https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

On Sinkhorn's Algorithm and Choice Modeling

Zhaonan Qu,[a,b,*] **Alfred Galichon,**[c,d,e] **Wenzhi Gao,**[f] **Johan Ugander**[g]

[a] Martin Tuchman School of Management, New Jersey Institute of Technology, Newark, New Jersey 07102; [b] Data Science Institute, Columbia University, New York, New York 10027; [c] Department of Mathematics, New York University, New York, New York 10012; [d] Department of Economics, New York University, New York, New York 10012; [e] Department of Economics, Sciences Po, 75337 Paris, France; [f] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, California 94305; [g] Department of Management Science and Engineering, Stanford University, Stanford, California 94305

*Corresponding author

Contact: zq2236@columbia.edu, https://orcid.org/0000-0003-1484-1217 (ZQ); alfred.galichon@nyu.edu (AG); gwz@stanford.edu (WG); jugander@stanford.edu, https://orcid.org/0000-0001-5655-4086 (JU)

Received: October 31, 2023
Revised: September 30, 2024
Accepted: April 6, 2025
Published Online in Articles in Advance: July 30, 2025

Area of Review: Optimization

https://doi.org/10.1287/opre.2023.0596

Abstract. For a broad class of models widely used in practice for choice and ranking data based on the Luce choice axiom, including the Bradley–Terry–Luce and Plackett–Luce models, we show that the associated maximum likelihood estimation problems are equivalent to a classic matrix-balancing problem with target row and column sums. This perspective opens doors between two seemingly unrelated research areas and allows us to unify existing algorithms in the choice-modeling literature as special instances or analogs of Sinkhorn's celebrated algorithm for matrix balancing. We draw inspirations from these connections and resolve some open problems on the study of Sinkhorn's algorithm. We establish the global linear convergence of Sinkhorn's algorithm for nonnegative matrices whenever finite scaling matrices exist and characterize its linear convergence rate in terms of the algebraic connectivity of a weighted bipartite graph. We further derive the sharp asymptotic rate of linear convergence, which generalizes a classic result of Knight. To our knowledge, these are the first quantitative linear convergence results for Sinkhorn's algorithm for general nonnegative matrices and positive marginals. Our results highlight the importance of connectivity and orthogonality structures in matrix balancing and Sinkhorn's algorithm, which could be of independent interest. More broadly, the connections that we establish in this paper between matrix balancing and choice modeling could also help motivate further transmission of ideas and lead to interesting results in both disciplines.

Funding: This work was supported in part by Stanford [an interdisciplinary graduate fellowship], the Division of Computing and Communication Foundations [Grant 2143176], and the European Research Council [Grant 866274].

Supplemental Material: All supplemental materials, including the code, data, and files required to reproduce the results, are available at https://doi.org/10.1287/opre.2023.0596.

Keywords: Luce choice models • matrix balancing • algebraic connectivity • linear convergence • optimal transport

1. Introduction

The modeling of choice and ranking data is an important topic across many disciplines. Given a collection of m objects, a universal problem is to aggregate choice or partial ranking data over them to arrive at a reasonable description of the behavior of decision makers, the intrinsic qualities of the objects, or both. Work on such problems dates back over a century at least to the work of Landau (1895), who considered m chess players and a record of their match results against one another, aiming to aggregate the pair-wise comparisons to arrive at a global ranking of all players (Landau 1895, Elo 1978). More generally, comparison data can result from choices from subsets of varying sizes, from partial or complete rankings of objects, or from mixtures of different data types.

The modern rigorous study of comparisons primarily builds on the foundational works of Thurstone (1927) and Zermelo (1929). Both proposed models are based on a numerical "score" for each item (e.g., chess player) but with different specifications of choice probabilities. Zermelo (1929) builds on the intuition that choice probability should be proportional to the score and proposes a iterative algorithm to estimate the scores from pairwise comparison data. As one of the foundational works in this direction, Luce (1959) formalized the multinomial logit (MNL) model of discrete choice starting from the axiom of independence of irrelevant alternatives (IIA).

