}_{\{X_t\in C\}}] = \mathbf{1}\mathbf{1}^\top \mathbb{P}[X_t \in C_1].$$

Putting the pieces together, we have

$$\begin{aligned}\mathbb{E}[X_0X_0^\top] &= A\mathbb{E}[X_0X_0^\top]A^\top + \mathbb{E}[D(V(X_0))] \\ &\quad +(\mathbb{E}^0[X_0X_0^\top] - p_1\mathbf{1}\mathbf{1}^\top)\mathbb{P}[X_0 \in C]\end{aligned}$$

where $p_1 := \mathbb{P}[X_0 \in C_1 \mid X_0 \in C]$, from which the statement of the lemma follows. □

By multiplying both sides in equation (1.5.22) with $\pi^\top$ and π from left and right respectively, we obtain the following corollary.

Corollary 1.5.1. *The following equation holds*

$$\begin{aligned}&((1-\mu(C))\mathbb{E}^0[\tau]+1)\sum_{u=1}^{n}\pi_u^2\mathbb{E}[V_{a_u}(X_0)] \\ =\ &\mathbb{E}^0[V_\pi(X_0)] + \mu(C_1) - \mu(C)\mathbb{E}^0[\pi^\top X_0].\end{aligned}$$

If $\mu(C) = 0$, then the last expression boils down to

$$(\mathbb{E}^0[\tau]+1)\sum_{u=1}^{n}\pi_u^2\mathbb{E}[V_{a_u}(X_0)] = \mathbb{E}^0[V_\pi(X_0)]$$

which asserted in (1.3.1).

The necessary and sufficient condition for the Lyapunov matrix equation (1.5.22) to have a unique solution M for any positive semi-definite matrix Q is that no two eigenvalues of A have product equal

to 1, i.e. $\lambda_i(A)\lambda_j(A) \neq 1$ for all $i, j = 1, \ldots, n$. Furthermore, it is known that $\rho(A) < 1$, where $\rho(A)$ is the spectral radius of A, holds if and only if for any positive definite Q, (1.5.22) has a a positive definite solution M.

The Lyapunov matrix equation of the voter model is such that Q is a positive semi-definite matrix as stated in the following lemma.

Lemma 1.5.4. *Q in (1.5.22) is a positive semi-definite matrix, with eigenvalue 0 associated with eigenvector π.*

Proof. Multiply both sides of equation (1.5.22) with $\pi^\top$ from the left and π from the right. Note that $\pi^\top AMA^\top \pi = \pi^\top M\pi$. It follows that $\pi^\top Q\pi = 0$, which shows that π is an eigenvector of Q with eigenvalue 0. □

1.5.14.2 Product-form Bernoulli initial state distribution

We consider the spectrum of matrix Q for initial state distribution μ that has product-form with Bernoulli (p) marginal distributions, with $0 < p < 1$. Note that

$$\mathbb{E}^0[X_0 \mid X_0 \notin C] = \frac{1-(1-p)^{n-1}}{1-p^n-(1-p)^n} p\mathbf{1}.$$

Under the given assumptions on distribution μ, Q is an $n \times n$ off-diagonal constant matrix with $q_{u,u} = \mathbb{E}[V_{a_u}(X_0)]$ and $q_{u,v} = -\alpha$, for $u \neq v$, where

$$\alpha := \frac{p(1-p)}{(1-p^n-(1-p)^n)\mathbb{E}^0[\tau]+1}. \tag{1.5.23}$$

In order to localize eigenvalues of Q, we will use the following lemma.

Lemma 1.5.5 (Gendreau [1986])**.** *Let S be an $n \times n$ off-diagonal constant matrix such that $s_{i,j} = d_i + \alpha$, for $i = j$ and $s_{i,j} = \alpha$, for $i \neq j$, where $d_1, d_2, \ldots, d_n$ and α are given real numbers with $\alpha \geq 0$. Let $e_1 < \cdots < e_m$ be distinct values in $\{d_1, \ldots, d_n\}$, and n_i be the number of occurrences of e_i. Then S has*

1. *one eigenvalue in (e_i, e_{i+1}) for $i = 1, \ldots, m-1$ and one eigenvalue in (e_m, ∞), all with multiplicity 1;*
2. *each e_i such that $n_i > 1$ is an eigenvalue of multiplicity $n_i - 1$.*

By Lemma 1.5.4 and Lemma 1.5.5, matrix Q in (1.5.22) has eigenvalue 0 and all other eigenvalues larger than or equal to $\min_u \mathbb{E}[V_{a_u}(X_0)] + \alpha > 0$. Hence, the smallest positive eigenvalue $\lambda_2(Q)$ of Q satisfies

$$\lambda_2(Q) \geq \min_{u \in V} \mathbb{E}[V_{a_u}(X_0)] + \alpha > 0. \tag{1.5.24}$$

1.5.14.3 Bounding smallest eigenvalue of the stationary correlation matrix

Let $R(S, x)$ denote the Rayleigh quotient, $R(S, x) = (x^\top S x)/x^\top x$, for some matrix S. Note that

$$R(M; \pi) = \frac{\mathbb{E}[(\pi^\top X_0)^2]}{||\pi||^2}.$$

Note that

$$\lambda_2(Q) = \min_{x: x \neq 0, \pi^\top x = 0} R(Q, x).$$

We claim that

$$\lambda_1(M) \geq \min\{R(M; \pi), \lambda_2(Q)\}.$$

To show this, we decompose any vector $x \in \mathbb{R}^n$ into orthogonal components $x = \gamma\pi + z$ for some $\gamma \in \mathbb{R}$ and $z \in \mathbb{R}^n$ such that $\pi^\top z = 0$. We have the following relations

$$\begin{aligned}
R(M, x) &= \frac{x^\top M x}{x^\top x} \\
&= \frac{(\gamma\pi + z)^\top M(\gamma\pi + z)}{(\gamma\pi + z)^\top(\gamma\pi + z)} \\
&= \frac{(\gamma\pi + z)^\top M(\gamma\pi + z)}{\gamma^2 \pi^\top \pi + z^\top z} \\
&\geq \frac{\gamma^2 \pi^\top M \pi + z^\top M z}{\gamma^2 \pi^\top \pi + z^\top z} \\
&\geq \frac{\gamma^2 \pi^\top \pi R(M; \pi) + z^\top z \lambda_2(Q)}{\alpha^2 \pi^\top \pi + z^\top z} \\
&\geq \min\{R(M; \pi), \lambda_2(Q)\}.
\end{aligned}$$

From this it follows that the smallest eigenvalue $\lambda_1(M)$ of M statisfies $\lambda_1(M) = \min_{x: x \neq 0} R(M, x) \geq \min\{R(M; \pi), \lambda_2(Q)\}$.

It readily follows from (1.1.3) that the Rayleigh quotient $R(M;\pi)$ can be lower bounded as follows

$$R(M;\pi) \geq \frac{1}{(1-\mu(C))\mathbb{E}[\tau]+1}\frac{\mathbb{E}^0[(\pi^\top X_0)^2]}{||\pi||^2}.$$

For the case when μ is a product-form distribution with Bernoulli (p) marginal distributions, we have

$$\mathbb{E}^0[(\pi^\top X_0)^2] = p^2 + p(1-p)||\pi||^2.$$

Hence, $R(M;\pi) \geq p(1-p+p/||\pi||^2)/(\mathbb{E}^0[\tau]+1)$. By combining with (1.5.24), we have

$$\lambda_1(M) \geq \min\left\{\frac{p\left(1-p+p\frac{1}{||\pi||^2}\right)}{\mathbb{E}^0[\tau]+1}, \min_{u\in V}\mathbb{E}[V_{a_u}(X_0)] + \frac{p(1-p)}{\mathbb{E}^0[\tau]+1}\right\}.$$

1.5.14.4 Complete graph example

We consider the case when $a_{u,u} = 0$ and $a_{u,v} = 1/(n-1)$ for all $u \neq v$, and when the initial state distribution μ is the product-form with Bernoulli (p) marginal distributions. Because of the symmetry, $m_{u,u} = a$ and $m_{u,v} = b$, for all $u \neq v$, for some a and b. Note

$$\begin{aligned}(AMA^\top)_{u,v} &= \sum_{i=1}^n\sum_{j=1}^n a_{u,i}a_{v,j}m_{i,j}\\ &= a\sum_{i=1}^n a_{u,i}a_{v,i} + b\sum_{i=1}^n\sum_{j\neq i} a_{u,i}a_{v,j}\\ &= (a-b)\sum_{i=1}^n a_{u,i}a_{v,i} + b\sum_{i=1}^n\sum_{j=1}^n a_{u,i}a_{v,j}.\end{aligned}$$

It follows

$$(AMA^\top)_{u,v} = (a-b)\left(\frac{1}{n-1}\mathbb{I}_{\{u=v\}} + \frac{n-2}{(n-1)^2}\mathbb{I}_{\{u\neq v\}}\right) + b.$$

Hence, we have

$$(M - AMA^\top)_{u,v} = \begin{cases}(a-b)\left(1-\frac{1}{n-1}\right) & \text{if } u = v\\ -(a-b)\frac{n-2}{(n-1)^2} & \text{if } u \neq v.\end{cases}$$

Now, we have $M - AMA^\top = Q$, and Q has diagonal elements of value $\mathbb{E}[V_{a_1}(X_0)]$ and off-diagonal elements of value $-\alpha$, where α is given in (1.5.23). It follows that

$$a - b = \frac{(n-1)^2}{n-2} \frac{1}{(1-p^n-(1-p)^n)\mathbb{E}^0[\tau]+1} p(1-p)$$

and

$$\mathbb{E}[V_{a_1}(X_0)] = \left(1 - \frac{1}{n-1}\right)(a-b).$$

Note that

$$a - b = \frac{np(1-p)}{\mathbb{E}^0[\tau]}(1+o(1))$$

and

$$\mathbb{E}[V_{a_1}(X_0)] = p(1-p)\frac{n}{\mathbb{E}^0[\tau]+1}(1+o(1)).$$

From Lemma 1.5.5, it follows that $\lambda_1(M) = a - b$. Hence,

$$\lambda_1(M) = p(1-p)\frac{n}{\mathbb{E}^0[\tau]+1}(1+o(1)).$$

The smallest eigenvalue of the correlation matrix K with respect to the stationary distribution of an extended voter process that includes final consensus states of individual voter model processes, $\lambda_1(K)$, such that K has constant diagonal elements and constant non-diagonal elements, $\lambda_1(K)$, is related to the smallest eigenvalue of the correlation matrix K' with respect to stationary distribution of the extended voter process that does not include final consensus states, $\lambda_1(K')$, as follows:

$$\lambda_1(K) = \frac{\mathbb{E}^0[\tau]}{\mathbb{E}^0[\tau]+1}\lambda_1(K'). \tag{1.5.25}$$

This easily follows from (1.5.21) and the fact that the smallest eigenvalue of a matrix with constant diagonal elements (say equal α) and constant non-diagonal elements (say equal to β) is equal to $\alpha - \beta$.

For the complete graph case considered in this section, from (1.5.25) and $\mathbb{E}^0[\tau]/(\mathbb{E}^0[\tau]+1) = 1+o(1)$, the correlation matrix M' with respect to stationary distribution of the extended voter process that does not include final consensus states of individual voter model processes is

$$\lambda_1(M') = p(1-p)\frac{n}{\mathbb{E}^0[\tau]+1}(1+o(1)).$$

1.5.15 Linear ϵ-noisy voter model

In this section we consider the linear ϵ-voter model defined as the linear discrete-time dynamical system (1.1.6). Let $A^\epsilon := (1-2\epsilon)A$.

Lemma 1.5.6. *For any linear ϵ-voter model with parameters A and $0 < \epsilon \leq 1/2$, the expected values of vertex states satisfy*

$$\mathbb{E}[X_0] = \frac{1}{2}\mathbf{1}.$$

Proof. From (1.1.6), we have

$$\mathbb{E}[X_{t+1}] = \mathbb{E}[Q_{t+1}Z_{t+1}X_t] + \mathbb{E}[R_t] = (1-2\epsilon)A\mathbb{E}[X_t] + \epsilon\mathbf{1}.$$

Hence,

$$(I - (1-2\epsilon)A)\mathbb{E}[X_0] = \epsilon\mathbf{1}.$$

Since $\rho((1-2\epsilon)A) = 1-2\epsilon < 1$, we have

$$\mathbb{E}[X_0] = (I - (1-2\epsilon)A)^{-1}\epsilon\mathbf{1}.$$

Now, note

$$\begin{aligned}(I - (1-2\epsilon)A)^{-1}\mathbf{1} &= \sum_{i=0}^{\infty}(1-2\epsilon)^i A^i\mathbf{1} \\ &= \sum_{i=0}^{\infty}(1-2\epsilon)^i\mathbf{1} \\ &= \frac{1}{2\epsilon}\mathbf{1}.\end{aligned}$$

Hence, it holds $\mathbb{E}[X_0] = (1/2)\mathbf{1}$. □

Lemma 1.5.7. *For any linear ϵ-voter model with parameters A and $0 < \epsilon \leq 1/2$, the following Lyapunov matrix equation holds*

$$M = A^\epsilon M {A^\epsilon}^\top + Q^\epsilon$$

where

$$\mathbb{E}[M] = \mathbb{E}[X_0 X_0^\top]$$

and

$$Q^\epsilon = \mathbb{E}[D(V^\epsilon(X_0))] + \epsilon^2 I + \epsilon(1-\epsilon)\mathbf{1}\mathbf{1}^\top.$$

Proof. From (1.1.6), we have

$$\begin{aligned} X_{t+1}X_{t+1}^\top &= (D(Q_{t+1})Z_{t+1}X_t + R_t)(D(Q_{t+1})Z_{t+1}X_t + R_t)^\top \\ &= (D(Q_{t+1})Z_{t+1}X_t + R_t)((D(Q_{t+1})Z_{t+1}X_t)^\top + R_t^\top) \\ &= D(Q_{t+1})Z_{t+1}X_tX_t^\top(D(Q_{t+1})Z_{t+1})^\top + D(Q_{t+1})Z_{t+1}X_tR_t^\top \\ &\quad + (D(Q_{t+1})Z_{t+1}X_tR_t^\top)^\top + R_tR_t^\top. \end{aligned}$$

Now, note

$$\begin{aligned} &\mathbb{E}[(D(Q_{t+1})Z_{t+1}X_t + R_t)((D(Q_{t+1})Z_{t+1}X_t)^\top + R_t^\top)] \\ = \;& A^\epsilon \mathbb{E}[X_tX_t^\top]A^{\epsilon\top} + \mathbb{E}[D(V^\epsilon(X_t))]. \end{aligned}$$

$$(D(Q_{t+1})Z_{t+1}X_tR_t^\top)_{u,v} = Q_{t+1,u}\sum_w Z_{t+1,u,w}X_{t,w}R_{t,v}$$

$$\mathbb{E}[(D(Q_{t+1})Z_{t+1}X_tR_t^\top)_{u,v}] = \epsilon(1-2\epsilon)a_u^\top\mathbb{E}[X_t]\mathbb{I}_{\{u\neq v\}}$$

$$\mathbb{E}[D(Q_{t+1})Z_{t+1}X_tR_t^\top] = \epsilon D(A^\epsilon\mathbb{E}[X_t])(\mathbf{1}\mathbf{1}^\top - I)$$

and

$$\mathbb{E}[R_tR_t^\top] = \epsilon((1-\epsilon)I + \epsilon\mathbf{1}\mathbf{1}^\top).$$

Putting the pieces together, we have

$$\begin{aligned} Q^\epsilon &= \mathbb{E}[D(V^\epsilon(X_0))] + \\ &\quad +\epsilon(D(A^\epsilon\mathbb{E}[X_0])(\mathbf{1}\mathbf{1}^\top - I) + (\mathbf{1}\mathbf{1}^\top - I)D(A^\epsilon\mathbb{E}[X_0])) \\ &\quad +\epsilon(1-\epsilon)I + \epsilon^2\mathbf{1}\mathbf{1}^\top. \end{aligned}$$

Since, by Lemma 1.5.6, $\mathbb{E}[X_0] = (1/2)\mathbf{1}$, we have

$$\begin{aligned} & \epsilon(D(A^\epsilon\mathbb{E}[X_0])(\mathbf{1}\mathbf{1}^\top - I) + (\mathbf{1}\mathbf{1}^\top - I)D(A^\epsilon\mathbb{E}[X_0])) + \epsilon((1-\epsilon)I + \epsilon\mathbf{1}\mathbf{1}^\top) \\ = \;& \epsilon(1-2\epsilon)(\mathbf{1}\mathbf{1}^\top - I) + \epsilon(1-\epsilon)I + \epsilon^2\mathbf{1}\mathbf{1}^\top \\ = \;& \epsilon^2 I + \epsilon(1-\epsilon)\mathbf{1}\mathbf{1}^\top. \end{aligned}$$

□

Lemma 1.5.8. *For any linear ϵ-voter model with parameters A and $0 < \epsilon \leq 1/2$, we have*

$$\lambda_{\min}(\mathbb{E}[X_0 X_0^\top]) \geq \frac{1}{1-\lambda_{\min}(A)^2}(\min_u \mathbb{E}[V_{a_u^\epsilon}(X_0)] + \epsilon^2) \geq \epsilon^2.$$

Proof. Let M and Q^ϵ be defined as in Lemma 1.5.7. It is known that (see, e.g. Yasuda and Hirai [1979]),

$$\lambda_{\min}(M) \geq \frac{1}{1-\lambda_{\min}(A)^2}\lambda_{\min}(Q^\epsilon).$$

Note that Q^ϵ is a matrix with constant off-diagonal elements equal to $\alpha := \epsilon(1-\epsilon)$ and diagonal elements equal to

$$\mathbb{E}[V_{a_u^\epsilon}(X_0)] + \epsilon^2 + \alpha.$$

By Lemma 1.5.5, it follows

$$\lambda_{\min}(Q^\epsilon) \geq \min_u \mathbb{E}[V_{a_u^\epsilon}(X_0)] + \epsilon^2.$$

□

1.5.16 Proof of Theorem 1.4.3

The proof follows similar steps as that for linear dynamical systems with additive Gaussian noise Jedra and Proutiere [2019]. The differences lie in steps that are needed to resolve technical points that arise due to Bernoulli random variables and underlying constraints on the model parameter.

For any substochastic $n \times n$ matrix A, $x_0, \ldots, x_t \in \{0,1\}^n$, and $t \geq 0$, let us define

$$p_A(x_0, \ldots, x_t) := \mathbb{P}_A[X_0 = x_0, \ldots, X_t = x_t]$$

and, for any $x, y \in \{0,1\}^n$,

$$p_A(y \mid x) := \mathbb{P}_A[X_{t+1} = y \mid X_t = x].$$

Let A^* be an $n \times n$ stochastic matrix and A be an $n \times n$ substochastic matrix such that $A \neq A^*$. The log-likelihood ratio of the observed voter model process states, under parameters A^* and A, is given by

$$L(X) = \sum_{i=1}^{m} \log \left(\frac{p_{A^*}(X_0^{(i)}, \ldots, X_{\tau_i}^{(i)})}{p_A(X_0^{(i)}, \ldots, X_{\tau_i}^{(i)})} \right).$$

For every $t \geq 1$, we have

$$p_{A^*}(x_0, \ldots, x_t) = \mu(x_0) p_{A^*}(x_1 \mid x_0) \cdots p_{A^*}(x_t \mid x_{t-1}).$$

Now, note

$$\begin{aligned} \mathbb{E}^0_{A^*}[L(X)] &= m\mathbb{E}^0_{A^*} \left[\sum_{t=0}^{\tau-1} \mathbb{E}^0_{A^*} \left[\log \left(\frac{p_{A^*}(X_{t+1} \mid X_t)}{p_A(X_{t+1} \mid X_t)} \right) \mid \mathcal{F}_t \right] \right] \\ &= m\mathbb{E}^0_{A^*} \left[\sum_{t=0}^{\tau-1} \sum_{u=1}^{n} \mathrm{KL}({a_u^*}^\top X_t \mid\mid {a_u}^\top X_t) \right]. \end{aligned}$$

For every $u \in V$, let S_u denote the support of a_u^*, $\mathcal{X}_u = \{x \in \{0,1\}^n : 0 < \sum_{j \in S_u} x_j < |S_u|\}$, and $\mathcal{A}_u^* = \{a \in \mathbb{R}_+^n \mid \sum_{v=1}^n a_v = 1, a_w \geq \alpha/2, \text{ for all } w \in S_u\}$. Then, for every $x \in \mathcal{X}_u$ and $a_u \in \mathcal{A}_u^*$, we have $\alpha/2 \leq x^\top a_u \leq 1 - \alpha/2$.

If $X_s \in \mathcal{X}_u$ and $a_u \in \mathcal{A}_u^*$, we have

$$\begin{aligned} \mathrm{KL}({a_u^*}^\top X_t \,||\, {a_u}^\top X_t) &\leq \frac{4}{\alpha}({a_u^*}^\top X_t - {a_u}^\top X_t)^2 \\ &= \frac{4}{\alpha} X_t^\top (a_u^* - a_u)(a_u^* - a_u)^\top X_t. \end{aligned}$$

In the remainder of the proof, we assume that A is such that $a_u \in \mathcal{A}_u^*$ for all $u \in \{1, \ldots, n\}$. It follows

$$\begin{aligned} \mathbb{E}_{A^*}^0[L(X)] &\leq m\frac{4}{\alpha}\mathbb{E}_{A^*}^0\left[\sum_{t=0}^{\tau-1}\sum_{u=1}^{n}({a_u^*}^\top X_t - {a_u}^\top X_t)^2\right] \\ &= m\frac{4}{\alpha}\mathbb{E}_{A^*}^0\left[\sum_{t=0}^{\tau-1}\sum_{u=1}^{n}\mathbb{E}_{A^*}^0[X_t^\top ({a_u}^* - a_u)(a_u^* - a_u)^\top X_t]\right] \\ &= m\frac{4}{\alpha}\mathbb{E}_{A^*}^0\left[\sum_{t=0}^{\tau-1} X_t^\top W X_t\right] \end{aligned}$$

where $W := (A^* - A)^\top (A^* - A)$.

By the elementary properties of the trace of a matrix, we have

$$\mathbb{E}_{A^*}^0\left[\sum_{t=0}^{\tau-1} X_t^\top W X_t\right] = \mathrm{tr}\left(W\mathbb{E}_{A^*}^0\left[\sum_{t=0}^{\tau-1} X_t X_t^\top\right]\right).$$

Hence, it holds

$$\mathbb{E}_{A^*}^0[L(X)] \leq m\frac{4}{\alpha}\mathrm{tr}\left(W\mathbb{E}_{A^*}^0\left[\sum_{t=0}^{\tau-1} X_t X_t^\top\right]\right).$$

By the Palm inversion formula (1.2.2), we have

$$\mathbb{E}_{A^*}^0\left[\sum_{t=0}^{\tau-1} X_t X_t^\top\right] = \mathbb{E}_{A^*}^0[\tau]\mathbb{E}_{A^*}[X_0 X_0^\top].$$

It follows that

$$\mathbb{E}_{A^*}^0[L(X)] \leq \frac{4}{\alpha} m \mathbb{E}_{A^*}^0[\tau]\mathrm{tr}(W\mathbb{E}_{A^*}[X_0 X_0^\top]). \tag{1.5.26}$$

Let $\mathcal{F}_m$ denote the σ-algebra of observations from m independent realizations of the voter model process with parameter A^* and initial state distribution μ. By the data processing inequality, we have

$$\mathbb{E}_{A^*}^0[L(X)] \geq \sup_{E \in \mathcal{F}_m} \mathrm{KL}(\mathbb{P}_{A^*}^0[E] \,||\, \mathbb{P}_A^0[E]). \tag{1.5.27}$$

Assume in addition that A satisfies $2\epsilon \leq ||A - A^*||_F \leq 3\epsilon$, and assume that $m \geq m_0$. Let E be the $\mathcal{F}_m$-measurable event defined as

$$E = \{||\hat{A} - A^*||_F \leq \epsilon\}.$$

Since the algorithm is (ϵ, δ)-locally stable, we have

$$\mathbb{P}_{A^*}[||\hat{A} - A^*||_F \leq \epsilon] \geq 1 - \delta$$

and

$$\mathbb{P}_A[||\hat{A} - A^*||_F \leq \epsilon] \leq \mathbb{P}_A[||\hat{A} - A||_F > \epsilon] \leq \delta.$$

Hence, it follows

$$\mathrm{KL}(\mathbb{P}^0_{A^*}(E) \,||\, \mathbb{P}^0_A(E)) \geq \mathrm{KL}(1 - \delta \,||\, \delta) \geq \log\left(\frac{1}{2.4\delta}\right). \tag{1.5.28}$$

Combining (1.5.26), (1.5.27) and (1.5.28), it follows that for any (ϵ, δ)-locally stable estimator in A^*, for all substochastic matrices A such that (a) $2\epsilon \leq ||A - A^*||_F \leq 3\epsilon$ and (b) $a_{u,v} \geq \alpha/2$ for every (u, v) in the support of A^*, and $m \geq m_0$, we have

$$m\mathbb{E}^0_{A^*}[\tau]\mathrm{tr}\left(W\mathbb{E}_{A^*}[X_0X_0^\top]\right) \geq \frac{\alpha}{4}\log\left(\frac{1}{2.4\delta}\right) \tag{1.5.29}$$

where, recall,

$$W = (A^* - A)^\top(A^* - A).$$

We need to show that there exists a substochastic matrix A that minimizes the left-hand side of inequality (1.5.29) under the given constraints.

Let $\mathbb{E}_{A^*}[X_0X_0^\top] = Q\Lambda Q^\top$ be the eigenvalue decomposition of the correlation matrix $\mathbb{E}_{A^*}[X_0X_0^\top]$, with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, and eigenvectors $Q = (q_1, \ldots, q_n)$. Hence, we have $\mathbb{E}_{A^*}[X_0X_0^\top] = \sum_{i=1}^n \lambda_i q_i q_i^\top$. Finding A that minimizes the left-hand side of the inequality (1.5.29) corresponds to

finding A that is a solution of the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \operatorname{tr}(W \textstyle\sum_{v=1}^{n} \lambda_v q_v q_v^\top) \\ \text{subject to} \quad & W = (A^* - A)^\top (A^* - A) \\ & 2\epsilon \leq \sqrt{\operatorname{tr}(W)} \leq 3\epsilon \\ & a_{u,v} \geq \alpha/2 \text{ for every } (u, v) \text{ in the support of } A^* \\ & A \text{ is a substochastic matrix.} \end{aligned}$$

By taking $W = 4\epsilon^2 q_1 q_1^\top$, we have

$$\operatorname{tr}(W \mathbb{E}_{A^*}[X_0 X_0^\top]) = 4\epsilon^2 \lambda_1$$

and

$$||A - A^*||_F = \sqrt{\operatorname{tr}(W)} = 2\epsilon.$$

We need to show that there exists a matrix A that satisfies the constraints of the above optimization problem. To show this, let A be such that for some fixed $u \in \{1, \ldots, n\}$, $a_v^* = a_v$ for all $v \neq u$, and a_u is given by

$$a_u^* - a_u = 2\epsilon q_1. \tag{1.5.30}$$

For such a matrix A, we have $W = (a_u^* - a_u)(a_u^* - a_u)^\top$, and clearly $W = 4\epsilon^2 q_1 q_1^\top$. Note that $||A^* - A||_F = ||a_u^* - a_u||_2 = 2\epsilon$.

From (1.5.30), we have

$$a_u^\top \mathbf{1} = 1 - 2\epsilon q_1^\top \mathbf{1}.$$

Hence, A is a substochastic matrix, if and only if,

$$2\epsilon |q_1^\top \mathbf{1}| \leq 1.$$

By Cauchy-Schwartz inequality $|q_1^\top \mathbf{1}| \leq ||\mathbf{1}||_2 ||q_1||_2 = \sqrt{n}$. Hence, if $\epsilon \leq 1/(2\sqrt{n})$, then A is a substochastic matrix.

From (1.5.30), for every (u, v) which is in the support of A^*, i.e. $a^*_{u,v} > 0$, it must hold

$$a^*_{u,v} - 2\epsilon q_{1,v} \geq \alpha/2.$$

Since $|q_{1,v}| \leq 1$ for all v, $a^*_{u,v} \geq \alpha$ for all (u, v) in the support of A^*, we have that the above condition holds if $\epsilon \leq \alpha/4$.

1.5.17 Asynchronous voter model on a path

We consider the asynchronous voter model process on a path of two or more vertices, where at each time step one vertex, chosen uniformly at random, updates its state. We assume that initial node states are such that k vertices on one end of the path are in state 1 and other vertices are in state 0. For any such initial state, at every time step, there are at most two vertices with a mixed neighborhood set. If such two vertices exist, they reside on the boundary separating the state-1 vertices from state-0 vertices. Note that an *informative interaction* for the parameter estimation problem occurs only when one of these boundary vertices samples a neighbour. See Figure 1.1 for an illustration.

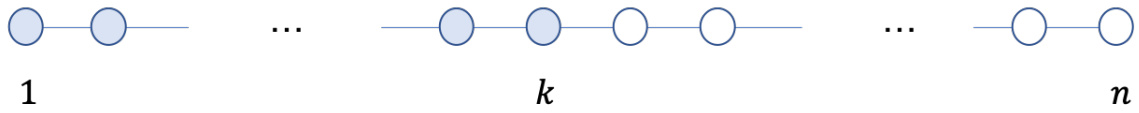


Figure 1.1: A path example: initial state is such that k leftmost vertices are in state 1 and the remaining $n - k$ rightmost vertices are in state 0. Each vertex samples a neighbour equiprobably.

The expected number of vertices that perform at least one informative interaction until the voter model process hits a consensus state can be characterized as asserted in the following proposition.

Proposition 1.5.1. *Consider the voter model on a path with $n \geq 2$ vertices such that each vertex samples a neighbor with equal probabilities, with initial states such that k vertices on one end of the path are in state 1 and other vertices are in state 0. Then, the number of vertices N that participate in at least one informative interaction has the expected value*

$$\mathbb{E}[N] = k \log\left(\frac{n}{k}\right) + (n-k)\log\left(\frac{n}{n-k+1}\right) + \Theta(1).$$

Note that if $k/n = o(1)$, then

$$\mathbb{E}[N] = k\left(\log\left(\frac{n}{k}\right) + \Theta(1)\right).$$

On the other hand, if k is a fixed constant, then

$$\mathbb{E}[N] = O(\log(n)).$$

This makes precise the intuition that only a small fraction of vertices will participate in at least one informative interaction if a small fraction of vertices on one end of the path are in state 1 and other vertices are in state 0, for asymptotically large n.

In the remainder of this section, we prove the proposition. Let I_t denote the vertex activated at time step t. Let p_u denotes the probability with which vertex u samples vertex $u+1$. Then, $1-p_u$ is the probability of vertex u sampling vertex $u-1$. Here $p = (p_1, p_2, \ldots, p_n)$ are parameters such that $p_u \in (0,1)$ for $u = 1, 2, \ldots, n$. In the proposition, we consider the case $p_1 = 1$, $p_n = 0$, and $p_u = 1/2$, for $u \in \{1, \ldots, n\} \setminus \{1, n\}$.

If vertex u is active at time step t, then the state of u is according to

$$\mathbb{P}[X_{t+1,u} = 1 \mid X_t = x] = \begin{cases} (1-p_1)x_1 + p_1 x_2 & \text{if } u = 1 \\ (1-p_u)x_{u-1} + p_u x_{u+1} & \text{if } 1 < u < n, \\ (1-p_n)x_{n-1} + p_n x_n & \text{if } u = n, \end{cases}$$

and the states of other vertices remain unchanged.

Given observed node states until absorption to a consensus state, we want to estimate the values of parameters $p_1, \ldots, p_n$. We consider this parameter estimation problem for the initial state such that $X_{0,u} = 1$ for $u \in \{1, \ldots, k\}$ and $X_{0,u} = 0$ for $u \in \{k+1, \ldots, n\}$, for some $k \in \{1, \ldots, n-1\}$.

Under given condition on the initial state, the system dynamics is fully described by Y_t defined as the number of vertices in state 1 at time step t. Note that $\{Y_t\}_{t \geq 0}$ is a Markov chain with state space

$\{0, \ldots, n\}$, initial state $Y_0 = k$ and the transition probabilities:

$$\begin{aligned}
\mathbb{P}[Y_{t+1} = y+1 \mid Y_t = y] &= \frac{1}{n}(1 - p_{y+1})\mathbb{I}_{\{0<y<n\}} \\
\mathbb{P}[Y_t = y-1 \mid Y_t = y] &= \frac{1}{n}p_y\mathbb{I}_{\{0<y<n\}} \\
\mathbb{P}[Y_{t+1} = y \mid Y_t = y] &= 1 - \frac{1}{n}(1 - p_{y+1})\mathbb{I}_{\{0<y<n\}} - \frac{1}{n}p_y\mathbb{I}_{\{0<y<n\}}.
\end{aligned}$$

This Markov chain has two absorbing states 0 and n.

The Markov chain $\{Y_t\}_{t\geq 0}$ has a jump point at time step t if, and only if, (a) vertex Y_t is active and this vertex samples vertex $Y_t + 1$ or (b) vertex $Y_t + 1$ is active and this vertex samples vertex Y_t. We refer to each jump point of Y as a useful interaction as only at a jump point we can observe outcome of a Bernoulli experiment of which vertex is sampled by the active node, with parameter in a strict interior of $(0, 1)$.

If at time step t, the active vertex is $I_t = Y_t$, then this vertex sampled vertex $Y_t + 1$ if we observe $Y_{t+1} - Y_t = -1$, otherwise this vertex sampled vertex $Y_t - 1$ if we observe $Y_{t+1} - Y_t = 0$. If at time step t, the active vertex is $I_t = Y_{t+1} + 1$, then this vertex sampled vertex Y_t if we observe $Y_{t+1} - Y_t = 1$ and otherwise this vertex sampled $Y_t + 2$ if we observe $Y_{t+1} - Y_t = 0$.

We next consider the probability that a given vertex has at least one informative interaction before Y gets absorbed in either state 0 or 1. This is of interest because the maximum likelihood estimate of p_u is well defined only if at least one informative interaction is performed for vertex u.

Let $h_{u,y}$ be the probability that vertex u has at least one informative interaction given that Y started at initial state $Y_0 = y$. We want to compute the values of $h_{u,k}$ for $1 \leq u \leq n$ and $1 < k < n$.

Case $k < u$ Vertex u has at least one informative interaction if, and only if, there exists a time step t such that $I_t = u$ and $Y_t = u - 1$. Let $\tilde{Y}$ be a Markov chain with state space $\{1, 2, \ldots, u\}$ and transition probabilities for vertices $0 \leq v < u - 1$ corresponding to those of Y and by definition

$$\begin{aligned}
\mathbb{P}[\tilde{Y}_{t+1} = u \mid \tilde{Y}_t = u-1] &= \frac{1}{n} \\
\mathbb{P}[\tilde{Y}_{t+1} = u-2 \mid \tilde{Y}_t = u-1] &= \frac{1}{n}p_{u-1} \\
\mathbb{P}[\tilde{Y}_{t+1} = u-1 \mid \tilde{Y}_t = u-1] &= 1 - \frac{1}{n} - \frac{1}{n}p_{u-1}
\end{aligned}$$

and

$$\mathbb{P}[\tilde{Y}_{t+1} = u \mid \tilde{Y}_t = u] = 1.$$

The value of $h_{u,k}$ corresponds to the probability of Markov chain $\tilde{Y}$ hitting state u by starting from state k. We have boundary conditions $h_{u,0} = 0$ and $h_{u,u} = 1$. By the first-step analysis of Markov chains, for $0 < y < u$,

$$h_{u,y} = \frac{1}{n} p_y h_{u,y-1} + \frac{1}{n}(1 - p_{y+1}) h_{u,y+1} + \left(1 - \frac{1}{n} p_y - \frac{1}{n}(1 - p_{y+1})\right) h_{u,y}$$

where we abuse the notation by assuming that $p_u = 0$. We can write

$$(p_y + 1 - p_{y+1}) h_{u,y} = p_y h_{u,y-1} + (1 - p_{y+1}) h_{u,y+1}, \text{ for } 0 < y < u.$$

Now, this can be equivalently written as

$$(1 - p_{y+1})(h_{u,y+1} - h_{u,y}) = p_y (h_{u,y} - h_{u,y-1}), \text{ for } 0 < y < u.$$

Let $\delta_{u,y} := h_{u,y+1} - h_{u,y}$. Then, note

$$\delta_{u,y} = \frac{p_y \cdots p_1}{(1 - p_{y+1}) \cdots (1 - p_2)} \delta_{u,0}$$

where $\delta_{u,0} = h_{u,1}$. Since $h_{u,y} = \delta_{u,y-1} + \cdots + \delta_{u,0}$ and $h_{u,u} = 1$, we have

$$\left(1 + \sum_{z=1}^{u-1} \frac{p_z \cdots p_1}{(1 - p_{z+1}) \cdots (1 - p_2)}\right) \delta_{u,0} = 1.$$

Hence, we have

$$h_{u,k} = \frac{1 + \sum_{z=1}^{k-1} \frac{p_z \cdots p_1}{(1-p_{z+1}) \cdots (1-p_2)}}{1 + \sum_{z=1}^{u-1} \frac{p_z \cdots p_1}{(1-p_{z+1}) \cdots (1-p_2)}}, \text{ for } 1 \leq k < u. \tag{1.5.31}$$

For the special when all the transition probabilities of Y of values in $(0, 1)$ are equal to $1/2$, we have

$$h_{u,k} = \frac{2k}{2u - 1}, \text{ for } 1 \leq k < u. \tag{1.5.32}$$

Case $k > u$ In this case, vertex u has at least one informative interaction if, and only, if there exits a time step t such that $I_t = u$ and $Y_t = u$. Let $\tilde{Y}$ be a Markov chain with state space $\{u - 1, u, \ldots, n\}$

and transition probabilities for vertices $u < v \leq n$ corresponding to those of Y and by definition

$$\begin{aligned}
\mathbb{P}[\tilde{Y}_{t+1} = u+1 \mid \tilde{Y}_t = u] &= \frac{1}{n}(1 - p_{u+1}) \\
\mathbb{P}[\tilde{Y}_{t+1} = u-1 \mid \tilde{Y}_t = u] &= \frac{1}{n} \\
\mathbb{P}[\tilde{Y}_{t+1} = u \mid \tilde{Y}_t = u] &= 1 - \frac{1}{n}(1 - p_{u+1}) - \frac{1}{n}
\end{aligned}$$

and

$$\mathbb{P}[\tilde{Y}_{t+1} = u-1 \mid \tilde{Y}_t = u-1] = 1.$$

The value of $h_{u,k}$ corresponds to $\tilde{Y}$ hitting state $u-1$ by starting from state k. We have boundary conditions $h_{u,u-1} = 1$ and $h_{u,n} = 0$. By same arguments as before, for $u - 1 < y < n$,

$$h_{u,y} = \frac{1}{n} p_y h_{u,y-1} + \frac{1}{n}(1 - p_{y+1}) h_{u,y+1} + \left(1 - \frac{1}{n} p_y - \frac{1}{n}(1 - p_{y+1})\right) h_{u,y}$$

where we abuse the notation by assuming $p_u = 1$.

Again, it follows

$$(1 - p_{y+1})(h_{u,y+1} - h_{u,y}) = p_y (h_{u,y} - h_{u,y-1}), \text{ for } u - 1 < y < n$$

and

$$\delta_{u,y} = \frac{p_y \cdots p_u}{(1 - p_{y+1}) \cdots (1 - p_{u+1})} \delta_{u,u-1}, \text{ for } u - 1 < y < n.$$

Since $h_{u,y} = -(\delta_{u,y} + \delta_{u,y+1} + \cdots + \delta_{u,n-1})$ and $h_{u-1,u} = 1$, we obtain

$$h_{u,k} = \frac{\sum_{z=k}^{n-1} \frac{p_z \cdots p_u}{(1-p_{z+1}) \cdots (1-p_{u+1})}}{1 + \sum_{z=i}^{n-1} \frac{p_z \cdots p_u}{(1-p_{z+1}) \cdots (1-p_{u+1})}}, \text{ for } u < k < n. \tag{1.5.33}$$

In particular, when all the transition probabilities of Y of values in $(0, 1)$ are equal to $1/2$, we have

$$h_{i,k} = \frac{2(n-k)}{2(n-i)+1}, \text{ for } i < k < n. \tag{1.5.34}$$

Case $k = i$ In this case, we have

$$h_{k,k} = \frac{1}{n} + \frac{1}{n}(1 - p_{k+1})h_{k,k+1} + \left(1 - \frac{1}{n} - \frac{1}{n}(1 - p_{k+1})\right) h_{k,k}$$

Hence,

$$h_{k,k} = \frac{1 + (1 - p_{k+1})h_{k,k+1}}{2 - p_{k+1}}. \tag{1.5.35}$$

In particular, when all the transition probabilities of Y of values in $(0, 1)$ are equal to $1/2$, we have

$$h_{k,k} = \frac{2}{3}\left(1 + \frac{(n-k)}{2(n-k)+1}\right). \tag{1.5.36}$$

We next discuss the results of the above analysis for the special case when the transition probabilities of Y of value in $(0, 1)$ are equal to $1/2$. From (1.5.31), we observe that vertex n has at least one informative interaction with probability

$$h_{n,k} = \frac{2n}{2n-1}\frac{k}{n}.$$

For large n, we have $h_{n,k} \sim k/n$. It follows that vertex n has a diminishing probability of having at least one informative interaction provided that $k/n = o(1)$. For instance, if k is a constant, then $h_{n,k} = \Theta(1/n)$.

Let N be the number of vertices with at least one informative interaction. We have

$$\mathbb{E}[N] = \sum_{u=1}^{n} h_{u,k}.$$

Note the following elementary identity

$$S_n := \sum_{u=1}^{n} \frac{1}{2u-1} = \frac{1}{2}(2H_{2n} - H_n)$$

and note that

$$T_n := \sum_{i=1}^{n} \frac{1}{2n+1} = S_{n+1} - 1.$$

We first compute

$$\begin{aligned}\sum_{u=k+1}^{n} h_{u,k} &= 2k \sum_{u=k+1}^{n} \frac{1}{2u-1} \\ &= 2k(S_n - S_k) \\ &= k(2H_{2n} - H_n - 2H_{2k} + H_k) \\ &= k\log(n/k)(1+o(1)).\end{aligned}$$

Then, we compute

$$\begin{aligned}\sum_{u=1}^{k-1} h_{u,k} &= 2(n-k) \sum_{u=1}^{k-1} \frac{1}{2(n-u)+1} \\ &= 2(n-k) \sum_{v=n-k+1}^{n-1} \frac{1}{2v+1} \\ &= 2(n-k)(T_{n-1} - T_{n-k}) \\ &= 2(n-k)(S_n - S_{n-k+1}) \\ &= (n-k)(2H_{2n} - H_n - 2H_{2(n-k+1)} + H_{n-k+1}) \\ &= (n-k)\log(n/(n-k+1))(1+o(1)).\end{aligned}$$

It follows that

$$\mathbb{E}[N] = k\log\left(\frac{n}{k}\right) + (n-k)\log\left(\frac{n}{n-k+1}\right) + \Theta(1).$$

Chapter 2

Accelerated MM Algorithms for Inference of Ranking Scores from Comparison Data

2.1 Introduction

Rank aggregation is an important task that arises in a wide-range of applications, including recommender systems, information retrieval, online gaming, sports competitions, and evaluation of machine learning algorithms. Given a set of items, rank aggregation aims to infer ranking scores of items or an ordering of items based on observed data containing partial orderings of items. A typical scenario is that of paired comparisons, where observations consist of information about which item is preferred in a pairwise comparison. For example, player A defeats player B in a game, product A is preferred over product B by a user, and machine learning algorithm A outperforms machine learning algorithm B in an evaluation. In such scenarios, a common goal is not only to compute an aggregate ranking of items, but also to compute ranking scores, which represent strengths of individual items. Such ranking scores are used for predicting outcomes of future ranking outcomes, such as predicting outcomes of matches in online games and sport contests, and predicting preferences of users in product shopping or movie watching scenarios, among others. Note that, importantly, observations are not restricted to paired comparisons, but may also include other types of comparison data, such as choice (e.g., product A chosen from a set of two or more products) or full ranking (e.g., a ranking list of players or teams participating in a competition).