It states that the relative likelihood of choosing an item j over another item k is independent of the presence of other alternatives. In other words, if S and S' are two subsets of the m alternatives, both containing j and k, and $\Pr(j,S)$ denotes the probability of choosing item j from S, then for $\Pr(k,S), \Pr(k,S') > 0$,

$$\frac{\Pr(j,S)}{\Pr(k,S)} = \frac{\Pr(j,S')}{\Pr(k,S')}.$$

This invariance property, together with a natural condition for zero probability alternatives, is often referred to as Luce's choice axioms. They guarantee that each alternative can be summarized by a nonnegative score s_j such that the probability of choice can be parameterized by

$$\Pr(j,S) = \frac{s_j}{\sum_{k\in S} s_k} \tag{1}$$

for any set S that contains j. The parameters s_j reflect the "intrinsic" value of item j and are unique up to a normalization, which can be set to $\sum_j s_j = 1$. This general choice model includes as a special case the Bradley–Terry–Luce model (BTL) for pair-wise comparisons (Bradley and Terry 1952) and also applies to ranking data when each k-way ranking is broken down into $k-1$ choice observations, where an item is chosen over the set of items ranked lower (Plackett 1975, Hausman and Ruud 1987, Critchlow et al. 1991). The many subsequent works that build on Luce's choice axioms speak to its fundamental importance in choice modeling. Other works have also sought to address the limitations of Luce choice models. Prominent among them are probit models (Thurstone 1927, Berkson 1944), random utility models (McFadden and Train 2000), context-dependent models (Batsell and Polking 1985, Seshadri et al. 2020), and behavioral models from psychology (Tversky 1972).

Matrix balancing (or scaling), meanwhile, is a seemingly unrelated mathematical problem with an equally long history. In its most common form that we study in this paper, the problem seeks positive diagonal matrices D^0, D^1 of a given (entry-wise) nonnegative matrix $A \geq 0$ such that the scaled matrix D^0AD^1 has row and column sums equal to some prescribed positive marginals $\boldsymbol{p}, \boldsymbol{q}$:

$$\begin{aligned}(D^0AD^1)\mathbf{1} &= \boldsymbol{p},\\ (D^0AD^1)^T\mathbf{1} &= \boldsymbol{q}.\end{aligned} \tag{2}$$

Over the years, numerous applications and problems across different domains, including statistics (Yule 1912, Deming and Stephan 1940, Ireland and Kullback 1968), economics (Stone 1962, Bacharach 1970, Galichon and Salanié 2021), transportation networks (Kruithof 1937, Lamond and Stewart 1981, Chang et al. 2024), optimization (Bregman 1967b, Ruiz 2001), and machine learning (Cuturi 2013, Peyré and Cuturi 2019), have found themselves essentially solving a new incarnation of the old matrix-balancing problem, which attests to its universality and importance.

A major appeal of the matrix-balancing problem lies in the simplicity and elegance of its popular solution method, widely known as Sinkhorn's algorithm (Sinkhorn 1964). Observe that it is easy to scale the rows or columns of A such that the resulting matrix satisfies one of the two marginal constraints in (2). However, it is more difficult to construct scalings D^0, D^1 that simultaneously satisfy both constraints. Sinkhorn's algorithm (Algorithm 1) simply alternates between updating the scalings D^0 and D^1 to satisfy one of the two marginal conditions in the hope of converging to a solution, leading to lightweight implementations that have proven to be effective for practical problems of massive size. In particular, Sinkhorn's algorithm has gained much popularity in the recent decade thanks to its empirical success at approximating optimal transport (OT) distances (Cuturi 2013, Altschuler et al. 2017), which are bedrocks of important recent topics in operations research, such as Wasserstein distributionally robust optimization (Esfahani and Kuhn 2018; Blanchet et al. 2019, 2022; Kuhn et al. 2019; Gao and Kleywegt 2023).