In this chapter, our goals are twofold. First, we aim to shed light on the efficiency of one of the most

popular iterative optimization methods for inferring ranking scores, namely the MM algorithm, where ranking scores correspond to parameter estimates of popular Bradley-Terry family of models. Second, we propose an accelerated MM algorithm that resolves a slow convergence issue found to hold for a classic MM algorithm.

2.1.1 Related work

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 sport games (Elo [1978a]), 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 (Rao and Kupper [1967]), choices from comparison sets of two or more items, e.g., Luce choice model (Luce [1959]), full ranking outcomes for comparison sets of two or more items, e.g., Plackett-Luce ranking model (Plackett [1975]), as well as group comparisons (Huang et al. [2006b, 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 (Thurstone [1927b]).

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 Elo [1978a]. Extensions to team competitions and tie outcomes were implemented in popular online gaming platforms, e.g. TrueSkill rating system Herbrich et al. [2006]. 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 Burges et al. [2006], Li [2011]. Statistical models of paired comparisons are used in timely applications such as evaluation of reinforcement learning algorithms Balduzzi et al. [2018].

An iterative optimization algorithm for the maximum likelihood (ML) parameter estimation (MLE) of the Bradley-Terry model has been known since the original work of Zermelo [1929]. Lange et al. [2000] 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 maximization or minimization of an objective function. Lange [2016] provided a book on MM algorithms and Hunter and Lange [2004] provided a tutorial. Mairal [2015] established some convergence results for incremental MM algorithms.

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 [2012a], 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, where $\alpha \geq 1$ is the *shape* parameter and $\beta > 0$ is the *rate* parameter. Importantly, unlike to the ML estimation, the MAP estimate is always guaranteed to exist, for any given observation data.

Algorithms for fitting Bradley-Terry model parameters are implemented in open source software packages, including BradleyTerry2 (Turner and Firth [2012]), BradleyTerryScalable (Kaye and Firth [2020]), and Choix (Maystre [2018]). The first package uses a Fisher scoring algorithm (a second-order optimization method), while the latter two use MM algorithms (a first-order optimization method). First-order methods are generally preferred over second-order methods for fitting high-dimensional models using large training datasets. Our focus in this chapter is on first-order optimization methods, specifically, gradient descent and MM algorithms.

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. [2014b], Rajkumar and Agarwal [2014], Chen and Suh [2015], Shah et al. [2016], Khetan and Oh [2016a], Vojnovic and Yun [2016b], Borkar et al. [2016], Negahban et al. [2017], Chen et al. [2019], Han et al. [2020], Wang et al. [2020], Hendrickx

et al. [2020], Ruijian Han and Chen [2022], Li et al. [2022]. Note that the question about statistical estimation accuracy and computation complexity tradeoff is out of the scope of our chapter, and this was studied in the above cited papers. The focus of our work is on *convergence properties* of *first-order* iterative optimization methods for parameter estimation of Bradley-Terry models. Here "first-order" refers to optimization methods that are restricted to value oracle access to gradients of the optimization objective function, thus not allowing access to second-order properties such as values of the Hessian matrix. Specifically, we are interested in convergence properties of first-order methods for ML and MAP estimation objectives. It is noteworthy that some recently proposed algorithms show empirically faster convergence rate than MM, e.g., Negahban et al. [2017], Maystre and Grossglauser [2015a], Agarwal et al. [2018], but it is hard to apply them for the MAP estimation objective. We thus restrict our attention to MM and gradient descent algorithms which are able to solve both MLE and MAP optimization problems.

While 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 Vojnovic et al. [2020], we closed this gap by providing tight characterizations of convergence rates. We presented the 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 showed that both gradient descent and MM algorithms have linear convergence with convergence rates differing only in constant factors. An iterative optimization algorithm that has linear convergence is generally considered to be fast in the space of first-order optimization algorithms, and many first-order algorithms cannot guarantee a linear convergence. For example, standard stochastic gradient descent algorithm is known to have sub-linear convergence, see, e.g., Bubeck [2015]. We provided explicit bounds on convergence rates that provide insights into which properties of observed comparison data play a key role for the rate of convergence.

Specifically, we showed that the rate of convergence critically depends on certain properties of the matrix of counts of item pair co-occurrences, $\mathbf{M}$, in input comparison data. We found that 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. Here $a(\mathbf{M})$ is the Fiedler value (eigenvalue), see Fiedler [1973], defined as the second smallest eigenvalue of the Laplacian matrix $\mathbf{L_M} = \mathbf{D_M} - \mathbf{M}$, where $\mathbf{D_M}$ is the diagonal matrix with diagonal elements equal to the row sums of $\mathbf{M}$. The Fiedler value of a

matrix of paired comparison counts is known to play a key role in determining the MLE accuracy, e.g., Hajek et al. [2014b], Shah et al. [2016], Khetan and Oh [2016c], Vojnovic and Yun [2016b], Negahban et al. [2017]. These works characterized the number of samples needed to estimate the true parameter value within a statistical estimation error tolerance. This is different from the problem of characterizing the number of iterations needed for an iterative optimization algorithm to compute a ML or a MAP parameter estimate satisfying an error tolerance condition, which is studied in this chapter.

Our results in Vojnovic et al. [2020] revealed the following facts about convergence time, defined as the number of iterations that an iterative optimization algorithm takes to reach the value of the underlying objective function within a given error tolerance parameter $\epsilon > 0$ of the optimum value.

For the ML objective, we showed that the convergence time satisfies

$$T^{\mathrm{ML}} = O\left(\frac{d(\mathbf{M})}{a(\mathbf{M})} \log\left(\frac{1}{\epsilon}\right)\right) \tag{2.1.1}$$

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

For the MAP estimation, we showed that the convergence time satisfies

$$T^{\mathrm{MAP}} = O\left(\left(\frac{d(\mathbf{M})}{\beta} + 1\right) \log\left(\frac{1}{\epsilon}\right)\right) \tag{2.1.2}$$

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. We observed that the convergence time for the MAP estimation problem can be arbitrarily large for small enough value of β, where small values of β correspond to less informative prior distributions.

From the convergence time results in Vojnovic et al. [2020], we identifed a slow rate of convergence issue for gradient descent and MM algorithms for the MAP estimation problem. While the MAP estimation alleviates the issue of the 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 (2.1.1) holds, while for the MAP estimation, the convergence time grows arbitrarily large as β approaches 0 from

above.

In the present chapter, we extend our prior work Vojnovic et al. [2020] by proposing new accelerated algorithms and establishing their theoretical guarantees (Section 2.4) as well as demonstrating their efficiency through numerical evaluations (Section 2.5).

2.1.2 Summary of our contributions for this chapter

We propose an acceleration method for the MAP estimation objective that has convergence time bounded as follows

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

This acceleration method resolves the slow convergence issue of classic MM algorithm for the MAP estimation for generalized Bradley-Terry models. 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 normalizes the parameter vector estimate in each iteration of the gradient descent or MM algorithm using a transformation that ensures (a) that the value of the objective function is non decreasing along the sequence of parameter estimates and (b) that the objective function satisfies certain smoothness and strong convexity properties that ensure high convergence rate. This amounts to a slight modification of the classical MM algorithm to resolve the slow convergence issue. This acceleration method is derived by using a theoretical framework that may be of general interest. This framework can be applied to other statistical models of ranking data and prior distributions for Bayesian inference of parameters of these models.

We present numerical evaluation of the convergence time 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 algorithm for MAP estimation and show significant speed ups achieved by our accelerated MM algorithm.

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 by using concepts of smooth and strongly convex functions. We derive accelerated iterative optimization algorithms based on a general approach that may be of independent interest. This approach transforms the parameter estimator in each iteration so that certain conditions are preserved for the gradient vector and the

Hessian matrix of the objective function. For generalized Bradley-Terry models, this transformation turns out to be simple, yielding practical algorithms.

2.1.3 Organization of the chapter

In Section 2.2, we present problem formulation and some background material. We summarise our prior results in Vojnovic et al. [2020] on characterization of convergence rates of gradient descent and MM algorithms in Section 2.3; for simplicity of exposition, we focus only on the Bradley-Terry model of paired comparisons. Section 2.4 presents our accelerated algorithms for MAP estimation. Section 2.5 contains our numerical results. We conclude in Section 2.6. Section 2.7 contains all our proofs, additional discussions, and extensions to generalized Bradley-Terry models.

2.2 Problem formulation

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 outcomes is given by

$$\Pr[i \succ j] = \frac{\theta_i}{\theta_i + \theta_j} \tag{2.2.1}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^\top \in \mathbb{R}^n_+$ is the 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 by classic books on categorical data analysis, e.g., Agresti [2002].

We will sometimes use the parametrization $\theta_i = e^{w_i}$ when it is simpler to express an equation or when we want to make a connection with the literature using this parametrization. Using parameterization $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}^n$, we have

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

All our convergence results are for the model with parameter $\mathbf{w}$. The Bradley-Terry type models for paired comparisons with ties, choice, and full ranking outcomes, we refer to as *generalized Bradley-Terry models*, are defined in Section 2.7.8. Our results apply to all these different models. In the main body of this chapter, we focus only on the Bradley-Terry model for paired comparisons in order to

keep the presentation simple.

Maximum likelihood estimation The maximum likelihood parameter estimation problem corresponds to finding $\mathbf{w}^\star$ that solves the following optimisation problem:

$$\max_{\mathbf{w}\in\mathbb{R}^n} \ell(\mathbf{w}) \tag{2.2.2}$$

where $\ell(\mathbf{w})$ is the log-likelihood function,

$$\ell(\mathbf{w}) = \sum_{i=1}^{n}\sum_{j\neq i} d_{i,j}\left(w_i - \log\left(e^{w_i} + e^{w_j}\right)\right) \tag{2.2.3}$$

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

The maximum likelihood optimisation problem (2.2.2) is a convex optimization problem. Note, however, that the objective function is not a strictly concave function as adding a common constant to each element of the parameter vector keeps the value of the objective function unchanged.

MAP estimation problem An alternative objective is obtained by using a *Bayesian inference framework*, which amounts to finding a maximum a posteriori estimate of the parameter vector under a given prior distribution. We consider the Bayesian method introduced by Caron and Doucet [2012a], which assumes the prior distribution to be of product-form with marginal distributions such that $\theta_i(= e^{w_i})$ has a Gamma(α, β) distribution where $\alpha \geq 1$ is the *shape* parameter and $\beta > 0$ is the *rate* parameter. Note that α and β affect the *scale* of the parameter vector as with respect to the Gamma(α, β) prior distribution, θ_i has the expected value and the mode equal to α/β and $(\alpha - 1)/\beta$, respectively. For any fixed $\alpha \geq 1$, the density of Gamma(α, β) distribution becomes more flat as β approaches zero which corresponds to a less informative prior. According to the assumed prior distribution, $\sum_{i=1}^n \theta_i \sim \text{Gamma}(n\alpha, \beta)$ and, hence, the mode of $\sum_{i=1}^n \theta_i$ is $(n\alpha - 1)/\beta$. We can interpret the mode of $\sum_{i=1}^n \theta_i$ as the *scale* of the parameter vector.

The log-a posteriori probability function can be written as

$$\rho(\mathbf{w}) = \ell(\mathbf{w}) + \ell_0(\mathbf{w}) \tag{2.2.4}$$

where ℓ is the log-likelihood function in (2.2.3) and ℓ_0 is the log-likelihood of the prior distribution

given by

$$\ell_0(\mathbf{w}) = \sum_{i=1}^{n} \left((\alpha - 1)w_i - \beta e^{w_i}\right). \tag{2.2.5}$$

Note that for $\alpha = 1$ and $\beta = 0$, the log-a posteriori probability function corresponds to the log-likelihood function. For these values of parameters α and β, MAP and ML estimation problems are equivalent.

MM algorithms The MM algorithm for minimizing a function f is defined by minimizing a surrogate function that *majorizes* 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 iterative updates:

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}} g(\mathbf{x}; \mathbf{x}^{(t)}). \tag{2.2.6}$$

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.

The classic MM algorithm for the Bradley-Terry model of paired comparisons uses the following minorization function of $\ell(\mathbf{x})$:

$$\underline{\ell}(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^{n} \sum_{j \neq i} \underline{\ell}_{ij}(\mathbf{x}; \mathbf{y}), \tag{2.2.7}$$

where

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

It is easy to observe that $\underline{\ell}(\mathbf{x}; \mathbf{y})$ is a minorization surrogate function of $\ell(\mathbf{x})$ by noting that $\log(x) \leq x - 1$ and that equality holds if, and only if, $x = 1$, which is used to break $\log(e^{x_i} + e^{x_j})$ terms in the log-likelihood function.

The *classic MM algorithm* for the ML parameter estimation of the Bradley-Terry model of paired comparisons (Ford [1957], Hunter [2004]), is defined by the following iterative updates, for $i =$

$1, 2, \ldots, n$,

$$\theta_i^{(t+1)} = \frac{\sum_{j=1}^{n} d_{i,j}}{\sum_{j=1}^{n} \frac{m_{i,j}}{\theta_i^{(t)} + \theta_j^{(t)}}}. \tag{2.2.8}$$

Following Caron and Doucet [2012a], the MM algorithm for the MAP parameter estimation of the Bradley-Terry model of paired comparisons is derived for the minorant surrogate function $\underline{\rho}$ of function ρ in (2.2.4), 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 (2.2.7) and ℓ_0 is the prior log-likelihood function (2.2.5).

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

$$\theta^{(t+1)} = \frac{\alpha - 1 + \sum_{j \neq i} d_{i,j}}{\beta + \sum_{j \neq i} \frac{m_{i,j}}{\theta_i^{(t)} + \theta_j^{(t)}}}. \tag{2.2.9}$$

Note that this iterative optimization algorithm corresponds to the classic MM algorithm for ML estimation (2.2.8) when $\alpha = 1$ and $\beta = 0$.

We also consider *gradient descent algorithm* with constant step size $\eta > 0$, which has iterative updates as

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \nabla f(\mathbf{x}^{(t)}). \tag{2.2.10}$$

Our goal in this chapter is to summarise the rate of convergence of gradient descent and MM algorithms for generalized Bradley-Terry models that we derived in our previous work and then propose new accelerated algorithms to avoid the observed slow convergence issue for MAP estimation. It is natural to consider gradient descent algorithms as they belong to the class of first-order optimization methods (not requiring second-order quantities such as Hessian of the objective function or its approximations). Intuitively, the rate of convergence of an iterative algorithm quantifies how fast the value of the objective function converges to the optimum value with the number of iterations.

For an iterative optimization method for minimizing function f, which outputs a sequence of points $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$, we say that there is *an α-improvement with respect to f at time step t* if

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^\star) \leq (1 - \alpha)(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star))$$

where $\mathbf{x}^\star$ is a minimizer of f. An iterative optimization method is said to have *linear convergence* if there exist positive constants α and t_0 such that the method yields an α-improvement at each time step $t \geq t_0$.

Background on convex analysis We define some basic concepts from convex analysis that we will use throughout the chapter.

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

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

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

Function f is μ-*smooth* on $\mathcal{X}$ if its gradient ∇f is μ-Lipschitz on $\mathcal{X}$, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \leq \mu ||\mathbf{x} - \mathbf{y}||.$$

For any μ-smooth function f on $\mathcal{X}$, we also have that, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \leq \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2. \tag{2.2.11}$$

A proof of the last claim can be found in Lemma 3.4 in Bubeck [2015].

Function f satisfies *the Polyak-Lojasiewicz inequality* on $\mathcal{X}$ (Polyak [1963]) if there exists $\gamma > 0$ such that for all $\mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\gamma}||\nabla f(\mathbf{x})||^2 \tag{2.2.12}$$

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 $\mathcal{X}$.

2.3 Prior results on convergence rates

In this section, we summarise our prior results on the rate of convergence for gradient descent and MM algorithms for ML and MAP estimation for the Bradley-Terry model of paired comparisons from Vojnovic et al. [2020]. We first show some 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 used to define the MM algorithm. These results are then used to derive convergence rate bounds for the Bradley-Terry model.

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 main steps as for the Bradley-Terry model of paired comparisons. The differences lie in the characterization of the strong-convexity and smoothness parameters. The resulting 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 Section 2.7.8.

2.3.1 General convergence theorems

We first present a well-known result on the convergence rate of a gradient descent algorithm, for the reader's convenience. A result of this type can be found in Nesterov [2013] and a simple proof can be found in Chapter 9.3 of Boyd and Vandenberghe [2004].

Theorem 2.3.1 (gradient descent)**.** *Assume f is a convex μ-smooth function on $\mathcal{X}_\mu$ satisfying the γ-PL inequality on $\mathcal{X}_\gamma \subseteq \mathcal{X}_\mu$, $\mathbf{x}^* \in \mathcal{X}_\gamma$ is a minimizer of f, and $\mathbf{x}^{(t)} \mapsto \mathbf{x}^{(t+1)}$ is according to the gradient descent algorithm (2.2.10) with step size $\eta = 1/\mu$.*

Then, if $\mathbf{x}^{(t)} \in \mathcal{X}_\gamma$ and $\mathbf{x}^{(t+1)} \in \mathcal{X}_\mu$, there is an γ/μ-improvement with respect to f at time step t.

Proof of Theorem 2.3.1 is provided in Section 2.7.10.1.

Note that if there exists $t_0 \geq 0$ such that $\mathbf{x}^{(t)} \in \mathcal{X}_\gamma$ for all $t \geq t_0$, then Theorem 2.3.1 implies a linear convergence rate with rate γ/μ. Such a t_0 indeed exists as it can be shown that $||\mathbf{x}^{(t)} - \mathbf{x}^*||$ is non-increasing in t and is decreasing for every t such that $||\nabla f(\mathbf{x}^{(t)})|| \neq 0$.

We next present a result in Vojnovic et al. [2020] which shows that the MM algorithm also has linear convergence, for any smooth and strongly convex function f that has a surrogate function g satisfying a certain condition.

Theorem 2.3.2 (MM)**.** *Assume f is a convex μ-smooth function on $\mathcal{X}_\mu$ satisfying the γ-PL inequality on $\mathcal{X}_\gamma \subseteq \mathcal{X}_\mu$, $\mathbf{x}^* \in \mathcal{X}_\gamma$ is a minimizer of f and $\mathbf{x}^{(t)} \mapsto \mathbf{x}^{(t+1)}$ is according to the MM algorithm*

(2.2.6). Let g be a majorant surrogate function of f such that for some $\delta > 0$,

$$g(\mathbf{x};\mathbf{y}) - f(\mathbf{x}) \leq \frac{\delta}{2}||\mathbf{x}-\mathbf{y}||^2 \text{ for all } \mathbf{x},\mathbf{y} \in \mathcal{X}_\mu.$$

Then, if $\mathbf{x}^{(t)} \in \mathcal{X}_\gamma$ and $\mathbf{x}^{(t)} - \frac{1}{\mu+\delta}\nabla f(\mathbf{x}^{(t)}) \in \mathcal{X}_\mu$, there is a $\gamma/(\mu+\delta)$-improvement with respect to f at time step t.

From Theorems 2.3.1 and 2.3.2, 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.

2.3.2 Maximum likelihood estimation

We consider the rate of convergence for the ML parameter estimation for the Bradley-Terry model of paired comparisons. This estimation problem amounts to finding a parameter vector that minimizes the negative log-likelihood function, with the log-likelihood function given in (2.2.3). Recall that $\mathbf{M}$ denotes the matrix of counts of item-pair co-occurrences and $\mathbf{L_M}$ denotes the corresponding Laplacian matrix. For any positive semidefinite matrix $\mathbf{A}$, we let $\lambda_i(\mathbf{A})$ denote the i-th smallest eigenvalue of $\mathbf{A}$.

Lemma 2.3.1. *For any $\omega \geq 0$, the negative log-likelihood function for the Bradley-Terry model of paired comparisons 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*

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

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

By Lemma 2.3.1, the smoothness parameter μ is proportional to the largest eigenvalue of the Laplacian matrix $\mathbf{L_M}$. By the 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 $\lambda_n(\mathbf{L_M}) \geq d(\mathbf{M}) + 1$ by 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 per item.

The following lemma will be useful for showing that a function f satisfies the γ-PL inequality if it satisfies a γ-strong convexity condition.

Lemma 2.3.2. *Assume that $\mathcal{X}$ is a convex set such that f is γ-strongly convex on $\mathcal{X}_0 = \mathcal{X} \cap \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{1} = 0\}$, and that for all $c \in \mathbb{R}$ and $\mathbf{x} \in \mathcal{X}$,*

(C1) $f(\Pi_c(\mathbf{x})) = f(\mathbf{x})$ *and*

(C2) $\nabla f(\Pi_c(\mathbf{x})) = \nabla f(\mathbf{x})$

where

$$\Pi_c(\mathbf{x}) = \mathbf{x} + c\mathbf{1}.$$

Then, f satisfies the γ-PL inequality on $\mathcal{X}$.

Since the negative log-likelihood function of the Bradley-Terry model of paired comparisons satisfies conditions (C1) and (C2) of Lemma 2.3.2, combining with Lemma 2.3.1, we observe that it satisfies the γ-PL inequality on $\mathcal{W}_\omega$ with $\gamma = c_\omega a(\mathbf{M})$. Furthermore, by Lemma 2.3.1, the negative log-likelihood function is μ-smooth on $\mathbb{R}^n$ with $\mu = d(\mathbf{M})/2$. Combining these facts with Theorem 2.3.1, we have the following corollary:

Corollary 2.3.1 (gradient descent)**.** *Assume that $\mathbf{w}^*$ is the maximum likelihood parameter estimate in $\mathcal{W}_\omega = \{\mathbf{w} : \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$, for some $\omega \geq 0$, and $\mathbf{w}^{(t)} \mapsto \mathbf{w}^{(t+1)}$ is according to gradient descent algorithm with step size $\eta = 2/d(\mathbf{M})$.*

Then, if $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$, there is an $\alpha_{\mathbf{M},\omega}$-improvement at time step t where

$$\alpha_{\mathbf{M},\omega} = 2c_\omega \frac{a(\mathbf{M})}{d(\mathbf{M})}.$$

The result in Corollary 2.3.1 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:

$$T = O\left(\frac{d(\mathbf{M})}{a(\mathbf{M})} \log\left(\frac{1}{\epsilon}\right)\right). \tag{2.3.1}$$

We next consider the classic MM algorithm for the ML estimation problem, which uses the surrogate function in (2.2.7). This surrogate function satisfies the following property:

Lemma 2.3.3. *For any $\omega \geq 0$, 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*

$$\delta = \frac{1}{2} e^{2\omega} d(\mathbf{M}).$$

By Theorem 2.3.2 and Lemmas 2.3.1, 2.3.2, and 2.3.3, we have the following corollary:

Corollary 2.3.2 (MM). *Assume that $\mathbf{w}^*$ is the maximum likelihood parameter estimate in $\mathcal{W}_\omega = \{\mathbf{w} : \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$, for some $\omega \geq 0$, and that $\mathbf{w}^{(t)} \mapsto \mathbf{w}^{(t+1)}$ is according to the MM algorithm.*

Then, if $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$, there is an $\alpha_{\mathbf{M},\omega}$-improvement with respect at time step t where

$$\alpha_{\mathbf{M},\omega} = 2c'_\omega \frac{a(\mathbf{M})}{d(\mathbf{M})}$$

and $c'_\omega = 1/[(e^{-\omega} + e^{\omega})^2(1 + e^{2\omega})]$.

From Corollaries 2.3.1 and 2.3.2, 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. Hence, both gradient descent and MM algorithm have a linear convergence, and the convergence time bound (2.3.1).

2.3.3 Maximum a posteriori probability estimation

We next consider the maximum a posteriori probability estimation problem. We first note that the negative log-a posteriori probability function has the following properties.

Lemma 2.3.4. *The negative log-a posteriori probability function for the Bradley-Terry model of paired comparisons and the prior distribution Gamma(α, β) is γ-strongly convex and μ-smooth on $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$ with*

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

Note that the strong convexity parameter γ is proportional to β while as shown in Lemma 2.3.1, for the ML objective γ is proportional to $\lambda_2(\mathbf{M})$. This has important implications on the rate of convergence which we discuss next.

By Theorem 2.3.1 and Lemma 2.3.4, we have the following corollary:

Corollary 2.3.3 (gradient descent)**.** *Assume* $\mathbf{w}^*$ *is the maximum a posteriori parameter estimate in* $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$*, for some* $\omega \geq 0$*, and* $\mathbf{w}^{(t)} \mapsto \mathbf{w}^{(t+1)}$ *is according to gradient descent algorithm (2.2.10) with step size* $\eta = 2/(d(\mathbf{M}) + 2\beta e^{\omega})$.

Then, if $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$*, there is an* $\alpha_{\mathbf{M},\omega}$*-improvement where*

$$\alpha_{\mathbf{M},\omega} = \frac{2e^{-\omega}\beta}{d(\mathbf{M}) + 2e^{\omega}\beta}.$$

The result in Corollary 2.3.3 implies a linear convergence with the convergence time bound

$$T = O\left(\left(1 + \frac{d(\mathbf{M})}{\beta}\right)\log\left(\frac{1}{\epsilon}\right)\right). \tag{2.3.2}$$

This bound can be arbitrarily large by taking parameter β to be small enough.

We next consider the MM algorithm. First, note that since $\underline{\rho}(\mathbf{x};\mathbf{y}) - \rho(\mathbf{x}) = \underline{\ell}(\mathbf{x};\mathbf{y}) - \ell(\mathbf{x})$, by Lemma 2.3.3, we have

Lemma 2.3.5. *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 2.3.2 and Lemmas 2.3.4 and 2.3.5, we have the following corollary:

Corollary 2.3.4 (MM)**.** *Assume* $\mathbf{w}^*$ *is the maximum a posteriori parameter estimate in* $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$*, for some* $\omega \geq 0$*, and* $\mathbf{w}^{(t)} \mapsto \mathbf{w}^{(t+1)}$ *is according to the MM algorithm.*

Then, if $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$*, there is an* $\alpha_{\mathbf{M},\omega}$*-improvement at time step* t *where*

$$\alpha_{\mathbf{M},\omega,\beta} = \frac{2e^{-\omega}\beta}{(1 + e^{2\omega})d(\mathbf{M}) + 2e^{\omega}\beta}.$$

From Corollaries 2.3.3 and 2.3.4, we observe that both gradient descent algorithm and MM algorithm have the rate of convergence bound $1 - \Omega(\beta/(\beta + d(\mathbf{M})))$, and hence both have linear convergence and both have the convergence time bound (2.3.2).

Note that Corollaries 2.3.1 to 2.3.4 rely on the radius parameter ω. For the gradient descent algorithm,

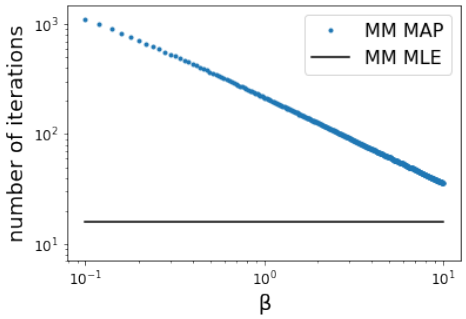


Figure 2.1: Simple illustrative example: number of iterations until convergence versus parameter β. Note that (a) the smaller the value of β, the slower the convergence for MAP and (b) the MM algorithm for MAP can be slower for several orders of magnitude than for ML.

$\mathbf{w}^{(t)} \in \mathcal{W}_\omega$ for all $t \geq 0$, given $\mathbf{w}^{(0)} = \mathbf{0}$ and $\omega = \|\mathbf{w}^*\|_\infty + \|\mathbf{w}^*\|_2$. This holds because $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$ is non-increasing in t and $\|\mathbf{w}^{(t)}\|_\infty \leq \|\mathbf{w}^*\|_\infty + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_\infty \leq \|\mathbf{w}^*\|_\infty + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$. For the MM algorithm, however, it seems hard to establish such theoretical bound for a sufficiently large radius. Instead, we empirically examine the radius ω for several different scenarios in Section 2.5. In all our experimental results, consistently, the radius ω is bounded and is scaled with $\|\mathbf{w}^*\|_\infty$.

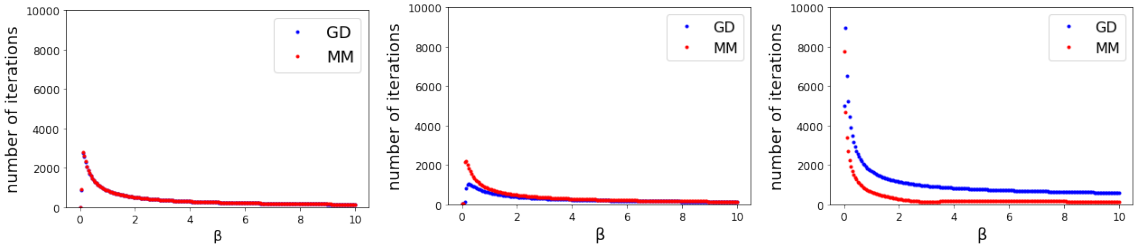
2.3.4 A simple illustrative numerical example

We illustrate the rate of convergence for a simple example, using randomly generated observations of paired comparisons. This allows us to demonstrate how the number of iterations grows as the value of parameter β becomes smaller, and how the number of iterations is affected by the value of parameter ω. Later, in Section 2.5, we provide further validation by using real-world datasets.

Our example is for 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 the parameter vector such that a half of items have parameter value $-\omega$ and the other half of items have parameter value ω, for a parameter $\omega > 0$. We define the convergence time T to be the smallest integer t such that $||\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}||_\infty \leq \xi$, for a fixed parameter $\xi > 0$. In our experiments, we set $\xi = 0.0001$.

The results in Figure 2.1, obtained for $\omega = 1/2$, demonstrate that the MM algorithm for the MAP estimation problem with $\beta > 0$ can be much slower than the MM algorithm for the ML estimation problem.

We further evaluate the convergence time of gradient descent and MM algorithms for different values of parameter ω, for each distinct pair of items compared 100 times. The numerical results in Figure 2.2



(a) $\omega = 0.01$ (b) $\omega = 1$ (c) $\omega = 5$

Figure 2.2: Number of iterations versus β for gradient descent and MM algorithms, for different values of ω.

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 the convergence time that increases with decreasing the value of β for strictly positive values of β. 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).

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: $f(x_1, x_2) = (x_1 - 1)^2 + c(x_2 - 1)^2$, for a parameter $c \geq 0$. Then, the gradient descent converges to the unique solution $(1, 1)$ slowly when c is close to 0. When $c = 0$, however, we just need to find the minimum point of $(x_1 - 1)^2$ which can be solved in a few iterations.

2.4 Accelerated MAP inference

In this section, we present a new accelerated algorithm for gradient descent and MM algorithms for MAP estimation. The key element is a transformation of the parameter vector estimate in each iteration of an iterative optimization algorithm that (a) ensures monotonic improvement of the optimization objective along the sequence of parameter vector estimates and (b) ensures certain second-order properties of the objective function hold along the sequence of parameter vector estimates.

We first introduce transformed versions of gradient descent and MM algorithms. Given a mapping $\Pi : \mathbb{R}^n \to \mathbb{R}^n$, we define the Π-*transformed gradient descent algorithm* by

$$\mathbf{x}^{(t+1)} = \Pi(\mathbf{x}^{(t)} - \eta f(\mathbf{x}^{(t)})). \tag{2.4.1}$$

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

$$\mathbf{x}^{(t+1)} = \Pi(\arg\min_{\mathbf{x}} g(\mathbf{x}; \mathbf{x}^{(t)})). \tag{2.4.2}$$

Importantly, function f and mapping Π have to satisfy certain conditions in order to provide a convergence rate guarantee, which we discuss in the following section.

2.4.1 General convergence theorems

Assume that f and Π satisfy the following conditions, for a convex set $\mathcal{X}$ that contains optimum point $\mathbf{x}^*$, and a vector $\mathbf{d} \in \mathbb{R}^n$:

(F1) f is μ-smooth on $\mathcal{X}$;

(F2) f satisfies the γ-PL inequality on

$$\mathcal{X}_0 = \mathcal{X} \cap \{\mathbf{x} \in \mathbb{R}^n : \nabla f(\mathbf{x})^\top \mathbf{d} = 0\} \tag{2.4.3}$$

and

(P1) $f(\Pi(\mathbf{x})) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$;

(P2) $\Pi(\mathbf{x}) \in \mathcal{X}_0$ when $\Pi(\mathbf{x}) \in \mathcal{X}$.

Condition (F1) is a standard smoothness condition imposed on $\mathcal{X}$. Condition (F2) is a standard γ-PL condition imposed on the subset of points in $\mathcal{X}$ at which the gradient of the function f is orthogonal to vector $\mathbf{d}$. Condition (P1) means that applying Π to a point cannot increase the value of function f. This condition is crucial to ensure a monotonic improvement of the objective function value when transformation Π is applied to an iterative optimization method. Condition (P2) is satisfied when at any Π-transformed point, the gradient of f is orthogonal to vector $\mathbf{d}$. This condition is crucial to ensure certain second-order properties hold when Π is applied to an iterative optimization method.

We have the following two theorems.

Theorem 2.4.1 (Gradient descent)**.** *Assume that f satisfies (F1) and (F2), Π satisfies (P1), and $\eta = 1/\mu$. Let $\mathbf{x}^{(t)} \mapsto \mathbf{x}^{(t+1)}$ be according to the Π-transformed gradient descent algorithm (2.4.1).*

Then, if $\mathbf{x}^{(t)}, \mathbf{x}^{(t+1)} \in \mathcal{X}_0$, there is an γ/μ-improvement with respect to f at time step t.

The theorem establishes the same rate of convergence as for the gradient descent algorithm in Theorem 2.3.1 but with the strong convexity condition restricted to points at which the gradient of f is orthogonal to vector $\mathbf{d}$.

Theorem 2.4.2 (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 $\mathbf{x}^{(t)} \mapsto \mathbf{x}^{(t+1)}$ be according to the Π-transformed MM algorithm (2.4.2).*

Then, if $\mathbf{x}^{(t)}, \mathbf{x}^{(t+1)} \in \mathcal{X}_0$, there is an $\gamma/(\gamma+\delta)$-improvement with respect to f at time step t.

The last theorem establishes the same rate of convergence as for classic MM algorithm in Theorem 2.3.2, but with a strong convexity condition imposed only at the points at which the gradient of f is orthogonal to vector $\mathbf{d}$.

We next present a lemma which will be instrumental in showing that the PL condition in (F1) holds for the MAP estimation problem.

Lemma 2.4.1. *Assume f is a convex, twice-differentiable function. Let $\mathcal{X}$ be a convex set, $\mathcal{X}_0$ be defined by (2.4.3) for a given vector $\mathbf{d}$, and $\mathbf{x}^* \in \mathcal{X}_0$ be a minimizer of f.*

If for some positive semidefinite matrix $\mathbf{A}_{\mathcal{X}}$,

(A1) $\nabla^2 f(\mathbf{x}) \succeq \mathbf{A}_{\mathcal{X}}$ *for all* $\mathbf{x} \in \mathcal{X}$, *and*

(A2) $\mathbf{u}^\top \mathbf{A}_{\mathcal{X}} \mathbf{v} = 0$ *for all* $\mathbf{u}, \mathbf{v}$ *such that*

$$\mathbf{u} = (\mathbf{I} - \mathbf{P}_{\mathbf{d}})\mathbf{z} \text{ and } \mathbf{v} = \mathbf{P}_{\mathbf{d}}\mathbf{z}, \text{ for } \mathbf{z} \in \mathbb{R}^n$$

where

$$\mathbf{P}_{\mathbf{d}} = \mathbf{I} - \frac{1}{||\mathbf{d}||^2}\mathbf{d}\mathbf{d}^\top$$

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

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

Note that $\mathbf{P}_{\mathbf{d}}$ is the projection matrix onto the space orthogonal to vector $\mathbf{d}$. The value of γ_0 is maximized when vector $\mathbf{d}$ is the eigenvector corresponding to the smallest eigenvalue of $\mathbf{A}_{\mathcal{X}}$. In this

case, γ_0 is the second smallest eigenvalue of $\mathbf{A}_{\mathcal{X}}$.

2.4.2 Convergence rate for the Bradley-Terry model

In this section, we apply the framework developed in the previous section to characterize the convergence rate for the MAP parameter estimation of the Bradley-Terry model of paired comparisons. The MAP parameter estimation problem amounts to finding a parameter vector that maximizes the log-a posteriori probability function ρ defined in (2.2.4). Let the transformation Π be defined as

$$\Pi(\mathbf{x}) = \mathbf{x} + c(\mathbf{x})\mathbf{1} \tag{2.4.4}$$

where

$$c(\mathbf{x}) = \log\left(\frac{\alpha - 1}{\beta} n\right) - \log\left(\sum_{i=1}^{n} e^{x_i}\right). \tag{2.4.5}$$

We next show that f and Π satisfy conditions (F1), (F2), (P1), and (P2) for the direction vector $\mathbf{d} = \mathbf{1}$. This will allow us to apply Theorems 2.4.1 and 2.4.2 to characterize the rate of convergence for Π-transformed gradient descent and MM algorithms.

We first show that f satisfies conditions (F1) and (F2) for the set $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$. Condition (F1) holds because, in Lemma 2.3.4, we have already shown that f is μ-smooth on $\mathcal{W}_\omega$ with $\mu = d(\mathbf{M})/2 + e^{\omega}\beta$. Condition (F2) can be shown to hold by Lemma 2.4.1 as follows. 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 2.4.1 hold: (A1) holds because $\mathbf{A}_{\mathcal{W}_\omega}$ 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 $\mathbf{u}^\top \mathbf{I}\mathbf{v} = \mathbf{u}^\top \mathbf{v} = 0$ ($\mathbf{u}$ and $\mathbf{v}$ are orthogonal). Since γ_0 is the smallest eigenvalue of $\mathbf{A}_{\mathcal{X}}$ on the subspace orthogonal to vector $\mathbf{1}$, we have

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

Hence, by Lemma 2.4.1, it follows that f satisfies condition (F2) with $\gamma = c_\omega \lambda_2(\mathbf{L}_\mathbf{M}) + e^{-\omega}\beta$.

We next show that Π, defined in (2.4.4), satisfies conditions (P1) and (P2). These two conditions are shown to hold in the following lemma.

Lemma 2.4.2. *For all* $\mathbf{w} \in \mathbb{R}^n$,

$$\rho(\Pi(\mathbf{w})) \geq \rho(\mathbf{w}) \tag{2.4.6}$$

and

$$\nabla\rho(\Pi(\mathbf{w}))^\top \mathbf{1} = 0. \tag{2.4.7}$$

From Theorem 2.4.1, we have the following corollary.

Corollary 2.4.1 (Gradient descent)**.** *Assume that* $\mathbf{w}^{(t)} \mapsto \mathbf{w}^{(t+1)}$ *is according to the* Π*-transformed gradient descent (2.4.1) for the negative log-a posteriori probability function of the Bradley-Terry model of paired comparisons, with product-form prior distribution such that* $e^{w_i} \sim$ *Gamma* (α, β)*,* $\alpha \geq 1$ *and* $\beta > 0$*, and* $\eta = 2/(||\mathbf{M}||_\infty + 2e^\omega\beta)$*.*

Then, there is an $\alpha_{\mathbf{M},\omega,\beta}$*-improvement with respect to* ρ *at time step* t *where*

$$\alpha_{\mathbf{M},\omega,\beta} = \frac{2c_\omega a(\mathbf{M}) + 2e^{-\omega}\beta}{d(\mathbf{M}) + 2e^\omega\beta}$$

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

From Theorem 2.4.2, we have the following corollary.

Corollary 2.4.2 (MM)**.** *Assume that iterates are according to the* Π*-transformed MM (2.4.1) for the negative log-a posteriori probability function of the Bradley-Terry model of paired comparisons, with product-form prior distribution such that* $e^{w_i} \sim$ *Gamma* (α, β)*,* $\alpha \geq 1$ *and* $\beta > 0$*.*

Then, there is an $\alpha_{\mathbf{M},\omega,\beta}$*-improvement with respect to* ρ *at time step* t *where*

$$\alpha_{\mathbf{M},\omega,\beta} = \frac{2c_\omega a(\mathbf{M}) + 2e^{-\omega}\beta}{(1 + e^{2\omega})d(\mathbf{M}) + 2e^\omega\beta}$$

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

Proof. 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 $\mathcal{W}_\omega$ and $\delta = e^{2\omega}d(\mathbf{M})/2$. Condition (F2) holds by Lemma 2.4.1 with $\gamma = c_\omega a(\mathbf{M}) + e^{-\omega}\beta$. Conditions (P1) and (P2) hold by (2.4.6) and (2.4.7) in Lemma 2.4.2, respectively. □

Note that in the limit of small β, the convergence rate bounds in Corollaries 2.4.1 and 2.4.2 correspond to the bounds for the ML estimation in Corollaries 2.3.1 and 2.3.2, respectively. From Corollaries 2.4.1 and 2.4.2, it follows that for accelerated gradient descent and accelerated MM algorithms, the

Algorithm 1 Accelerated MM algorithm

Initialization: $\epsilon, \theta, \theta^{\text{prev}}$
while $||\theta - \theta^{\text{prev}}||_\infty > \epsilon$ **do**
  $\theta^{\text{prev}} \leftarrow \theta$
  for $i = 1, 2, \ldots n$ **do**
    $\theta_i^{\text{temp}} = \frac{\alpha - 1 + \sum_{j \neq i} d_{i,j}}{\beta + \sum_{j \neq i} \frac{m_{i,j}}{\theta_i + \theta_j}}$ ▷ standard MM
  end for
  for $i = 1, 2, \ldots, n$ **do**
    $\theta_i = \frac{\theta_i^{\text{temp}}}{\sum_{j=1}^n \theta_j^{\text{temp}}} \frac{\alpha - 1}{\beta} n$ ▷ rescaling
  end for
end while

convergence time satisfies

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

For the Bradley-Terry model of paired comparisons with parametrization $\theta = (\theta_1, \ldots, \theta_n)^\top$, where $\theta_i = e^{w_i}$ for $i = 1, 2, \ldots, n$, the transformation Π given by (2.4.4) is equivalent to a *rescaling* as shown in a procedural form in Algorithm 1. This algorithm first performs the standard MM update in Eq. (2.2.9), which is followed by rescaling the resulting intermediate parameter vector such that the parameter vector θ at every iteration satisfies $\sum_{i=1}^n \theta_i = c$ where $c = n(\alpha - 1)/\beta$. This can be interpreted as fixing the scale of parameters to a carefully chosen scale that is dependent on the choice of the prior distribution. Note that the scaling factor c cannot be arbitrarily fixed while still preserving good convergence properties. In particular, selecting the scale $c = 1$ can result in undesired convergence properties. We demonstrate this in Section 2.5 by numerical examples.