Despite the widespread popularity of Sinkhorn's algorithm, its convergence behavior is yet to be fully understood. In particular, although there have been extensive studies of convergence, many focus on the setting when the matrix $A > 0$ (i.e., entry-wise strictly positive), which includes most OT problems. In contrast, other applications of matrix balancing, particularly those with network structures, have $A \geq 0$ with zero elements, and are, therefore, potentially sparse. In this setting, quantitative analyses are less common and more fragmented, employing different assumptions whose connections and distinctions remain less clear. On one hand, works such as Kalantari et al. (2008), Chakrabarty and Khanna (2021), and Léger (2021) have established global (that is, true for all iterations $t \geq 1$) sublinear convergence results (i.e., convergence to an ε accuracy solution requires a total number of iterations that is polynomial in $1/\varepsilon$). On the other hand, Knight (2008) establishes local (and more specifically, asymptotic) linear convergence for square matrix $A \geq 0$ and uniform marginals $\boldsymbol{p}, \boldsymbol{q}$. In other words, as $t \to \infty$, solution accuracy at iteration $t+1$ improves over that at iteration t with a constant factor. Furthermore, a general result in Luo and Tseng (1992) implies global linear convergence of Sinkhorn's algorithm (i.e., convergence to an ε accuracy solution requires iterations polynomial in $\log(1/\varepsilon)$). However, their result is implicit and does not characterize the dependence on problem parameter and structure as those in the sublinear results.

These results leave open several questions on the convergence of Sinkhorn's algorithm. First, when does a quantitative global linear convergence result exist for

$A \geq 0$, and how do we characterize the global convergence rate in terms of the problem primitives? Second, how do we characterize the sharp (i.e., best-possible) asymptotic linear convergence rate λ that is applicable to general nonnegative A and nonuniform $\boldsymbol{p}, \boldsymbol{q}$? Third, how do we reconcile and clarify the linear versus sublinear convergence results under different assumptions on the problem structure? Given that many applications of matrix balancing with network structures correspond to the setting with sparse $A \geq 0$, such as transportation and trade, it is, therefore, important to better understand the convergence of Sinkhorn's algorithm in this setting.

In this paper, we provide answers to these open questions in matrix balancing. Surprisingly, the inspirations for our solutions come from results in the seemingly unrelated topic of choice modeling. Our main contributions are summarized below.

Our first set of contributions, which we detail in Section 4, is recognizing Luce choice models as yet another instance where a central problem reduces to that of matrix balancing. More precisely, we formally establish the equivalence between the maximum likelihood estimation of Luce choice models and matrix-balancing problems with an $A \geq 0$ with binary elements (Theorem 1). We also clarify the relations and distinctions between problem assumptions in the two literatures (Proposition 1). More importantly, we demonstrate that classic and new algorithms from the choice literature, including those of Zermelo (1929), Dykstra (1956), Ford (1957), Hunter (2004), Maystre and Grossglauser (2017), and Agarwal et al. (2018), can be viewed as special cases or analogs of Sinkhorn's algorithm when applied to various problems in the choice setting (Theorem 2). These intimate mathematical and algorithmic connections allow us to provide a unifying perspective on works from both areas. More broadly, they enable researchers to import insights and tools from one domain to the other. In particular, recent works on choice modeling (Shah et al. 2015, Seshadri et al. 2020, Vojnović et al. 2020) have highlighted the importance of algebraic connectivity (Fiedler 1973, Spielman 2007) of the data structure for efficient parameter learning, which motivates us to also consider this quantity in the convergence analysis of Sinkhorn's algorithm.

Our next set of contributions, detailed in Section 5, is establishing novel convergence bounds on Sinkhorn's algorithm, drawing from the connections to choice modeling that we establish. First, we provide a global linear convergence bound for Sinkhorn's algorithm whenever the matrix-balancing problem has a finite solution pair D^0, D^1 (Theorem 3). We characterize the global convergence rate in terms of the algebraic connectivity of the weighted bipartite graph whose biadjacency matrix is precisely A. To our knowledge, this result is the first to highlight the fundamental role of algebraic connectivity in the study of matrix balancing with sparse matrices. In addition, we characterize the asymptotic linear rate of convergence in terms of the scaled matrix D^0AD^1 with target marginals $\boldsymbol{p}, \boldsymbol{q}$, generalizing a result of Knight (2008) for uniform marginals and square matrices (Theorem 4). This result employs a more explicit analysis that exploits an intrinsic orthogonality structure of Sinkhorn's algorithm. We also clarify the convergence behavior of Sinkhorn's algorithm under two regimes. When a finite scaling pair D^0, D^1 exists, Sinkhorn's algorithm converges linearly; otherwise, it only converges sublinearly under the minimal conditions required for convergence (Proposition 3).