It turns out that the rescaling in Algorithm 1 is roughly of the same order as the random rescaling suggested in Caron and Doucet [2012a]. Therein, the authors suggested using independent identically distributed random rescaling factors across different iteration steps with distribution Gamma$(n\alpha, \beta)$. This ensures the distribution of $\sum_{i=1}^n \theta_i$ to remain invariant across different iterations, equal to Gamma$(n\alpha, \beta)$. The mode of this rescaling factor is $(n\alpha - 1)/\beta$. This bears a similarity with the rescaling in Algorithm 1, in particular, with respect to the dependence on parameter β. Our results show that it suffices to use a simple deterministic rescaling factor to ensure linear convergence. Moreover, using a different rescaling than the one used in Algorithm 1 can result in a lack or slow convergence, which is shown by numerical experiments in Section 2.5.

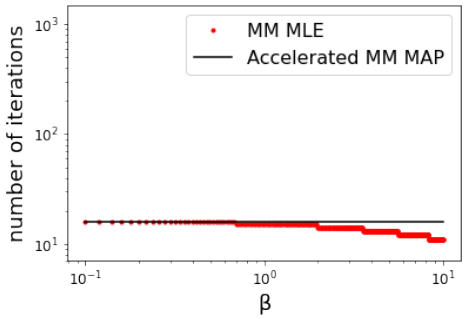


Figure 2.3: The illustrative example revisited: accelerated MM resolves the convergence issue for MAP estimation: it has faster or equal convergence than for ML estimation.

Our numerical example revisited We ran the accelerated MM algorithm for our numerical example and obtained the results shown in Figure 2.3. By comparing with the corresponding results obtained by the MM algorithm with no acceleration, shown in Figure 2.1, we observe that the acceleration resolves the slow convergence issue and that it can yield a significant reduction of the convergence time.

2.5 Numerical results

In this section we present evaluation of convergence times of gradient descent and MM algorithms for different generalized Bradley-Terry models for a collection of 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 2.4. The code and datasets for reproducing our experiments are available online at: `https://github.com/GDMMBT/AcceleratedBradleyTerry`.

2.5.1 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 Rich et al. [2018]. This service presents a user with a pair of images and asks to select one that better expresses a given metric, or select none. The dataset contains 1,048,576 observations and covers 17 metrics. We used this dataset to evaluate convergence of MM algorithms for the Bradley-Terry model of paired comparisons. We did this for each of 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 ratings competition Sonas [2010]. 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, consisting of a ranking of drivers in increasing order of their race finish times. The dataset is available from a 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] and more recently by Caron and Doucet [2012a]. We used this dataset to evaluate convergence times of MM algorithms for the Plackett-Luce ranking model.

Table 2.1: 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

We summarise some key statistics for each dataset in Table 2.1. We use the 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 at least one item in one group is not compared with any item in the other group, i.e., the algebraic connectivity $a(\mathbf{M})$ of matrix $\mathbf{M}$ is zero. In this case, there exists no ML estimate, while an MAP

estimate always exists. In order to consider cases when an MLE exists, we consider sampled datasets by restricting to the set of items such that the algebraic connectivity for this subset of items is strictly positive. This subsampling was done by selecting the largest connected component of items.

2.5.2 Experimental results

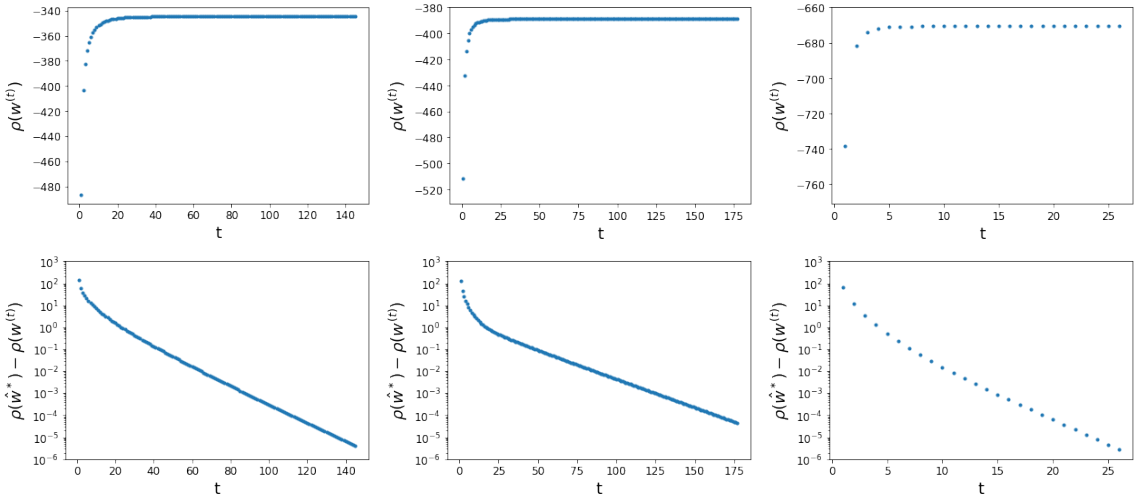
We evaluated the convergence time defined as the number of iterations that an algorithm takes until a convergence criteria is satisfied. We use the standard convergence criteria based on the difference of successive parameter vector estimates. Specifically, the convergence time T is defined as the smallest integer $t > 0$ such that $||\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}||_\infty \leq \xi$, for fixed value of parameter $\xi > 0$, with initial value $\mathbf{w}^{(0)} = \mathbf{0}$. In our experiments, we used 10^{-4} as the default value for parameter ξ. For NASCAR dataset, we also present results for several other values of ξ to demonstrate how the convergence time changes. In our experiments, we also evaluated the convergence time measured in real processor time units. We noted that they validate all the observations derived from the convergence times measured in the number of iterations, and hence we do not further discuss them.

In our experiments we varied the value of parameter β and, unless specified otherwise, we set the value of parameter α such that $\alpha - 1 = \beta$. This corresponds to fixing the mode of the Gamma prior marginal distributions to value 1. Note that the case $\beta = 0$ corresponds to ML estimation.

Before discussing numerical convergence time results, we first show results validating that the MM algorithm converges and that this convergence is linear. This is shown in Figure 2.4 for GIFGIF A dataset for three different values of parameter β. For space reasons, we only include results for this dataset. We observe that in all cases the log-a posteriori probability monotonically increases with the number of iterations, thus validating convergence. We also observe that the gap between the maximum log-a posteriori probability and the log-a posteriori probability decreases with the number of iterations in a linear fashion for sufficiently large number of iterations, when plotted using the logarithmic scale for the y axis, thus validating linear convergence.

We next discuss our numerical results for convergence time evaluated for the MM algorithm and accelerated MM algorithm for different datasets and choice of parameters. Our numerical results are shown in Table 2.2.

For GIFGIF datasets, we observe that the convergence time increases as the value of parameter β decreases for $\beta > 0$. For the values of β considered, this increase can be for as much as two orders



(a) $\beta = 0$ (b) $\beta = 0.1$ (c) $\beta = 1$

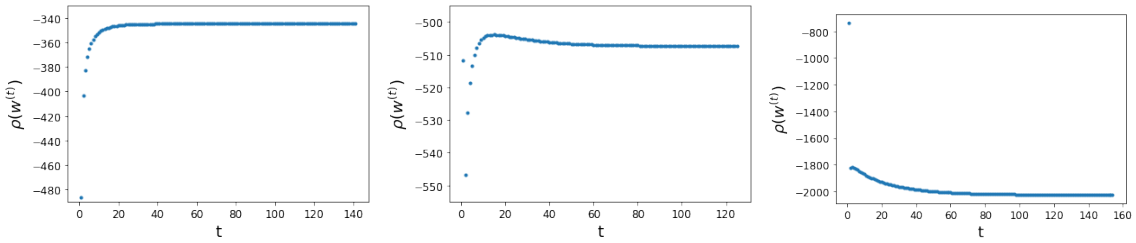
Figure 2.4: Convergence of the MM algorithm with the input GIFGIF A training dataset for different values of parameter β: (top) log-a posteriori probability versus the number of iterations, and (bottom) the difference between the maximum log-a posteriori probability and the log a-posteriori probability versus the number of iterations. The maximum a posteriori probability is approximated by taking the parameter vector $\hat{\mathbf{w}}^*$ output by the MM algorithm after a large number of iterations. The plots in the top row indicate that the algorithm converges. The plots in the bottom row indicate linear convergence. From the plots, we also observe that the convergence is slower for smaller values of parameter β.

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 for 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 order 10% of the convergence time of the MM algorithm without acceleration. These empirical results validate our theoretical results.

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

For NASCAR dataset, we show results for different values of parameter ξ, 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 pronounced for smaller values of ξ. For the cases considered, this can be for as much as three orders of magnitude. Similarly, the accelerated MM algorithm converges much faster than the classical MM algorithm. We also compare the quality of estimates obtained from the classical MM algorithm and the accelerated MM algorithm through a

toy simulation, which is included in Section 2.7.9 due to space constraints.



(a) $\beta = 0$ (b) $\beta = 0.1$ (c) $\beta = 1$

Figure 2.5: The log-a posteriori probability versus the number of iterations for the MM algorithm with normalization such that $\sum_{i=1}^{n} e^{w_i} = 1$ at each iteration, using GIFGIF A (sample) as input dataset, for different values of β. The results indicate that the log-a posteriori probability is not guaranteed to monotonically increase with the number of iterations.

Table 2.2: Number of iterations for the MM algorithm and accelerated MM algorithm (AccMM).

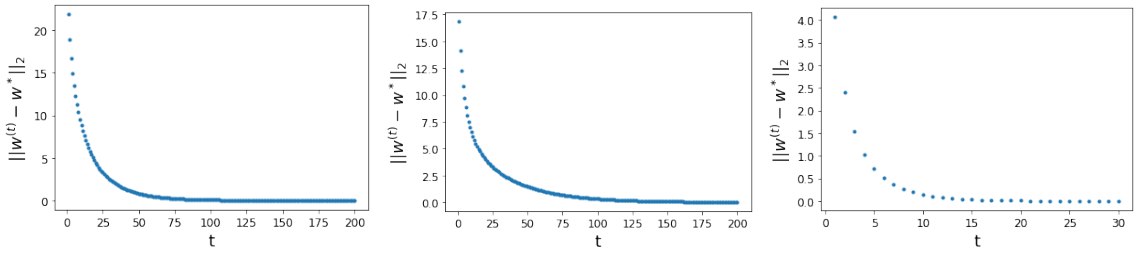
Dataset	Algorithm	$\beta = 0$	0.01	0.1	1	10
GIFGIF: A (full)	MM	MLE	572	125	70	16
	AccMM	non-existant	509	123	42	13
GIFGIF: C (full)	MM	MLE	733	150	49	13
	AccMM	non-existant	551	93	37	13
GIFGIF: H (full)	MM	MLE	1,127	149	98	21
	AccMM	non-existant	1,044	159	51	18
GIFGIF: A (sample)	MM	145	854	177	26	7
	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-existant	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 ($\xi = 10^{-5}$)	MM	14	1,528	2,069	105	16
	AccMM		14	14	12	**7**
NASCAR ($\xi = 10^{-6}$)	MM	17	2,362	3,223	157	23
	AccMM		17	16	14	8
NASCAR ($\xi = 10^{-8}$)	MM	22	4,029	5,544	261	36
	AccMM		22	21	18	11

We next discuss the importance of carefully changing the scale of the parameter vector in each iteration, as done in our accelerated MM algorithm, Algorithm 1, as otherwise the monotonic convergence may not be guaranteed or the convergence may be slow. To demonstrate this, we examine the alternative

Table 2.3: Number of iterations for the MM algorithm with normalization such that $\sum_{i=1}^{n} e^{w_i} = 1$ in each iteration.

Dataset	Algorithm	$\beta = 0$	0.01	0.1	1	10
GIFGIF: A (sample)	MM norm	141	136	125	154	No convergence
GIFGIF: C (sample)	MM norm	128	130	144	350	No convergence
GIFGIF: H (sample)	MM norm	205	203	184	124	No convergence

change of scale such that the parameter vector $\mathbf{w}$ in each iteration satisfies $\sum_{i=1}^{n} e^{w_i} = 1$. We present the results for the GIFGIF (sample) datasets. From Figure 2.5, we observe that the algorithm does not guarantee a monotonic increase of the a posteriori probability with the number of iterations, which is unlike to our accelerated MM algorithm for which this always holds. In Table 2.3, we show the same quantities as in Table 2.2 but for the MM algorithm with the alternative change of scale under consideration. We observe that our acceleration method can converge much faster, and that there are cases for which the alternative change of scale results in no convergence within a bound on the maximum number of iterations.



(a) $\beta = 0$ (b) $\beta = 0.1$ (c) $\beta = 1$

Figure 2.6: Euclidean distance between the MM algorithm parameter vector estimator and the true parameter vector versus the number of iterations, for GIFGIF:A (Sample) dataset. The results demonstrate that the distance monotonically decreases with the number of iterations.

Finally, we demonstrate that the distance between the parameter vector estimator of the MM algorithm and the true parameter vector decreases with the number of iterations. We show this in Figure 2.6 for different values of parameter β and GIFGIF:A (sample) dataset. We also verified that the monotonicity holds for other values of parameter β and other datasets, which is not shown for space reasons.

2.6 Further discussion

We have shown that for generalized Bradley-Terry models, gradient descent and MM algorithms for the ML estimation problem have a linear convergence with the convergence time bound $O(d(\mathbf{M})/a(\mathbf{M}))$,

Table 2.4: Parameters $d(\mathbf{M})$ and $a(\mathbf{M})$ when $\mathbf{M}$ 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	$2\left(1-\cos\left(\frac{\pi}{n}\right)\right) \sim \frac{4\pi^2}{n^2}$	$\Theta(n^2)$
path	2	$2\left(1-\cos\left(\frac{2\pi}{n}\right)\right) \sim \frac{\pi^2}{n^2}$	$\Theta(n^2)$

where $d(\mathbf{M})$ is the maximum number of observed comparisons per item and $a(\mathbf{M})$ is the algebraic connectivity of the matrix $\mathbf{M}$ of the observed counts of item-pair co-occurrences. We have also shown that for generalized Bradley-Terry models, gradient descent and MM algorithms for the MAP estimation problem, with the prior product-form distribution with Gamma(α, β) marginal distributions, the convergence time is also linear but with the convergence time bound $O(d(\mathbf{M})/\beta)$. This bound is shown to be tight. Our results identify a slow convergence issue for gradient descent and MM algorithms for the MAP estimation problem, which occurs for small values of parameter β. The small values of parameter β correspond to more vague prior distributions. Our results identify a discontinuity of the convergence time at $(\alpha, \beta) = (1, 0)$, which corresponds to ML estimation. The proposed acceleration method for the MAP estimation problem resolves the slow convergence issue, and yields a convergence time that is bounded by the best of what can be achieved for the ML and MAP estimation problems.

Our results provide insights into how the observed comparison data affect the rate of convergence of gradient descent and MM algorithms. The two key parameters affecting the rate of convergence are $d(\mathbf{M})$ and $a(\mathbf{M})$. For illustration purposes, in Table 2.4 we show values of $d(\mathbf{M})$ and $a(\mathbf{M})$ for examples of matrix $\mathbf{M}$ with 0-1 valued entries, which correspond to graph adjacency matrices. We observe that when each distinct pair is compared the same number of times, i.e. for the complete graph case, the convergence time is $T = O(\log(1/\epsilon))$. For other cases, the convergence time is $T = O(n^c \log(1/\epsilon))$, for some $c \geq 1$.

We further consider the case of random design matrices where each distinct pair of items is either compared once or not compared at all, and this is according to independent Bernoulli random variables with parameter p across all distinct pairs of items. In other words, the item pair co-occurrence is according to the Erdös-Rényi random graph $G_{n,p}$ and $\mathbf{M}$ is its adjacency matrix. $d(\mathbf{M})$ corresponds to the maximum degree of $G_{n,p}$ which has been extensively studied, with precise results obtained for different scalings of p with n. In particular, by Bollobás [2001] (Corollary 3.4), $d(\mathbf{M}) = pn +$

$O(\sqrt{pn\log(n)})$ with probability $1-1/n$ provided that $p=\omega(\log(n)^3/n)$. The algebraic connectivity for Erdös-Rényi graphs has been studied as well. By Coja-Oghlan [2007] (Theorem 1.3), $a(\mathbf{M}) = pn + O(\sqrt{pn\log(n)})$ with probability $1-o(1)$, provided that $p=\omega(\log(n)^2/n)$. Intuitively, if the expected degree np is large enough, $d(\mathbf{M})/a(\mathbf{M}) = \Theta(1)$.

We can derive an upper bound for the convergence time, which depends only on some simple properties of the graph associated with matrix $\mathbf{M}$. Let $\mathbf{A}$ be the adjacency matrix of a graph G which has edge (i,j) if, and only if, $m_{i,j}>0$. Let $r=\overline{m}/\underline{m}$ where $\overline{m}=\max_{i,j} m_{i,j}$ 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 and MM algorithms for the ML estimation, we have the convergence time bound (shown in Section 2.7.7):

$$T = O(rd(n)D(n)n\log(1/\epsilon)). \tag{2.6.1}$$

This implies that $T=O(rn^3\log(1/\epsilon))$ for every connected graph G, which follows by trivial facts $d(n)\leq n$ and $D(n)\leq n$. The bound in (2.6.1) follows from the lower bound on the algebraic connectivity of a Laplacian matrix $\lambda_2(\mathbf{L_A}) \geq 4/(nD(n))$, see Theorem 3.4 in Merris [1994].

2.7 Proofs and additional results

2.7.1 Comparison of Theorem 2.3.2 with Proposition 2.7 in Mairal [2015]

Theorem 2.7.1 (Proposition 2.7 in Mairal [2015])**.** *Suppose that f is a strongly convex function on $\mathcal{X}_\gamma$ and $\mathbf{x}^*$ is a minimizer of f and that it holds $\mathbf{x}^*\in\mathcal{X}_\gamma$. Assume that g is a first-order surrogate function of f on $\mathcal{X}_\mu$ with parameter $\mu_0>0$. Let $\mathbf{x}^{(t+1)}$ be the output of the MM algorithm for input $\mathbf{x}^{(t)}$. Then, if $\mathbf{x}^{(t)}\in\mathcal{X}_\gamma$ and $\mathbf{x}^{(t+1)}\in\mathcal{X}_\mu$, then we have*

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

where

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

Proof. Proof If g is a first-order surrogate function on $\mathcal{X}_\mu$ with parameter μ_0, then

$$f(\mathbf{x}') \leq f(\mathbf{z}) + \frac{\mu_0}{2}||\mathbf{z} - \mathbf{y}||^2$$

where $\mathbf{x}' = \arg\min_{\mathbf{z}'} g(\mathbf{z}'; \mathbf{y})$.

From this, it follows that

$$\begin{aligned} & f(\mathbf{x}') \\ \leq & \min_{\mathbf{z}} \left\{ f(\mathbf{z}) + \frac{\mu_0}{2}||\mathbf{z} - \mathbf{x}^*||^2 \right\} \\ \leq & \min_{a \in [0,1]} \left\{ f(a\mathbf{x}^* + (1-a)\mathbf{x}) + \frac{\mu_0 a^2}{2}||\mathbf{x} - \mathbf{x}^*||^2 \right\} \\ \leq & \min_{a \in [0,1]} \left\{ af(\mathbf{x}^*) + (1-a)f(\mathbf{x}) + \frac{\mu_0 a^2}{2}||\mathbf{x} - \mathbf{x}^*||^2 \right\} \end{aligned}$$

where the last inequality is by convexity of f.

We have established the following inequality

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ \leq & \min_{a \in [0,1]} \left\{ (1-a)(f(\mathbf{x}) - f(\mathbf{x}^*)) + \frac{\mu_0 a^2}{2}||\mathbf{x} - \mathbf{x}^*||^2 \right\}. \end{aligned}$$

By assumption that f is γ-strongly convex on $\mathcal{X}_\gamma$ and $\mathbf{x} \in \mathcal{X}_\gamma$, we have

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\gamma}{2}||\mathbf{x} - \mathbf{x}^*||^2.$$

It follows that

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ \leq & \min_{a \in [0,1]} \left\{ 1 - a + \frac{\mu_0 a^2}{\gamma} \right\} (f(\mathbf{x}) - f(\mathbf{x}^*)). \end{aligned}$$

It remains only to note that

$$\min_{a \in [0,1]} \left\{ 1 - a + \frac{\mu_0 a^2}{\gamma} \right\} = c.$$

□

The rate of convergence bound derived from Theorem 2.3.2 can be tighter than the rate of convergence bound derived from Theorem 2.7.1.

To show this consider the Bradley-Terry model for which we have shown in Lemma 2.3.3 that the surrogate function $\underline{\ell}$ of the log-likelihood function ℓ satisfies condition of Theorem 2.3.2 on $[-\omega, \omega]^n$ with $\delta = \frac{1}{2}e^{2\omega}d(\mathbf{M})$. It also holds that surrogate function $\underline{\ell}$ is also a first-order surrogate function of ℓ on $[-\omega, \omega]^n$ with $\mu_0 = \frac{1}{2}e^{2\omega}d(\mathbf{M})$. Hence in this case, we have $\delta = \mu_0$.

The convergence rate bound of Theorem 2.3.2 is tighter than the convergence rate bound of Theorem 2.7.1 if and only if $\mu + \delta < 4\mu_0$. Since $\delta = \mu_0$, this is equivalent to $\mu < 3\delta$. Since by Lemma 2.3.1 we can take $\mu = \frac{1}{2}d(\mathbf{M})$, the latter condition reads as

$$1 < 3e^{\omega}$$

which indeed holds true.

2.7.2 Surrogate function (2.2.7) for the Bradley-Terry model is a first-order surrogate function

We show that the surrogate function $\underline{\ell}$ of the log-likelihood function ℓ of the Bradley-Terry model, given by (2.2.7), is a first-order surrogate function on $\mathcal{X}_\omega = [-\omega, \omega]^n$ with $\mu_0 = \frac{1}{2}e^{2\omega}d(\mathbf{M})$.

We need to show that the error function $h(\mathbf{x}; \mathbf{y}) = \ell(\mathbf{x}) - \underline{\ell}(\mathbf{x}; \mathbf{y})$ is a μ_0-smooth function on $\mathcal{X}_\omega$.

By a straightforward calculus, we note

$$\nabla^2 h(\mathbf{x}; \mathbf{y}) = \nabla^2 \ell(\mathbf{x}) + D(\mathbf{x}, \mathbf{y})$$

where $D(\mathbf{x}, \mathbf{y})$ is a diagonal matrix with diagonal elements

$$d_u = \sum_{j \neq u} m_{u,j} \frac{e^{x_u}}{e^{y_u} + e^{y_j}}.$$

We can take

$$\mu_0 = \max_{\mathbf{x}, \mathbf{y} \in \mathcal{X}_\omega} \max\{|\lambda_1(\nabla^2 h(\mathbf{x}; \mathbf{y}))|, |\lambda_n(\nabla^2 h(\mathbf{x}; \mathbf{y}))|\}.$$

For any $A = B + D$ where B is a $n \times n$ matrix and D is a $n \times n$ diagonal matrix with diagonal

elements $d_1, d_2, \ldots, d_n$, we have

$$\lambda_1(B) + \min_u d_u \leq \lambda_i(A) \leq \lambda_n(B) + \max_u d_u.$$

It thus follows that

$$\begin{aligned} \mu_0 \quad \leq \quad & \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}_\omega} \max\{|\lambda_1(\nabla^2\ell(\mathbf{x}))| \\ & + \min_u d_u|, |\lambda_n(\nabla^2\ell(\mathbf{x})) + \max_u d_u|\}. \end{aligned}$$

Now note that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}_\omega$,

$$-\frac{1}{2}d(\mathbf{M}) \leq \lambda_1(\nabla^2\ell(\mathbf{x})) \leq \lambda_n(\nabla^2\ell(\mathbf{x})) = 0$$

and

$$\frac{1}{2}e^{-2\omega}\min_u \sum_{j\in u} m_{u,j} \leq \min_u d_u \leq \max_u d_u \leq \frac{1}{2}e^{2\omega}d(\mathbf{M}).$$

We have

$$|\lambda_n(\nabla^2\ell(\mathbf{x})) + \max_u d_u| = \max_u d_u \leq \frac{1}{2}e^{2\omega}d(\mathbf{M})$$

and

$$\begin{aligned} & |\lambda_1(\nabla^2\ell(\mathbf{x})) + \min_u d_u| \\ = \quad & (\lambda_1(\nabla^2\ell(\mathbf{x})) + \min_u d_u)\mathbb{I}_{\lambda_1(\nabla^2\ell(\mathbf{x}))+\min_u d_u\geq 0} \\ & +(-\lambda_1(\nabla^2\ell(\mathbf{x})) - \min_u d_u)\mathbb{I}_{\lambda_1(\nabla^2\ell(\mathbf{x}))+\min_u d_u<0} \\ \leq \quad & \min_u d_u\mathbb{I}_{\lambda_1(\nabla^2\ell(\mathbf{x}))+\min_u d_u\geq 0} \\ & -\lambda_1(\nabla^2\ell(\mathbf{x}))\mathbb{I}_{\lambda_1(\nabla^2\ell(\mathbf{x}))+\min_u d_u<0} \\ \leq \quad & \frac{1}{2}e^{2\omega}d(\mathbf{M})\mathbb{I}_{\lambda_1(\nabla^2\ell(\mathbf{x}))+\min_u d_u\geq 0} \\ & +\frac{1}{2}d(\mathbf{M})\mathbb{I}_{\lambda_1(\nabla^2\ell(\mathbf{x}))+\min_u d_u<0} \\ \leq \quad & \frac{1}{2}e^{2\omega}d(\mathbf{M}). \end{aligned}$$

2.7.3 Proof of Theorem 2.4.1

Since $f(\Pi(\mathbf{x})) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} f(\mathbf{x}^{(t+1)}) &= f(\Pi(\mathbf{x}^{(t)} - \eta \nabla f(\mathbf{x}^{(t)}))) \\ &\leq f(\mathbf{x}^{(t)} - \eta \nabla f(\mathbf{x}^{(t)})). \end{aligned}$$

By the same steps as those in the proof of Theorem 2.3.1, we can show that

$$\begin{aligned} & f(\mathbf{x}^{(t)} - \eta \nabla f(\mathbf{x}^{(t)})) - f(\mathbf{x}^*) \\ \leq & \left(1 - \frac{\gamma}{\mu}\right) (f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)). \end{aligned}$$

Hence, it follows that

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

2.7.4 Proof of Lemma 2.4.1

By a limited Taylor expansion, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\begin{aligned} & f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\ & + \tfrac{1}{2} \min_{a \in [0,1]} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(a\mathbf{y} + (1-a)\mathbf{x})(\mathbf{y} - \mathbf{x}). \end{aligned} \tag{2.7.1}$$

Let

$$\mathbf{u} = (\mathbf{I} - \mathbf{P_d})(\mathbf{y} - \mathbf{x}) \text{ and } \mathbf{v} = \mathbf{P_d}(\mathbf{y} - \mathbf{x})$$

where

$$\mathbf{P_d} = \mathbf{I} - \frac{1}{||\mathbf{d}||^2} \mathbf{d}\mathbf{d}^\top.$$

Notice that

(i) $\mathbf{u} + \mathbf{v} = \mathbf{y} - \mathbf{x}$, and

(ii) $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, i.e., $\mathbf{u}^\top \mathbf{v} = 0$.

From now on, assume that $\mathbf{x}$ and $\mathbf{y}$ are such that $\mathbf{x}, \mathbf{y} \in \mathcal{X}_0$ and $\mathbf{y} = \mathbf{x}^*$.

By definition of $\mathcal{X}_0$, we have $\mathbf{d}^\top \nabla f(\mathbf{x}) = 0$, which together with $\mathbf{u} + \mathbf{v} = \mathbf{x}^* - \mathbf{x}$, implies

$$\nabla f(\mathbf{x})^\top (\mathbf{x}^* - \mathbf{x}) = \nabla f(\mathbf{x})^\top \mathbf{v}. \tag{2.7.2}$$

Now, note that for any $a \in [0, 1]$, we have the following relations:

$$\begin{aligned}
& (\mathbf{x}^* - \mathbf{x})^\top \nabla^2 f(a\mathbf{x}^* + (1-a)\mathbf{x})(\mathbf{x}^* - \mathbf{x}) \\
= \; & (\mathbf{u} + \mathbf{v})^\top \nabla^2 f(a\mathbf{x}^* + (1-a)\mathbf{x})(\mathbf{u} + \mathbf{v}) \\
\overset{(a)}{\geq} \; & (\mathbf{u} + \mathbf{v})^\top \mathbf{A}_{\mathcal{X}} (\mathbf{u} + \mathbf{v}) \\
\overset{(b)}{\geq} \; & \mathbf{v}^\top \mathbf{A}_{\mathcal{X}} \mathbf{v} \\
\geq \; & \left(\min_{\mathbf{y}: \mathbf{d}^\top \mathbf{y} = 0} \frac{\mathbf{y}^\top \mathbf{A}_{\mathcal{X}} \mathbf{y}}{||\mathbf{y}||^2} \right) ||\mathbf{v}||^2 \\
\geq \; & \gamma ||\mathbf{v}||^2
\end{aligned}$$

where (a) is by assumption (A1) and (b) is by assumption that $\mathbf{A}_{\mathcal{X}}$ is a positive semidefinite matrix and (A2). Hence, we have shown that, for all $a \in [0, 1]$,

$$(\mathbf{x}^* - \mathbf{x})^\top \nabla^2 f(a\mathbf{x}^* + (1-a)\mathbf{x})(\mathbf{x}^* - \mathbf{x}) \geq \gamma ||\mathbf{v}||^2. \tag{2.7.3}$$

Next, note that

$$\begin{aligned}
& \nabla f(\mathbf{x})^\top \mathbf{v} + \frac{1}{2}\gamma ||\mathbf{v}||^2 \\
\geq \; & \min_{\mathbf{z} \in \mathbb{R}^n} \left(\nabla f(\mathbf{x})^\top \mathbf{z} + \frac{1}{2}\gamma ||\mathbf{z}||^2 \right) \\
\geq \; & -\frac{1}{2\gamma} ||\nabla f(\mathbf{x})||^2.
\end{aligned}$$

Combining with (2.7.1)-(2.7.3), we obtain

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\gamma} ||\nabla f(\mathbf{x})||^2.$$

2.7.5 Proof of Lemma 2.4.2

Proof of (2.4.6) Since $\ell(\mathbf{w}) = \ell(\Pi(\mathbf{w}))$ for all $\mathbf{w} \in \mathbb{R}^n$, we have that $\rho(\Pi(\mathbf{w})) \geq \rho(\mathbf{w})$ is equivalent to $\ell_0(\Pi(\mathbf{w})) \geq \ell_0(\mathbf{w})$.

Now, note

$$\begin{aligned}
& \ell_0(\Pi(\mathbf{w})) - \ell_0(\mathbf{w}) \\
= \; & \ell_0(\mathbf{w} + c(\mathbf{w})\mathbf{1}) - \ell_0(\mathbf{w}) \\
= \; & (\alpha - 1)nc(\mathbf{w}) - \beta e^{c(\mathbf{w})} \sum_{i=1}^{n} e^{w_i} + \beta \sum_{i=1}^{n} e^{w_i} \\
= \; & \beta \left(\sum_{i=1}^{n} e^{w_i} \right) \left(\frac{(\alpha - 1)n}{\beta \sum_{i=1}^{n} e^{w_i}} c(\mathbf{w}) - e^{c(\mathbf{w})} + 1 \right) \\
= \; & \beta \left(\sum_{i=1}^{n} e^{w_i} \right) e^{c(\mathbf{w})} \left(c(\mathbf{w}) - 1 + e^{-c(\mathbf{w})} \right) \\
\geq \; & 0
\end{aligned}$$

where the last inequality holds by the fact $x - 1 + e^{-x} \geq 0$ for all $x \in \mathbb{R}$.

Proof of (2.4.7) Indeed, $\nabla \rho(\mathbf{w}) = \nabla \ell(\mathbf{w}) + \nabla \ell_0(\mathbf{w})$. It is readily checked that $\nabla \ell(\mathbf{w})^\top \mathbf{1} = 0$ for all $\mathbf{w} \in \mathbb{R}^n$. We next show that $\nabla \ell_0(\Pi(\mathbf{w}))^\top \mathbf{1} = 0$ for all $\mathbf{w} \in \mathbb{R}^n$.

Note that

$$\frac{\partial}{\partial w_i} \ell_0(\mathbf{w}) = \alpha - 1 - \beta e^{w_i} \text{ for } i = 1, 2, \ldots, n.$$

Hence,

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

Now, by definition of the mapping Π given by (2.4.4) and (2.4.5), for all $\mathbf{w} \in \mathbb{R}^n$,

$$\nabla \ell_0(\Pi(\mathbf{w}))^\top \mathbf{1} = (\alpha - 1)n - \beta e^{c(\mathbf{w})} \sum_{i=1}^{n} e^{w_i} = 0.$$

2.7.6 Proof of Lemma 2.7.1

Let $t_{i,j}$ be the number of paired comparisons in the input data with tie outcome for items i and j. Note that $t_{i,j} = t_{j,i}$. The log-likelihood function can be written as follows:

$$\begin{aligned}\ell(\mathbf{w}) &= \sum_{i=1}^{n}\sum_{j\neq i} d_{i,j}\left(w_i - \log(e^{w_i} + \theta e^{w_j})\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i} t_{i,j}\left(w_i + w_j - \log(e^{w_i} + \theta e^{w_j})\right. \\ &\quad \left. - \log(\theta e^{w_i} + e^{w_j}) + \log(\theta^2 - 1)\right).\end{aligned}$$

Let $\bar{d}_{i,j}$ be the number of paired comparisons of items i and j such that $i \succeq j$, i.e., $\bar{d}_{i,j} = d_{i,j} + t_{i,j}$. By a straightforward calculus, we can write

$$\begin{aligned}\ell(\mathbf{w}) &= \sum_{i=1}^{n}\sum_{j\neq i} \bar{d}_{i,j}\left(w_i - \log(e^{w_i} + \theta e^{w_j})\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{n} t_{i,j}\log(\theta^2 - 1).\end{aligned}$$

Now, we note when $i \neq j$,

$$\begin{aligned}&\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \\ = \; & -\bar{d}_{i,j}\frac{\theta e^{w_i}e^{w_j}}{(e^{w_i} + \theta e^{w_j})^2} - \bar{d}_{j,i}\frac{\theta e^{w_i}e^{w_j}}{(\theta e^{w_i} + e^{w_j})^2}\end{aligned}$$

and

$$\frac{\partial^2}{\partial w_i^2}(-\ell(\mathbf{w})) = -\sum_{j\neq i}\frac{\partial^2}{\partial w_u \partial w_j}(-\ell(\mathbf{w})).$$

For any $i \neq j$, it indeed holds

$$\frac{\theta e^{w_i}e^{w_j}}{(e^{w_i} + \theta e^{w_j})^2} \leq \frac{1}{4}.$$

Hence, when $i \neq j$,

$$\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \geq -\frac{1}{4}(\bar{d}_{i,j} + \bar{d}_{j,i}) \geq -\frac{1}{2}m_{i,j}.$$

It follows that $\frac{1}{2}\mathbf{L_M} \succeq \nabla^2(-\ell(\mathbf{w}))$ for all $\mathbf{w} \in \mathbb{R}^n$. Hence,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \leq \frac{1}{2}\lambda_n(\mathbf{L_M}) \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

This implies that $-\ell$ is a $\frac{1}{2}\lambda_n(\mathbf{L_M})$-smooth function on $\mathbb{R}^n$.

On the other hand, we can show that for all $\mathbf{w} \in [-\omega, \omega]^n$,

$$\frac{\theta e^{w_i} e^{w_j}}{(e^{w_i} + \theta e^{w_j})^2} \geq \frac{\theta}{(\theta e^{-\omega} + e^{\omega})^2} := c_{\theta,\omega}.$$

This can be noted as follows. Let $z = \theta e^{w_j}/(e^{w_i} + \theta e^{w_j})$. Note that

$$\frac{\theta e^{w_i} e^{w_j}}{(e^{w_i} + \theta e^{w_j})^2} = z(1-z)$$

and that $z \in \Omega := [1/(1+\theta e^{2\omega}), 1/(1+\theta e^{-2\omega})]$. The function $z(1-z)$ is convex and thus achieves its minimum value over the interval Ω at one of its boundary points. It can be readily checked that the minimum is achieved at $z^* = 1/(1+\theta e^{2\omega})$, which yields $z^*(1-z^*) = c_{\theta,\omega}$.

Hence, when $i \neq j$,

$$\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \leq -c_{\theta,\omega}(\bar{d}_{i,j} + \bar{d}_{j,i}) \leq -c_{\theta,\omega} m_{i,j}.$$

It follows that $\nabla^2(-\ell(\mathbf{w})) \succeq c_{\theta,\omega}\mathbf{L_M}$. From this, we have that for all $\mathbf{w} \in [-\omega, \omega]^n$ and $\mathbf{x} \in \mathcal{X}$,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \geq c_{\theta,\omega}\lambda_2(\mathbf{L_M})$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x}||_\infty \leq \omega \text{ and } \mathbf{x}^\top \mathbf{1} = 0\}$. This implies that $-\ell$ is $c_{\theta,\omega}\lambda_2(\mathbf{L_M})$-strongly convex on $\mathcal{X}$.

2.7.7 Derivation of the convergence time bound (2.6.1)

First note that

$$\underline{m}\mathbf{A} \leq \mathbf{M} \leq \bar{m}\mathbf{A}$$

where the inequalities hold elementwise. From this, it follows that $\mathbf{L_M} \succeq \underline{m}\mathbf{L_A}$ and $\bar{m}\mathbf{L_A} \succeq \mathbf{L_M}$, where recall $\mathbf{A}$ is the adjacency matrix induced by matrix $\mathbf{M}$. Now, note

$$d(\mathbf{M}) = ||\mathbf{M}||_\infty \leq \bar{m}d(n)$$

and

$$a(\mathbf{M}) = \lambda_2(\mathbf{L_M}) \geq \underline{m}\lambda_2(\mathbf{L_A})$$

where $d(n)$ is the maximum degree of a node in graph G.

Hence, we have

$$\frac{d(\mathbf{M})}{a(\mathbf{M})} \leq \frac{rd(n)}{\lambda_2(\mathbf{L_A})}.$$

By Theorem 3.4 in Merris [1994], for any graph G with adjacency matrix $\mathbf{A}$ and diameter $D(n)$, $\lambda_2(\mathbf{L_A}) \geq 4/(nD(n))$.

It thus follows that

$$\frac{d(\mathbf{M})}{a(\mathbf{M})} \leq \frac{1}{4}rd(n)D(n)n$$

which implies the convergence time bound $T = O(rd(n)D(n)n\log(1/\epsilon))$.

2.7.8 Generalized Bradley-Terry models

In this section, we discuss how the results for Bradley-Terry model of paired comparisons can be extended to other instances of generalized Bradley-Terry models. In particular, we show this for the Rao-Kupper model of paired comparisons with tie outcomes, the Luce choice model and the Plackett-Luce ranking model.

We discuss only the characterization of the strong-convexity and smoothness parameters as the convergence rate bounds for gradient descent and MM algorithms follow similarly as in Section 2.3.1, from Theorems 2.3.1 and 2.3.2, respectively. Similarly, the rate of convergence bounds for accelerated gradient descent and MM algorithms follow readily, similarly to as in Section 2.4.1, from Theorems, 2.4.1 and 2.4.2, respectively.

2.7.8.1 Model definitions

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 $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}^n$ 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 $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}^n$ 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 $S \subseteq N = \{1, 2, \ldots, n\}$ 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 $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}_n$ 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 $y = (y_1, y_2, \ldots, y_k) \in T$, y_1 is an arbitrary item and $y_2, \ldots, y_k$ are sorted in lexicographical order. We can interpret each $y = (y_1, y_2, \ldots, y_k) \in T$ as a choice of item

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

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

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 $S \subseteq N = \{1, 2, \ldots, n\}$, 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 $y = (y_1, y_2, \ldots, y_k) \in T$ corresponds to a permutation of the set of items $S = \{y_1, y_2, \ldots, y_k\}$. The probability of outcome y is given by

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

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

In this section, we discuss how the results for Bradley-Terry model of paired comparisons can be extended to other instances of generalized Bradley-Terry models. In particular, we show this for the Rao-Kupper model of paired comparisons with tie outcomes, the Luce choice model and the Plackett-Luce ranking model.

2.7.8.2 Rao-Kupper model

The probability distribution of outcomes according to the Rao-Kupper model is defined in Section 2.7.8.1. The log-likelihood function can be written as

$$\begin{aligned}\ell(\mathbf{w}) &= \sum_{i=1}^{n}\sum_{j\neq i}\bar{d}_{i,j}\left(w_i - \log(e^{w_i}+\theta e^{w_j})\right)\\ &\quad +\frac{1}{2}\sum_{i=1}^{n} t_{i,j}\log(\theta^2-1)\end{aligned}$$

where $\bar{d}_{i,j}$ is the number of observed paired comparisons of items i and j such that either i wins against j or there is a tie outcome, and $t_{i,j}$ is the number of observed paired comparisons of items i and j with tie outcomes.

Lemma 2.7.1. *The negative log-likelihood function for the Rao-Kupper model of paired comparisons with parameter $\theta > 1$ is γ-strongly convex on $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega$ and $\mathbf{w}^\top \mathbf{1} = 0\}$ and μ-smooth on $\mathbb{R}^n$ with*

$$\gamma = c_{\theta,\omega}\lambda_2(\mathbf{L_M}) \text{ and } \mu = \frac{1}{2}\lambda_n(\mathbf{L_M})$$

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

Proof of Lemma 2.7.1 is provided in Section 2.7.6.