Besides the contributions above, we further discuss connections between Sinkhorn's algorithm and topics in optimization and choice modeling in Online Appendix B. For example, interpreting Sinkhorn's algorithm as a distributed optimization algorithm on a bipartite graph helps explain the importance of the spectral properties of the graph on its convergence. Inspired by Bayesian regularization of Luce choice models using gamma priors, we also design a regularized Sinkhorn's algorithm in Online Appendix C that is guaranteed to converge even when the standard algorithm does not, which is not uncommon when the data are very sparse and there are measurement errors.

We believe that the connections that we establish in this paper between choice modeling and matrix balancing can lead to further interesting results in both disciplines and are, therefore, relevant to researchers working on related topics. In particular, the fundamental role of algebraic connectivity in the study of matrix balancing for sparse matrices goes beyond quantifying the algorithmic efficiency of Sinkhorn's algorithm. See, for example, Chang et al. (2024), which quantifies the statistical efficiency of a network traffic model using algebraic connectivity.

2. Related Work

This section includes an extensive review of related works in choice modeling and matrix balancing. Well-versed readers may skip ahead to the mathematical preliminaries (Section 3) and our core results (Sections 4 and 5).

2.1. Choice Modeling

Methods for aggregating choice and comparison data usually take one of two closely related approaches: maximum likelihood estimation of a statistical model or ranking according to the stationary distributions of a random walk on a Markov chain. Recent connections between maximum likelihood and spectral methods have put these two classes of approaches in increasingly close conversation with each other.

2.1.1. Spectral Methods. The most well-known spectral method for rank aggregation is perhaps the PageRank

algorithm (Page et al. 1999), which ranks web pages based on the stationary distribution of a random walk on a hyperlink graph. The use of stationary distributions also features in the work of Dwork et al. (2001); the rank centrality (RC) algorithm (Negahban et al. 2012, 2017), which generates consistent estimates for the Bradley–Terry–Luce pair-wise comparison model under assumptions on the sampling frame; and the Luce spectral ranking (LSR) and iterative LSR algorithms of Maystre and Grossglauser (2015) for choices from pairs as well as larger sets. Following that work, Agarwal et al. (2018) proposed the accelerated spectral ranking algorithm with provably faster mixing times than RC and LSR and better sample complexity bounds than Negahban et al. (2017). Knight (2008) is an intriguing work partially motivated by Page et al. (1999) that applies Sinkhorn's algorithm, which is central to the current work, to compute authority and hub scores similar to those proposed by Kleinberg (1999) and Tomlin (2003), although the focus in Knight (2008) is on Markov chains rather than maximum likelihood estimation of choice models. For ranking data, Soufiani et al. (2013) decompose rankings into pair-wise comparisons and develop consistent estimators for Plackett–Luce models based on a generalized method of moments. Other notable works that make connections between Markov chains and choice modeling include Blanchet et al. (2016) and Ragain and Ugander (2016).

2.1.2. Maximum Likelihood Methods. Maximum likelihood estimation of the Bradley–Terry–Luce model dates back to Zermelo (1929), Dykstra (1956), and Ford (1957), which all give variants of the same iterative algorithm and prove its convergence to the maximum likelihood estimator (MLE) when the directed comparison graph is strongly connected. Much later, Hunter (2004) observed that their algorithms are instances of a class of minorization-maximization or majorization-minimization (MM) algorithms and developed MM algorithms for the Plackett–Luce model for ranking data among others. Vojnović et al. (2020, 2023) further investigated the convergence rate of the MM algorithm for choice models, quantifying it in terms of the algebraic connectivity of the comparison graph. Newman (2023) proposes an alternative to the classical iterative algorithm for pair-wise comparisons based on a reformulated moment condition, achieving impressive empirical speedups. Negahban et al. (2012) is arguably the first work that connects maximum likelihood estimation to Markov chains followed by Maystre and Grossglauser (2015), whose spectral method is based on a balance equation interpretation of the optimality condition. Kumar et al. (2015) consider the problem of inverting the stationary distribution of a Markov chain and embed the maximum likelihood problem of the Luce choice model into this framework, where the MLEs parameterize the desired transition matrix. Maystre and Grossglauser (2017) consider the estimation of a network choice model with similarly parameterized random walks. Lastly, a vast literature in econometrics on discrete choice also considers different aspects of the ML estimation problem. In particular, the present paper is related to the Berry–Levinsohn–Pakes (BLP) framework of Berry et al. (1995), which is well known in econometrics. The matrix-balancing interpretation of maximum likelihood estimation of choice models that we develop in this paper connects many of the aforementioned works.