A surrogate minorant function for the log-likelihood function of the Rao-Kupper model is given as follows:

$$\begin{aligned}&\underline{\ell}(\mathbf{x};\mathbf{y})\\ =\ &\sum_{i=1}^{n}\sum_{j\neq i}\bar{d}_{i,j}\left(x_i - \frac{e^{x_i}+\theta e^{x_j}}{e^{y_i}+\theta e^{y_j}} - \log(e^{y_i}+\theta e^{y_j}) + 1\right)\\ &+\frac{1}{2}\sum_{i=1}^{n} t_{i,j}\log(\theta^2-1).\end{aligned}$$

The MM algorithm is defined by, for $i = 1, 2, \ldots, n$,

$$w_i^{(t+1)} = \log\left(\sum_{j \neq i} \bar{d}_{i,j}\right) - \log\left(\sum_{j \neq i}\left(\frac{\bar{d}_{i,j}}{e^{w_i^{(t)}} + \theta e^{w_j^{(t)}}} + \frac{\theta \bar{d}_{j,i}}{e^{w_j^{(t)}} + \theta e^{w_i^{(t)}}}\right)\right).$$

Lemma 2.7.2. *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*

$$\delta = e^{2\omega} d(\mathbf{M}).$$

2.7.8.3 Luce choice model

The probability distribution of outcomes according to the Luce choice model is defined in Section 2.7.8.1. The log-likelihood function can be written as:

$$\ell(\mathbf{w}) = \sum_{y \in T} d_y \left(w_{y_1} - \log\left(\sum_{j \in y} e^{w_j}\right)\right).$$

Lemma 2.7.3. *The negative log-likelihood function for the Luce choice model with comparison sets of size* $k \geq 2$ *is* γ*-strongly convex and* μ*-smooth on* $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega$ *and* $\mathbf{w}^\top \mathbf{1} = 0\}$ *with*

$$\gamma = c_{\omega,k} \lambda_2(\mathbf{L_M}) \text{ and } \mu = d_{\omega,k} \lambda_n(\mathbf{L_M})$$

where

$$c_{\omega,k} = \begin{cases} 1/(e^{-\omega} + e^{\omega})^2, & \text{if } k = 2 \\ 1/((k-2)e^{2\omega} + 2)^2, & \text{if } k > 2 \end{cases}$$

and

$$d_{\omega,k} = \frac{1}{((k-2)e^{-2\omega} + 2)^2}.$$

Note that for every fixed $\omega > 0$, (a) $c_{\omega,k}/d_{\omega,k}$ is decreasing in k, (b) $1/e^{8\omega} \leq c_{\omega,k}/d_{\omega,k} \leq 1/e^{2\omega}$, and (c) $1/e^{8\omega}$ is the limit value of $c_{\omega,k}/d_{\omega,k}$ as k goes to infinity.

A minorant surrogate function for the log-likelihood function of the Luce choice model is given by

$$\underline{\ell}(\mathbf{x};\mathbf{y}) = \sum_{y\in T} d_y \left(x_{y_1} - \frac{\sum_{j\in y} e^{x_j}}{\sum_{j\in y} e^{y_j}} - \log\left(\sum_{j\in y} e^{y_j}\right) + 1 \right).$$

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

$$\begin{aligned} w_i^{(t+1)} &= \log\left(\sum_{y\in T} d_y \mathbb{I}_{i=y_1}\right) \\ &\quad - \log\left(\sum_{y\in T} d_y \mathbb{I}_{i\in y} \frac{1}{\sum_{j\in y} e^{w_j^{(t)}}}\right) \end{aligned}$$

where $\sum_{y\in T} d_y \mathbb{I}_{i=y_1}$ is the number of observed comparisons in which item i is the chosen item.

Lemma 2.7.4. *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*

$$\delta = \frac{1}{k(k-1)} e^{2\omega} d(\mathbf{M}).$$

2.7.8.4 Plackett-Luce ranking model

The probability distribution of outcomes according to the Plackett-Luce ranking model is defined in Section 2.7.8.1. The log-likelihood function can be written as follows:

$$\ell(\mathbf{w}) = \sum_{y\in T} d_y \sum_{r=1}^{|y|-1} \left(w_{y_r} - \log\left(\sum_{j=r}^{|y|} e^{w_{y_j}}\right)\right).$$

Lemma 2.7.5. *The negative log-likelihood function for the Plackett-Luce ranking model with comparison sets of size* $k \geq 2$ *is* γ*-strongly convex and* μ*-smooth on* $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega$ *and* $\mathbf{w}^\top \mathbf{1} = 0\}$ *with*

$$\gamma = \tilde{c}_{\omega,k} \lambda_2(\mathbf{L_M}) \text{ and } \mu = \tilde{d}_{\omega,k} \lambda_n(\mathbf{L_M})$$

where

$$\tilde{c}_{\omega,k} = \frac{1}{k^2} e^{-4\omega} \text{ and } \tilde{d}_{\omega,k} = \left(2 - \frac{1}{k}\right) e^{4\omega}.$$

Proof of Lemma 2.7.5 is provided in Section 2.7.10.6.

Note that for fixed values of ω and k, Lemma 2.7.5 implies the convergence time $\log(d(\mathbf{M})/a(\mathbf{M}))$. Note, however, that for fixed $\omega > 0$, $\tilde{c}_{\omega,k}/\tilde{d}_{\omega,k}$ decreases to 0 with k and is of the order $1/k^2$. This is because in the derivation of parameters $\tilde{c}_{\omega,k}$ and $\tilde{d}_{\omega,k}$ we use (conservative) deterministic bounds. Following Hajek et al. [2014b], one can derive bounds for γ and μ that hold with high probability, which are such that $\tilde{c}_{\omega,k}$ and $\tilde{d}_{\omega,k}$ scale with k in the same way.

The log-likelihood function of the Plackett-Luce ranking model admits the following minorization function:

$$\begin{aligned} & \underline{\ell}(\mathbf{x};\mathbf{y}) \\ = & \sum_{y\in T} d_y \sum_{r=1}^{|y|-1} \left(x_{y_r} - \frac{\sum_{j=r}^{|y|} e^{x_{y_j}}}{\sum_{j=r}^{|y|} e^{y_{y_j}}} - \log\left(\sum_{j=r}^{|y|} e^{y_{y_j}} \right) + 1 \right). \end{aligned}$$

The MM algorithm is given by: for $i = 1, 2, \ldots, n$,

$$\begin{aligned} w_i^{(t+1)} = & \log\left(\sum_{y\in T} d_y \mathbb{I}_{i\in S_{1,|y|-1}(y)} \right) \\ & - \log\left(\sum_{y\in T} d_y \sum_{r=1}^{|y|-1} \mathbb{I}_{i\in S_{r,|y|}(y)} \frac{1}{\sum_{j=r}^{|y|} e^{w_{y_j}^{(t)}}} \right) \end{aligned}$$

where $S_{a,b}(y) = \{y_a, y_{a+1}, \ldots, y_b\}$.

Lemma 2.7.6. *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*

$$\delta = \frac{1}{2} e^{2\omega} d(\mathbf{M}).$$

2.7.9 Additional numerical results

GD v.s. MM

Numerical results presented in Table 2.7 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 $\beta > 0$.

Table 2.5: 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

The comparison of the quality of obtained estimates from MM and AccMM.

In order to demonstrate that our accelerated algorithm also guarantees the quality of the obtained estimates, I conducted a toy simulation as follows: I considered a scenario with 6 items, where each distinct pair of items was compared 5 times. The input data was generated according to the Bradley-Terry model of paired comparisons with the parameter vector $\mathbf{w}$, such that a half of the items have parameters ω, and the other half have parameters $-\omega$. We choose $\omega = 1/2$.

We defined the convergence time (T) as the smallest integer t such that $\|\mathbf{w}^t - \mathbf{w}^{(t-1)}\| \leq 10^{-4}$, and $\hat{\mathbf{w}}^*$ represents the estimated parameter vector obtained from either the MM algorithm or the AccMM algorithm. The results on the number of iterations for the MM algorithm and the accelerated MM algorithm (AccMM) are summarized in Table 1. It can be observed that the number of iterations increases as the value of β decreases. Additionally, Table 2 presents the results of $\|\hat{\mathbf{w}}^* - \mathbf{w}\|_F^2$ for MM and AccMM algorithms for different values of α. It is evident that AccMM is much faster than MM when β becomes smaller. We also observed that $\|\hat{\mathbf{w}}^* - \mathbf{w}\|_F^2$ for AccMM is slightly larger than that for MM for different β values. This might suggest that the quality of the estimates obtained

Table 2.6: The convergence time (T) for the MM algorithm and accelerated MM algorithm (AccMM).

Algorithm	$\beta = 0$	0.1	1	10
MM	15	67	35	7
AccMM		15	13	7

Table 2.7: $\|\hat{\omega}^* - \omega\|_F^2$ for the MM algorithm and accelerated MM algorithm (AccMM).

Algorithm	$\beta = 0$	0.1	1	10
MM	1.593867	1.601168	1.572034	1.506924
AccMM		1.621299	1.573552	1.506955

from the MM algorithm is slightly better than those from the AccMM algorithm, but the difference is negligible. I repeated the simulation with different values of ω, and all gave me similar results to Table 2. Sometimes, $\|\hat{\mathbf{w}}^* - \mathbf{w}\|_F^2$ for AccMM is even slightly larger than MM, but again, the difference is negligible. Hence, this might suggest that the quality of estimates obtained from MM and AccMM is not significantly different. In this scenario, our AccMM algorithm is superior because it converges faster than MM while providing similar quality of estimates compared to the MM algorithm.

2.7.10 Background proofs for Chapter 2.3 and Chapter 2.7.8

2.7.10.1 Proof of Theorem 2.3.1

Let $\mathbf{x}'$ be the output of the gradient descent iteration update for input $\mathbf{x}$ with step size η.

If $\mathbf{x} \in \mathcal{X}_\gamma$ and $\mathbf{x}' \in \mathcal{X}_\mu$, then

$$\begin{aligned}
& f(\mathbf{x}') - f(\mathbf{x}^*) \\
= \; & f(\mathbf{x} - \eta \nabla f(\mathbf{x})) - f(\mathbf{x}^*) \\
\leq \; & f(\mathbf{x}) - \eta ||\nabla f(\mathbf{x})||^2 + \frac{\mu}{2}\eta^2 ||\nabla f(\mathbf{x})||^2 - f(\mathbf{x}^*) \\
= \; & f(\mathbf{x}) - f(\mathbf{x}^*) - \left(\eta - \frac{\mu}{2}\eta^2\right) ||\nabla f(\mathbf{x})||^2 \\
\leq \; & f(\mathbf{x}) - f(\mathbf{x}^*) - 2\gamma \left(\eta - \frac{\mu}{2}\eta^2\right) (f(\mathbf{x}) - f(\mathbf{x}^*)) \\
= \; & (1 - 2\gamma\eta + \gamma\mu\eta^2)(f(\mathbf{x}) - f(\mathbf{x}^*))
\end{aligned}$$

where the first inequality is by the assumption that f is μ-smooth on $\mathcal{X}_\mu$ and the second inequality is by the assumption that f satisfies the γ-PL inequality on $\mathcal{X}_\gamma$. Taking $\eta = 1/\mu$, which minimizes the above bound, establishes the claim of the theorem.

2.7.10.2 Proof of Theorem 2.3.2

Let $\mathbf{x}'$ be the output of the MM algorithm iteration update for input $\mathbf{x}$.

By the facts $f(\mathbf{x}') \leq g(\mathbf{x}';\mathbf{x})$ and $g(\mathbf{x}';\mathbf{x}) \leq g(\mathbf{z};\mathbf{x})$ for all $\mathbf{z}$, for any $\eta \geq 0$,

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ \leq \; & g(\mathbf{x}';\mathbf{x}) - f(\mathbf{x}^*) \\ \leq \; & g(\mathbf{x} - \eta\nabla f(\mathbf{x});\mathbf{x}) - f(\mathbf{x}^*) \\ = \; & f(\mathbf{x} - \eta\nabla f(\mathbf{x})) - f(\mathbf{x}^*) \\ & + g(\mathbf{x} - \eta\nabla f(\mathbf{x});\mathbf{x}) - f(\mathbf{x} - \eta\nabla f(\mathbf{x})). \end{aligned}$$

Now, by the same arguments as in the proof of Theorem 2.3.1, if $\mathbf{x} \in \mathcal{X}_\gamma$ and $\mathbf{x} - \eta\nabla f(\mathbf{x}) \in \mathcal{X}_\mu$, we have

$$\begin{aligned} & f(\mathbf{x} - \eta\nabla f(\mathbf{x})) - f(\mathbf{x}^*) \\ \leq \; & (1 - 2\gamma\eta + \gamma\mu\eta^2)(f(\mathbf{x}) - f(\mathbf{x}^*)). \end{aligned}$$

Next, if $\mathbf{x} \in \mathcal{X}_\gamma$ and $\mathbf{x} - \eta\nabla f(\mathbf{x}) \in \mathcal{X}_\mu$,

$$\begin{aligned} & g(\mathbf{x} - \eta\nabla f(\mathbf{x});\mathbf{x}) - f(\mathbf{x} - \eta\nabla f(\mathbf{x})) \\ \leq \; & \frac{\delta}{2}\eta^2 ||\nabla f(\mathbf{x})||^2 \\ \leq \; & \delta\eta^2\gamma(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

where the first inequality is by the smoothness condition on the majorant surrogate function and the second inequality is by the assumption that f satisfies the PL inequality with parameter γ on $\mathcal{X}_\gamma$.

Putting the pieces together, we have

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ \leq \; & \left(1 - 2\gamma\eta + \gamma(\mu + \delta)\eta^2\right)(f(\mathbf{x}) - f(\mathbf{x}^*)). \end{aligned}$$

Taking $\eta = 1/(\mu + \delta)$ (which minimizes the factor involving η in the last inequality) yields the asserted

result.

2.7.10.3 Proof of Lemma 2.3.1

The Hessian of the negative log-likelihood function has the following elements:

$$\nabla^2(-\ell(\mathbf{w}))_{i,j} = \begin{cases} \sum_{v \neq i} m_{i,v} \frac{e^{w_i} e^{w_v}}{(e^{w_i}+e^{w_v})^2}, & \text{if } i = j \\ -m_{i,j} \frac{e^{w_i} e^{w_j}}{(e^{w_i}+e^{w_j})^2}, & \text{if } i \neq j. \end{cases} \tag{2.7.4}$$

We will show that for all $i \neq j$,

$$\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \leq -c_\omega m_{i,j} \text{ for all } \mathbf{w} \in [-\omega, \omega]^n \tag{2.7.5}$$

and

$$-\frac{1}{4} m_{i,j} \leq \frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \text{ for all } \mathbf{w} \in \mathbb{R}^n. \tag{2.7.6}$$

From (2.7.5), we have $\nabla^2(-\ell(\mathbf{w})) \succeq c_\omega \mathbf{L_M}$ for all $\mathbf{w} \in [-\omega, \omega]^n$. Hence, for all $\mathbf{w} \in [-\omega, \omega]^n$ and $\mathbf{x} \in \mathcal{X}$,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \geq c_\omega \lambda_2(\mathbf{L_M})||\mathbf{x}||^2$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{1} = 0\}$. This shows that $-\ell$ is $c_\omega \lambda_2(\mathbf{L_M})$-strongly convex on $\mathcal{X}$.

From (2.7.6), we have $\frac{1}{4}\mathbf{L_M} \succeq \nabla^2(-\ell(\mathbf{w}))$ for all $\mathbf{w} \in \mathbb{R}^n$. Hence, for all $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \leq \frac{1}{4}\lambda_n(\mathbf{L_M})||\mathbf{x}||^2.$$

This shows that $-\ell$ is $\frac{1}{4}\lambda_n(\mathbf{L_M})$-smooth on $\mathbb{R}^n$.

It remains to show that (2.7.5) and (2.7.6) hold. For (2.7.5), we need to show that $c_\omega \leq x_i x_j/(x_i+x_j)^2$ for all $\mathbf{x} \in [-\omega, \omega]^n$. Note that $x_i x_j/(x_i + x_j)^2 = z(1-z)$ where $z := x_i/(x_i + x_j)$. Note that $z \in \Omega := [e^{-\omega}/(e^{-\omega} + e^{\omega}), 1 - e^{-\omega}/(e^{-\omega} + e^{\omega})]$ for all $\mathbf{x} \in [-\omega, \omega]^n$. The function $z(1-z)$ achieves its minimum over the interval Ω at a boundary of Ω. Thus, it holds $\min_{z\in\Omega} z(1-z) = c_\omega$. For (2.7.6), we can immediately note that for all $\mathbf{w} \in \mathbb{R}^n$,

$$\frac{w_i w_j}{(w_i + w_j)^2} = \frac{w_i}{w_i + w_j}\left(1 - \frac{w_i}{w_i + w_j}\right) \leq \frac{1}{4}.$$

2.7.10.4 Proof of Lemma 2.3.3

Let $\mathbf{y}$ be an arbitrary vector in $[-\omega, \omega]^n$. Let $r(\mathbf{x}; \mathbf{y}) = \underline{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x})$ for $\mathbf{x} \in [-\omega, \omega]^n$. Then, we have

$$\begin{aligned} r(\mathbf{y}; \mathbf{y}) = 0, \nabla_{\mathbf{x}} r(\mathbf{y}; \mathbf{y}) = 0, \text{ and} \\ \nabla_x^2 r(\mathbf{x}; \mathbf{y}) = \nabla^2(-\ell(\mathbf{x})) + A \end{aligned} \tag{2.7.7}$$

where A is a $n \times n$ diagonal matrix with diagonal elements

$$A_{i,i} = -\sum_{j \in i} m_{i,j} \frac{e^{x_i}}{e^{y_i} + e^{y_j}} \geq -\frac{1}{2} e^{2\omega} ||\mathbf{M}||_\infty.$$

Since $\nabla^2(-\ell(\mathbf{x}))$ is a positive semi-definite matrix and A is a diagonal matrix, for all $\mathbf{x}, \mathbf{y} \in [-\omega, \omega]^n$ and $\mathbf{w} \in [-\omega, \omega]^n$, we have

$$\mathbf{x}^\top \nabla_{\mathbf{x}}^2 r(\mathbf{w}; \mathbf{y}) \mathbf{x} \geq -||\mathbf{M}||_\infty \frac{e^{2\omega}}{2} ||\mathbf{x}||^2 = -\delta ||\mathbf{x}||^2.$$

By limited Taylor expansion, for all $\mathbf{x} \in [-\omega, \omega]^n$,

$$\begin{aligned} & r(\mathbf{x}; \mathbf{y}) \\ \geq\ & r(\mathbf{y}; \mathbf{y}) + (\mathbf{x} - \mathbf{y})^\top \nabla_{\mathbf{x}} r(\mathbf{y}; \mathbf{y}) \\ & + \frac{1}{2} \min_{0 \leq a \leq 1} (\mathbf{x} - \mathbf{y})^\top \nabla_{\mathbf{x}}^2 r(a\mathbf{x} + (1-a)\mathbf{y}; \mathbf{y})(\mathbf{x} - \mathbf{y}) \\ =\ & \frac{1}{2} \min_{0 \leq a \leq 1} (\mathbf{x} - \mathbf{y})^\top \nabla_{\mathbf{x}}^2 r(a\mathbf{x} + (1-a)\mathbf{y})(\mathbf{x} - \mathbf{y}; \mathbf{y}) \\ \geq\ & -\frac{\delta}{2} ||\mathbf{x} - \mathbf{y}||^2. \end{aligned}$$

By the definition of $r(\mathbf{x}; \mathbf{y})$, we have $\bar{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x}) \geq -\frac{\delta}{2} ||\mathbf{x} - \mathbf{y}||^2$.

2.7.10.5 Proof of Lemma 2.3.4

We consider the log-a posteriori probability function $\rho(\mathbf{w}) = \ell(\mathbf{w}) + \ell_0(\mathbf{w}) + \text{const}$ where ℓ is the log-likelihood function given by (2.2.3) and ℓ_0 is the prior log-likelihood function given by (2.2.5). Note that $\nabla^2(-\ell_0(\mathbf{w}))$ is a diagonal matrix with diagonal elements equal to βe^{w_i}, for $i = 1, 2, \ldots, n$. It can be readily shown that for $\mathbf{w} \in \mathcal{W}_\omega$,

$$c_\omega \mathbf{L}_\mathbf{M} + e^{-\omega} \beta \mathbf{I}_n \preceq \nabla^2(-\rho(\mathbf{w})) \preceq \frac{1}{4} \mathbf{L}_\mathbf{M} + e^{\omega} \beta \mathbf{I}_n. \tag{2.7.8}$$

From (2.7.8), for all $\mathbf{w} \in \mathcal{W}_\omega$ and $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top \nabla^2(-\rho(\mathbf{w}))\mathbf{x} \geq \lambda_1(e^{-\omega}\beta \mathbf{I}_n)||\mathbf{x}||^2 = e^{-\omega}\beta||\mathbf{x}||^2.$$

Hence, $-\rho$ is $e^{-\omega}\beta$-strongly convex on $\mathcal{W}_\omega$.

Similarly, from (2.7.8), for all $\mathbf{w} \in \mathcal{W}_\omega$, and $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned}
\mathbf{x}^\top \nabla^2(-\rho(\mathbf{w}))\mathbf{x} &\leq \lambda_n(\frac{1}{4}\mathbf{L_M} + e^{\omega}\beta \mathbf{I}_n)||\mathbf{x}||^2 \\
&\leq (\lambda_n(\frac{1}{4}\mathbf{L_M}) + \lambda_n(e^{\omega}\beta \mathbf{I}_n))||\mathbf{x}||^2 \\
&= (\frac{1}{4}\lambda_n(\mathbf{L_M}) + e^{\omega}\beta)||\mathbf{x}||^2.
\end{aligned}$$

Hence, $-\rho$ is μ-smooth on $\mathcal{W}_\omega$ with $\mu = \frac{1}{4}\lambda_n(\mathbf{L_M}) + e^{\omega}\beta$.