Besides optimization problems related to maximum likelihood estimation, there have also been extensive studies on the statistical properties of maximum likelihood estimates themselves (Hajek et al. 2014, Rajkumar and Agarwal 2014). In particular, a line of recent works has highlighted the importance of algebraic connectivity—as quantified by the Fiedler eigenvalue (Fiedler 1973, Spielman 2007)—on the statistical efficiency of the MLEs. Shah et al. (2015) is the first to recognize this significance of data structure for the statistical efficiency of the BTL model, which they refer to as "topology dependence." As a by-product of analysis for a context-dependent generalization of the Luce choice model, Seshadri et al. (2020) obtain tight expected risk and tail risk bounds for the MLEs of Luce choice models (which they call MNL) and Plackett–Luce ranking models in terms of the algebraic connectivity, extending and improving upon previous works by Hajek et al. (2014), Shah et al. (2015), and Vojnović and Yun (2016). Other works with tight risk bounds on the BTL model include the works of Hendrickx et al. (2020) and Bong and Rinaldo (2022), who also provide the first high-probability guarantees for the existence of finite MLEs of the BTL model under conditions on a Fiedler eigenvalue. Interestingly, the statistical significance of algebraic connectivity has also been highlighted in models of networks in econometrics and machine learning by the works of De Paula (2017), Jochmans and Weidner (2019), and Chang et al. (2024) among others. Our present work is primarily concerned with the optimization aspects of the maximum likelihood estimation of choice models. Nevertheless, the statistical importance of algebraic connectivity in the aforementioned works also provides motivation for us to leverage it in our convergence analysis of Sinkhorn's algorithm for matrix balancing.

Lastly, a short note on terminology. Even though a choice model based on (1) is technically a "multinomial logit model" with only intercept terms (McFadden 1973), there are subtle differences. When (1) is applied to model ranking and choice data with distinct items, each observation i usually consists of a possibly different subset S_i of the universe of all alternatives so that there is a large number of different configurations of the choice menu in the data set. On the other hand, common applications of multinomial logit models, such as classification models in statistics and machine learning (Bishop and Nasrabadi 2006) and discrete choice models in

econometrics (McFadden 1973), often deal with repeated observations consisting of the same number of alternatives. However, these alternatives now possess "characteristics" that vary across observations, which are often mapped parametrically to the scores in (1). In this paper, we primarily use the term Luce choice model to refer to Model (4), although it is also called an MNL model in some works. We refrain from using the term MNL to avoid confusion with parametric models for featurized items used in ML and econometrics.

2.2. Matrix Balancing

Matrix balancing (or scaling) is an important topic in optimization and numerical linear algebra that underlies a diverse range of applications. The particular question of scaling rows and columns of a matrix A so that the resulting matrix has target row and column norms $\boldsymbol{p}, \boldsymbol{q}$ was studied as early as the 1930s and continues to interest researchers from different disciplines today. The present paper only contains a partial survey of the vast literature on this topic. Online Appendix D provides a summary of some popular applications to illustrate the ubiquity of the matrix-balancing problem. Schneider and Zenios (1990) and Idel (2016) also provide excellent discussions of many applications.

The standard iterative algorithm for the matrix-balancing problem that we study in this paper has been rediscovered independently quite a few times. As a result, it has domain-dependent names, including the iterative proportional fitting procedure (Deming and Stephan 1940), biproportional fitting (Bacharach 1965), and the RAS (the origin of the name is unknown) algorithm (Stone 1962), but it is perhaps most widely known as Sinkhorn's algorithm (Sinkhorn 1964). A precise description can be found in Algorithm 1. Sinkhorn's algorithm is also closely related to relaxation and coordinate descent-type methods for solving the dual of entropy optimization problems (Bregman 1967b, Cottle et al. 1986, Tseng and Bertsekas 1987, Luo and Tseng 1992) as well as message passing and belief propagation algorithms in distributed optimization (Balakrishnan et al. 2004, Agarwal et al. 2018).

The convergence behavior of Sinkhorn's algorithm in different settings has been extensively studied by Sinkhorn (1964), Bregman (1967a), Lamond and Stewart (1981), Franklin and Lorenz (1989), Ruschendorf (1995), Kalantari et al. (2008), Knight (2008), Pukelsheim and Simeone (2009), Altschuler et al. (2017), Dvurechensky et al. (2018), Di Marino and Gerolin (2020), Chakrabarty and Khanna (2021), Léger (2021), and Carlier (2022) among many others. For A with strictly positive entries, Franklin and Lorenz (1989) establish the global linear convergence of Sinkhorn's algorithm in the Hilbert projective metric d (Bushell 1973). More precisely, if $\boldsymbol{r}^{(t)}$ denotes the row sum of the scaled matrix after t iterations of Sinkhorn's algorithm that enforce column constraints, then

$$d(\boldsymbol{r}^{(t)}, \boldsymbol{p}) \leq \lambda^t \cdot d(\boldsymbol{r}^{(0)}, \boldsymbol{p}) \tag{3}$$

for some $\lambda \in (0,1)$ dependent on A. On the other hand, works such as Kalantari and Khachiyan (1993), Altschuler et al. (2017), and Dvurechensky et al. (2018) develop complexity bounds on the number of iterations required for the ℓ^1 distance $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1 \leq \varepsilon$ for a given $\varepsilon > 0$. Although these bounds imply a convergence that is sublinear (i.e., $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1 = \mathcal{O}(1/t)$), their focus is on optimal dependence on problem size and dimension. An important class of problems with $A > 0$ is entropy-regularized optimal transport (Cuturi 2013), where A is of the form $A = \exp(-c/\gamma)$ with a finite cost function (matrix) c (i.e., A strictly positive everywhere). In this setting, convergence of Sinkhorn's algorithm in discrete and continuous problems has been studied by Altschuler et al. (2017), Di Marino and Gerolin (2020), Léger (2021), and Ghosal and Nutz (2025) among others. The linear convergence of Sinkhorn's algorithm for $A > 0$ has also been extended to the multimarginal continuous setting by Carlier (2022), building on the work of Di Marino and Gerolin (2020).

However, the matter of convergence is more delicate when the matrix contains zero entries, and additional assumptions on the problem structure are required to guarantee the existence of scalings D^0, D^1 and the convergence of Sinkhorn's algorithm. For nonnegative A, convergence is first established by Sinkhorn and Knopp (1967) in the special case of square $A \geq 0$ and uniform $\boldsymbol{p} = \boldsymbol{q} = \mathbf{1}_n = \mathbf{1}_m$. Their necessary and sufficient condition is that A has support (i.e., there exists a permutation σ such that the "diagonal" $(A_{1\sigma(1)}, A_{2\sigma(2)}, \ldots, A_{n\sigma(n)})$ is strictly positive). Soules (1991) and Achilles (1993) further show that the convergence is linear if and only if the stronger condition of total support holds (i.e., any nonzero entry of A must be in $(A_{1\sigma(1)}, A_{2\sigma(2)}, \ldots, A_{n\sigma(n)})$ for some permutation σ). Knight (2008) provides a tight asymptotic linear convergence rate in terms of the subdominant (second-largest) singular value of the scaled doubly stochastic matrix $D^0 A D^1$. The convergence in Knight (2008) is measured by some implicit distance to the target marginals $\boldsymbol{p}, \boldsymbol{q}$. However, no asymptotic linear convergence rate is previously known for nonsquare $A \geq 0$ and nonuniform marginals.