2.7.10.6 Proof of Lemma 2.7.5

It can be easily shown that for all $\mathbf{w} \in [-\omega, \omega]^n$, $S \subseteq N$ such that $|S| \geq 2$, and $u, v \in S$ such that $u \neq v$, we have

$$\frac{e^{-4\omega}}{|S|^2} \leq \frac{e^{w_u}e^{w_v}}{(\sum_{j\in S} e^{w_j})^2} \leq \frac{e^{4\omega}}{|S|^2}.$$

Combining with (2.7.4), we have

$$\begin{aligned}
& \frac{\partial^2}{\partial w_u \partial w_v}(-\ell(\mathbf{w})) \\
\leq & -\sum_{y\in T} d_y \frac{w_u w_v}{(\sum_{j=1}^k e^{w_{y_j}})^2} 1_{u,v\in\{y_1,y_2,\ldots,y_k\}} \\
\leq & -\frac{e^{-4\omega}}{k^2} \sum_{y\in T} d_\pi 1_{u,v\in\{y_1,y_2,\ldots,y_k\}} \\
= & -\frac{e^{-4\omega}}{k^2} m_{u,v}.
\end{aligned}$$

From this it follows that for all $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{x}^\top \mathbf{1} = 0$,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \geq \frac{e^{-4\omega}}{k^2}\lambda_2(\mathbf{L_M})||\mathbf{x}||^2. \tag{2.7.9}$$

Similarly, we have

$$\begin{aligned}
& \frac{\partial^2}{\partial w_u \partial w_v}(-\ell(\mathbf{w})) \\
\geq & -\sum_{y \in T} d_y \sum_{l=1}^{k-1} \frac{w_u w_v}{(\sum_{j=l}^{k} e^{w_{y_j}})^2} 1_{u,v \in \{y_1, y_2, \ldots, y_k\}} \\
\geq & -e^{4\omega} \sum_{l=1}^{k-1} \frac{1}{(k-l+1)^2}\, m_{u,v} \\
= & -e^{4\omega} \sum_{l=2}^{k} \frac{1}{l^2}\, m_{u,v} \\
\geq & -e^{4\omega} \left(1 + \int_1^k \frac{dx}{x^2}\right) m_{u,v} \\
= & -e^{4\omega} \left(2 - \frac{1}{k}\right) m_{u,v}.
\end{aligned}$$

From this it follows that for all $\mathbf{x}$,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \leq e^{4\omega} \left(2 - \frac{1}{k}\right) \lambda_n(\mathbf{L_M})||\mathbf{x}||^2. \tag{2.7.10}$$

Bibliography

B. Abrahao, F. Chierichetti, R. Kleinberg, and A. Panconesi. Trace complexity of network inference. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 491–499, 2013.

I. Adler and S. M. Ross. The coupon subset collection problem. *Journal of Applied Probability*, 38(3): 737–746, 2001. doi: 10.1239/jap/1005091036.

A. Agarwal, P. Patil, and S. Agarwal. Accelerated spectral ranking. In *International Conference on Machine Learning*, pages 70–79, 2018.

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

A. E. Alaoui and A. Montanari. On the computational tractability of statistical estimation on amenable graphs, 2019.

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984. ISSN 00063444. URL http://www.jstor.org/stable/2336390.

D. Aldous. Interacting particle systems as stochastic social dynamics. *Bernoulli*, 19(4):1122–1149, 09 2013.

D. Aldous and J. A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. Unfinished monograph, 2002.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *Information and Inference: A Journal of the IMA*, 3: 224–294, 2014.

D. Balduzzi, K. Tuyls, J. Pérolat, and T. Graepel. Re-evaluating evaluation. In *Proceedings of the 32nd*

Conference on Neural Information Procesing Systems (NeurIPS '18), Montreal, Canada, December 2018.

O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley series in probability and mathematical statistics, 1978.

S. Barnett and C. Storey. *Matrix Methods in Stability Theory*. Nelson, 1970.

S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, 08 2015.

P. H. Baxendale. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.*, 15(1B):700–738, 02 2005.

P. Berenbrink, G. Giakkoupis, A.-M. Kermarrec, and F. Mallmann-Trenn. Bounds on the Voter Model in Dynamic Networks. In I. Chatzigiannakis, M. Mitzenmacher, Y. Rabani, and D. Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 146:1–146:15, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

J. Blitzstein and P. Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Math.*, 6(4):489–522, 2010. URL https://projecteuclid.org:443/euclid.im/1318514519.

B. Bollobás. *Random Graphs*. Cambridge University Press, 2 edition, 2001.

B. Bollob as. *Random Graphs*. Cambridge University Press, 2 edition, 2001.

B. Bollobás and O. Riordan. Counting dense connected hypergraphs via the probabilistic method. *Random Structures & Algorithms*, 53(2):185–220, 2018. doi: 10.1002/rsa.20762. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20762.

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

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

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

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.

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.

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

L. D. Brown. *Fundamental of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Lecture Notes - Monograph Series, 1986.

S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4): 231–357, Nov. 2015. ISSN 1935-8237.

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.

E. J. Candes and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv e-prints*, art. arXiv:1804.09753, Apr 2018.

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

F. Caron and A. Doucet. Efficient bayesian inference for generalized bradley–terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012b.

S. Chatterjee and E. Seneta. Towards consensus: some convergence theorems on repeated averaging. *Journal of Applied Probability*, 14(1):89–97, 1977.

S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *Ann. Appl. Probab.*, 21(4):1400–1435, 08 2011. doi: 10.1214/10-AAP728. URL https://doi.org/10.1214/10-AAP728.

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.

Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *The Annals of Statistics*, 20(2204–2235), 2019.

F. R. K. Chun. *Spectral Graph Theory*, volume 92. Americal Mathematical Society and Conference Board of the Mathematical Sciences, 1997.

F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

A. Coja-Oghlan. On the laplacian eigenvalues of $G_{n,p}$. *Combinatorics, Probability and Computing*, 16:923–946, 2007.

O. Cooley, M. Kang, and C. Koch. The size of the giant high-order component in random hypergraphs. *Random Structures & Algorithms*, 53(2):238–288, 2018a. doi: 10.1002/rsa.20761. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20761.

O. Cooley, M. Kang, and Y. Person. Largest components in random hypergraphs. 27(5):741–762, 2018b. doi: DOI:10.1017/S096354831800010X. URL https://www.cambridge.org/core/article/largest-components-in-random-hypergraphs/7DD8B2411550207E3D92F8745D5AE47D.

O. Cooley, M. Kang, and C. Koch. The size of the giant component in random hypergraphs: a short proof. In *The Electronic Journal of Combinatorics*, volume 26, 2019.

C. Cooper and N. Rivera. The linear voting model. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 144:1–144:12, 2016.

C. Cooper, A. Frieze, and W. Pegden. On the rank of a random binary matrix. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 946–955.

C. Cooper, A. Frieze, and T. Radzik. Multiple random walks in random regular graphs. *SIAM Journal on Discrete Mathematics*, 23(4):1738–1761, 2010.

C. Cooper, R. Elsässer, H. Ono, and T. Radzik. Coalescing random walks and voting on connected graphs. *SIAM Journal on Discrete Mathematics*, 27(4):1748–1758, 2013.

C. Cooper, M. Dyer, A. Frieze, and N. Rivera. Discordant voting processes on finite graphs. *SIAM Journal on Discrete Mathematics*, 32(4):2398–2420, 2018.

K. P. Costello and V. Vu. On the rank of random sparse matrices. *Combinatorics, Probability and Computing*, 19(3):321–342, 2010. doi: 10.1017/S0963548309990447.

K. P. Costello and V. H. Vu. The rank of random graphs. *Random Structures & Algorithms*, 33 (3):269–285, 2008. doi: 10.1002/rsa.20219. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20219.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2 edition, 2006.

J. T. Cox. Coalescing Random Walks and Voter Model Consensus Times on the Torus in $\mathbb{Z}^d$. *The Annals of Probability*, 17(4):1333 – 1366, 1989a. doi: 10.1214/aop/1176991158. URL https://doi.org/10.1214/aop/1176991158.

J. T. Cox. Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. *Annals of Probability*, 17(4):1333–1366, 10 1989b.

M. Daltayanni, L. de Alfaro, and P. Papadimitriou. Workerrank: Using employer implicit judgements to infer worker reputation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 263–272. ACM, 2015.

H. A. David. *The Method of Paired Comparisons*. Charles Griffin and Company, London, 1963.

R. R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.

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

M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974.

M. Desai and V. Rao. A characterization of the smallest eigenvalue of a graph. *Journal of Graph Theory*, 18(2):181–194, 1994. doi: 10.1002/jgt.3190180210. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jgt.3190180210.

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.

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.

D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

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

A. E. Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978b.

P. Erdös and T. Gallai. Graphen mit punkten vorgenschriebenen grades. *Mat. Lapok*, 11:264–274, 1960.

P. Erdös and A. Rényi. On random graphs. i. *Publicationes Mathematicae*, 6:290–297, 1959.

J. Fernley and M. Ortgiese. Voter models on subcritical scale-free random graphs. *Random Structures & Algorithms*, n/a(n/a).

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

M. Fiedler. Laplacian of graphs and algebraic connectivity. *Combinatorics and Graph Theory*, 25: 57–70, 1989.

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

Z. Gajic and M. T. J. Qureshi. *Lyapunov matrix equation in system stability and control*. Dover, 1995.

M. Gendreau. On the location of eigenvalues of off-diagonal constant matrices. *Linear Algebra and its Applications*, 79:99 – 102, 1986.

C. Godsil and G. Royle. *Algebraic Connectivity of Graphs*. Springer, 2001.

G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 4 edition, 2013.

M. Gomez-Rodriguez, L. Song, H. Daneshm, and B. Schölkopf. Estimating diffusion networks:

Recovery conditions, sample complexity and soft-thresholding algorithm. *Journal of Machine Learning Research*, 17(90):1–29, 2016.

S. M. Goodreau, J. A. Kitts, and M. Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103–125, 02 2009.

T. Graepel, T. Minka, and R. Herbrich. Trueskill(tm): A bayesian skill rating system. In *Proc. of NIPS 2006*, volume 19, pages 569–576, 2006.

M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6): 1420–1443, 1978.

B. L. Granovsky and N. Madras. The noisy voter model. *Stochastic Processes and their Applications*, 55(1):23 – 43, 1995.

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.

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.

B. Hajek, S. Oh, and J. Xu. Minimax-optimal inference from partial rankings. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1475–1483. Curran Associates, Inc., 2014a. URL http://papers.nips.cc/paper/5361-minimax-optimal-inference-from-partial-rankings.pdf.

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, 2014b.

E. C. Hall, G. Raskutti, and R. Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv e-prints*, art. arXiv:1605.02693, May 2016.

E. C. Hall, G. Raskutti, and R. M. Willett. Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory*, 65(4):2401–2422, April 2019.

R. Han, R. Ye, C. Tan, and K. Chen. Asymptotic theory of sparse Bradley–Terry model. *The*

Annals of Applied Probability, 30(5):2491 – 2515, 2020. doi: 10.1214/20-AAP1564. URL https://doi.org/10.1214/20-AAP1564.

Y. Hassin and D. Peleg. Distributed probabilistic polling and applications to proportionate agreement. *Information and Computation*, 171(2):248–268, 2001.

T. Hastie and R. Tibshirani. Classification by pairwise coupling. *Ann. Statist.*, 26(2):451–471, 04 1998. doi: 10.1214/aos/1028144844. URL http://dx.doi.org/10.1214/aos/1028144844.

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

T. P. Hayes. A large-deviation inequality for vector-valued martingales, 2005.

J. Hendrickx, A. Olshevsky, and V. Saligrama. Minimax rate for learning from pairwise comparisons in the BTL model. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4193–4202. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hendrickx20a.html.

R. Herbrich, T. Minka, and T. Graepel. Trueskill™: A Bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS '06)*, pages 569–576. Canada, 2006.

C. Hillar and A. Wibisono. Maximum entropy distributions on graphs, 2013.

J. Hoffmann and C. Caramanis. Learning graphs from noisy epidemic cascades. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(2), June 2019.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann. Probab.*, 3(4):643–663, 08 1975.

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

T.-K. Huang, C.-J. Lin, and R. C. Weng. Ranking individuals by group comparisons. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 425–432, New

York, NY, USA, 2006a. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143898. URL http://doi.acm.org/10.1145/1143844.1143898.

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. 2006b.

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. 2006c. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248547.1248551.

T.-K. Huang, C.-J. Lin, and R. C. Weng. Ranking individuals by group comparisons. *J. Mach. Learn. Res.*, 9:2187–2216, 2008.

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

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

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

M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

Y. Jedra and A. Proutiere. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2676–2681, 2019.

Y. Jedra and A. Proutière. Sample complexity lower bounds for linear system identification. *CoRR*, abs/1903.10343, 2019.

F. Juhász. The asymptotic behaviour of fielder's algebraic connectivity for random graphs. *Discrete Mathematics*, 96:59–63, 1991.

V. Kanade, F. Mallmann-Trenn, and T. Sauerwald. On coalescence time in graphs: When is coalescing as fast as meeting?: Extended abstract. In *Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 956–965, 2019.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken,

editors, *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46128-1.

M. Karoń ski and T. Łuczak. The phase transition in a random hypergraph. *Journal of Computational and Applied Mathematics*, 142(1):125 – 135, 2002. ISSN 0377-0427. doi: https://doi.org/10.1016/S0377-0427(01)00464-2. URL http://www.sciencedirect.com/science/article/pii/S0377042701004642. Probabilistic Methods in Combinatorics and Combinatorial Optimization.

D. Katselis, C. L. Beck, and R. Srikant. Mixing times and structural inference for Bernoulli autoregressive processes. *IEEE Transactions on Network Science and Engineering*, 6(3):364–378, 2019.

E. Kaye and D. Firth. BradleyTerryScalable, 2020. URL https://github.com/EllaKaye/BradleyTerryScalable.

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

A. Khetan and S. Oh. Computational statistical tradeoffs in learning to rank. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS '16)*, pages 739–747, Barcelona, Spain, 2016b.

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

A. Khetan and S. Oh. Computational and statistical tradeoffs in learning to rank. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 739–747. Curran Associates, Inc., 2016d. URL http://papers.nips.cc/paper/6442-computational-and-statistical-tradeoffs-in-learning-to-rank.pdf.

M. Kokkodis, P. Papadimitriou, and P. G. Ipeirotis. Hiring behavior models for online labor markets. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 223–232. ACM, 2015.

E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

A. Kontorovich and R. Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *J. Appl. Probab.*, 51(4):1100–1113, 12 2014.

L. A. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Ann. Probab.*, 36(6):2126–2158, 11 2008.

G. E. Kreindler and H. P. Young. Rapid innovation diffusion in social networks. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10881–10888, 2014.

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

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.

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

W. Li, S. Shrotriya, and A. Rinaldo. ℓ-bounds of the mle in the btl model under general comparison graphs. In J. Cussens and K. Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1178–1187. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/li22g.html.

X. Li, M. Wang, and A. Zhang. Estimation of Markov chain via rank-constrained likelihood. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3033–3042, 10–15 Jul 2018.

T. M. Liggett. *Interacting Particle Systems*. New York: Springer Verlag, 1985.

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

R. B. Lund and R. L. Tweedie. Geometric convergence rates for stochastically ordered Markov chains. *Mathematics of Operations Research*, 21(1):182–194, 1996.

N. V. R. Mahadev and U. N. Peled. *Thereshold Graphs and Related Topics*, volume Annals of Discrete Mathematics. North-Holland, 1995.

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

B. Mark, G. Raskutti, and R. Willett. Network estimation from point process data. *IEEE Transactions on Information Theory*, 65(5):2953–2975, May 2019.

B. Mark, G. Raskutti, and R. Willett. Estimating network structure from incomplete event data. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89, pages 2535–2544, 2019.

A. Maydeau-Olivares. Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, 64(3):325–340, 1999.

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

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

L. Maystre and M. Grossglauser. Robust active ranking from sparse noisy comparisons. *arXiv preprint arXiv:1502.05556*, 2015b.

L. Maystre and M. Grossglauser. Choicerank: Identifying preferences from node traffic in networks. In *Proc. of ICML 2017*, 2017.

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

I. Melnyk and A. Banerjee. Estimating structured vector autoregressive models. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, pages 830–839, 2016.

R. Merris. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 197-198:143 – 176, 1994.

R. Mukherjee, S. Mukherjee, and S. Sen. Detection Thresholds for the β-Model on Sparse Graphs. *ArXiv e-prints*, Aug. 2016.

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

T. Nakata, H. Imahayashi, and M. Yamashita. Probabilistic local majority voting for the agreement problem on finite graphs. In S. ichi Nakano, H. Imai, D. Lee, T. Tokuyama, and T. Asano, editors, *Computing and Combinatorics - 5th Annual International Conference, COCOON 1999, Proceedings*,

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 330–338, Germany, 1999. Springer Verlag.

S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2012a.

S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *Proc. of NIPS 2012*, pages 2483–2491, 2012b.

S. Negahban, S. Oh, and D. Shah. Rankcentrality: Ranking from pair-wise comparisons. *Operations Research*, 65:266–287, 2016.

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

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012c.

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

Y. Nesterov. Gradient methods for minimizing composite objective functions. *Math. Program.*, 140: 125–161, 2013.

P. Netrapalli and S. Sanghavi. Learning the graph of epidemic cascades. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, pages 211–222, 2012.

R. I. Oliveira. On the coalescence time of reversible random walks. *Transactions of the American Mathematical Society*, 364(4):2109–2128, 2012.

R. I. Oliveira and Y. Peres. Random walks on graphs: new bounds on hitting, meeting, coalescing and returning. In *2019 Proceedings of the Meeting on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 119–126, 2019.

P. Pandit, M. Sahraee-Ardakan, A. Amini, S. Rangan, and A. K. Fletcher. Sparse multivariate Bernoulli processes in high dimensions. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89, pages 457–466, 2019.

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

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

J. Pouget-Abadie and T. Horel. Inferring graphs from cascades: A sparse recovery framework. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 977–986, 2015a.

J. Pouget-Abadie and T. Horel. Inferring graphs from cascades: A sparse recovery framework. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 977–986, Lille, France, 07–09 Jul 2015b. PMLR. URL https://proceedings.mlr.press/v37/pouget-abadie15.html.

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, 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.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls, 2009. URL https://arxiv.org/abs/0910.2042.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

T. Rich, K. Hu, and B. Tome. GIFGIF - mapping the emotional language of gifs, 2018. URL http://gifgif.media.mit.edu.

A. Rinaldo, S. Petrovic, and S. E. Fienberg. Maximum likelihood estimation in the β-model. *Ann. Statist.*, 41(3):1085–1110, 06 2013. doi: 10.1214/12-AOS1078. URL https://doi.org/10.1214/12-AOS1078.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

J. S. Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

Y. X. Ruijian Han and K. Chen. A general pairwise comparison model for extremely sparse networks. *Journal of the American Statistical Association*, 0(0):1–11, 2022. doi: 10.1080/01621459.2022.2053137. URL https://doi.org/10.1080/01621459.2022.2053137.

F. Salehi, E. Abbasi, and B. Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems 32*, pages 11982–11992. Curran Associates, Inc., 2019.

J. Schmidt-Pruzan and E. Shamir. Component structure in the evolution of random hypergraphs. *Combinatorica*, 5(1):81–94, Mar 1985. ISSN 1439-6912. doi: 10.1007/BF02579445. URL https://doi.org/10.1007/BF02579445.

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. *Journal Machine Learning Research*, 17(1):2049–2095, Jan. 2016.

M. J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(3):310–313, 1981. ISSN 00359246. URL http://www.jstor.org/stable/2984941.

M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473, 06–09 Jul 2018.

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.

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

A. Soufiani, D. Parkes, and L. Xia. Computing parametric ranking models via rank-breaking. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 360–368, 2014.

H. A. Soufiani, W. Chen, D. C. Parkes, and L. Xia. Generalized method-of-moments for rank aggregation. In *Proc. of NIPS 2013*, 2013.

W. Stadje. The collector's problem with group drawings. *Advances in Applied Probability*, 22(4): 866–882, 1990. doi: 10.2307/1427566.

D. Stasi, K. Sadeghi, A. Rinaldo, S. Petrovic, and S. E. Fienberg. β models for random hypergraphs with a given degree sequence. *ArXiv e-prints*, July 2014.

H. Stern. A continuum of paired comparison models. *Biometrika*, 27(2):265–273, 1990.

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

P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

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

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

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015a. ISSN 1935-8237. URL http://dx.doi.org/10.1561/2200000048.

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

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

V. Verroios, P. Papadimitriou, R. Johari, and H. Garcia-Molina. Client clustering for hiring modeling in work marketplaces. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 2187–2196, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2788589. URL http://doi.acm.org/10.1145/2783258.2788589.

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

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 - Volume 48*, ICML'16, pages 498–506. JMLR.org, 2016a. URL http://dl.acm.org/citation.cfm?id=3045390.3045444.

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, 2016b.

M. Vojnovic, S.-Y. Yun, and K. Zhou. Convergence rates of gradient descent and MM algorithms for Bradley-Terry models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020*, volume 108, Palermo, Italy, 2020.

M. J. Wainwright. *High Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

J. Wang, N. Shah, and R. Ravi. Stretching the effectiveness of mle from accuracy to bias for pairwise comparisons. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 66–76. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/wang20a.html.

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.

R. Wu, J. Xu, R. Srikant, L. Massoulié, M. Lelarge, and B. Hajek. Clustering and inference from pairwise comparisons. *arXiv preprint arXiv:1502.04631*, 2015.

T. Yan and J. Xu. A central limit theorem in the -model for undirected random graphs with a diverging number of vertices. *Biometrika*, 100(2):519–524, 2013.

T. Yan, H. Qin, and H. Wang. Asymptotics in undirected random graph models parameterized by the strengths of vertices. *Statistica Sinica*, 26(1):273–293, 2016.

K. Yasuda and K. Hirai. Upper and lower bounds on the solution of the algebraic Riccati equation. *IEEE Transactions on Automatic Control*, 24(3):483–487, 1979.

J. I. Yellott. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement and the double exponential distribution. *Journal of Mathematical Psychology*, 15: 109–144, 1977.

S.-Y. Yun and A. Proutiere. Community detection via random and adaptive sampling. In *COLT*, pages 138–175, 2014.

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

X.-D. Zhang. The laplacian eigenvalues of graphs: a survey. *arXiv preprint arXiv:1111.2897*, 2011.

X. Zhu and R. Pan. Grouped network vector autoregression. *Statistica Sinica*, 30:1437–1462, 2020.

X. Zhu, R. Pan, G. Li, Y. Liu, and H. Wang. Network vector autoregression. *Ann. Statist.*, 45(3): 1096–1123, 06 2017.

Z. Zhu, X. Li, M. Wang, and A. Zhang. Learning markov models via low-rank optimization, 2019.