For general nonnegative matrices and nonuniform marginals, the necessary and sufficient conditions on A in the matrix-balancing problem that generalize that of Sinkhorn and Knopp (1967) have been studied by Thionet (1964), Bacharach (1965), Brualdi (1968), Menon (1968), Djoković (1970), Sinkhorn (1974), Balakrishnan et al. (2004), and Pukelsheim and Simeone (2009) among others, and convergence of Sinkhorn's algorithm under these conditions is well known. Connecting Sinkhorn's algorithm to dual coordinate descent for entropy optimization, Luo and Tseng (1992) show that the dual

optimality gap, defined in Equation (28), converges linearly globally with some unknown rate λ when finite scalings D^0, D^1 exist. However, their result is implicit, and there are no results that quantify the global linear rate λ, even for special classes of nonnegative matrices. When convergence results for $A > 0$ in previous works are applied to nonnegative matrices, the bounds often blow up or become degenerate as soon as $\min_{ij} A_{ij} \downarrow 0$. For example, in (3), the contraction factor $\lambda \to 1$ when A contains zero entries. When $\min_{ij} A_{ij} = 0$, complexity bounds on $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1$ and $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_2$ have been established for Sinkhorn's algorithm (for example, by the works of Kalantari et al. 2008 and Chakrabarty and Khanna 2021), with polynomial dependence on $1/\varepsilon$ (i.e., sublinear convergence). Under the minimal condition that guarantees the convergence of Sinkhorn's algorithm, Léger (2021) gives a quantitative global sublinear bound on the KL (Kullback-Leibler) divergence between $\boldsymbol{r}^{(t)}$ and $\boldsymbol{p}$ in the continuous setting for general probability distributions, which include nonnegative matrices $A \geq 0$.

It, therefore, remains to reconcile the various results on Sinkhorn's algorithm for $A \geq 0$ and characterize the global and asymptotic linear convergence rates for nonnegative A. Our results precisely fill these gaps left by previous works. The global linear convergence result in Theorem 3 establishes a contraction like (3) for the optimality gap whenever finite scalings D^0, D^1 exist and characterize the convergence rate λ in terms of the algebraic connectivity. Moreover, the asymptotic linear rate in Theorem 4 directly extends the result of Knight (2008). See Table 1 for a detailed summary and comparison of the convergence results in previous works and this paper. The dependence of Sinkhorn's convergence rate on spectral properties of graphs can be compared with convergence results in the literature on decentralized optimization and gossip algorithms, where a spectral gap quantifies the convergence rate (Boyd et al. 2006, Xiao et al. 2007).

Lastly, we note that other algorithms with better complexities have been developed for the matrix-balancing problem utilizing, for example, the ellipsoid algorithm (Kalantari and Khachiyan 1996) and geometric programming (Nemirovski and Rothblum 1999), interior point algorithms (Cohen et al. 2017, Chen et al. 2022), or customized first-/second-order techniques (Linial et al. 1998, Allen-Zhu et al. 2017). However, despite having better theoretical complexities, most of these algorithms have yet to be implemented practically. Sinkhorn's algorithm, on the other hand, remains an attractive choice in practice because of its simplicity, robustness, and parallelization capabilities.

3. Preliminaries on Choice Modeling and Matrix Balancing

We start by providing brief but self-contained introductions to the two main subjects of this paper, choice modeling and matrix balancing, including their respective underlying mathematical problems and assumptions. Then, we formally establish their equivalence in Section 4.

3.1. Maximum Likelihood Estimation of Luce Choice Models

In the Luce choice-modeling framework, we have n observations $\{(j_i, S_i)\}_{i=1}^n$, each consisting of a choice set $S_i \subseteq \{1, \dots, m\} = [m]$ that is a subset of the total m alternatives/items/objects and the alternative selected, denoted by $j_i \in S_i$. The choice probability is prescribed by Luce's axiom of choice given model parameter $s \in \mathbb{R}^m_{++}$ in the interior of the probability simplex Δ_m,

$$\Pr(j_i, S_i) = \frac{s_{j_i}}{\sum_{k \in S_i} s_k},$$

and the likelihood of the observed data is thus given by

$$L(s; \{(j_i, S_i)\}_{i=1}^n) := \prod_{i=1}^n \frac{s_{j_i}}{\sum_{k \in S_i} s_k}. \tag{4}$$

A popular method to estimate $s = \{s_1, \dots, s_m\}$ is the maximum likelihood estimation approach, which maximizes the log likelihood

$$\ell(s) := \log L(s) = \sum_{i=1}^n \left(\log s_{j_i} - \log \sum_{k \in S_i} s_k \right) \tag{5}$$

over the interior of the probability simplex. Note that the choice sets S_i can vary across i. In other words, in each observation, the choice is made from a potentially distinct set of alternatives. This feature of the problem turns out to be important for both the algorithmic efficiency of computing the maximizers to (5) as well as the statistical efficiency of the resulting MLEs, which can be quantified by a measure of connectivity of the data structure. We will elaborate on these points shortly. For now, we focus on the existence and uniqueness of MLE.

If we reparameterize $\exp(u_j) = s_j$, it is obvious that (5) is concave in u. However, to ensure that the log-likelihood (5) has a unique maximizer in the interior of the simplex, additional assumptions on the comparison structure of the data set $\{(j_i, S_i)\}_{i=1}^n$ are needed. The following classic condition is necessary and sufficient for the maximum likelihood problem to be well posed.

Assumption 1 (Strong Connectivity). *In any partition of $[m]$ into two nonempty subsets S and its complement S^C, some $j \in S$ is selected at least once over some $k \in S^C$. Equivalently, the directed comparison graph, with items as vertices and an edge $j \to k$ if and only if k is selected in some S_i for which $j, k \in S_i$, is strongly connected.*

Assumption 1 is standard in the literature (Hunter 2004, Noothigattu et al. 2020) and appeared as early as the works of Zermelo (1929) and Ford (1957) for pairwise comparisons. Hunter (2004) shows that Assumption 1 is necessary and sufficient for the upper

Table 1. Summary of Some Convergence Results on Sinkhorn's Algorithm

Relevant work	Convergence statement	λ	A	p, q
Franklin and Lorenz (1989)	$d_{\text{Hilbert}}(\boldsymbol{r}^{(t)}, \boldsymbol{p}) \le \lambda^t d_{\text{Hilbert}}(\boldsymbol{r}^{(0)}, \boldsymbol{p})$	$\kappa^2(A)$	$A > 0$, rectangular	Uniform
Luo and Tseng (1992)	$g^{(t)} - g^* \le \lambda^t (g^{(0)} - g^*)$	Unknown	$A \ge 0$, rectangular	General
Knight (2008)	$\|\boldsymbol{r}^{(t+1)} - \boldsymbol{p}\|_* / \|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_* \to \lambda$	$\sigma_2^2(\hat{A})$	$A \ge 0$, square	Uniform
Altschuler et al. (2017)	$\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1 \le 2\sqrt{\frac{\lambda}{t}}$	$\log\left(\frac{\sum_{ij} A_{ij}}{\min_{ij} A_{ij}}\right)$	$A > 0$, rectangular	General
Léger (2021)	$D_{\text{KL}}(\boldsymbol{r}^{(t)} \| \boldsymbol{p}) \le \frac{\lambda}{t}$	$D_{\text{KL}}(\hat{A} \| A)$	$A \ge 0$, continuous	General
Current work, asymptotic	$\|\frac{\boldsymbol{r}^{(t+1)}}{\sqrt{\boldsymbol{p}}} - \sqrt{\boldsymbol{p}}\|_2 / \|\frac{\boldsymbol{r}^{(t)}}{\sqrt{\boldsymbol{p}}} - \sqrt{\boldsymbol{p}}\|_2 \to \lambda$	$\lambda_2(\tilde{A}^T \tilde{A})$	$A \ge 0$, rectangular	General
Current work, global	$g^{(t)} - g^* \le \lambda^t (g^{(0)} - g^*)$	$1 - c_B \lambda_{-2}(\mathcal{L})/l$	$A \ge 0$, rectangular	General

Notes. Throughout, assume that $\|\boldsymbol{p}\|_1 = \|\boldsymbol{q}\|_1 = 1$. Define $r^{(t)} := A^{(t)} \mathbf{1}_m$, where $A^{(t)}$ is the scaled matrix after t Sinkhorn iterations. In Franklin and Lorenz (1989), $\kappa(A) = \frac{\theta(A)^{1/2} - 1}{\theta(A)^{1/2} + 1}$, where $\theta(A)$ is the diameter of A in the Hilbert metric. The norm in Knight (2008) is not explicitly specified, and $\sigma_2(\hat{A})$ denotes the second-largest singular value of the scaled doubly stochastic matrix $\hat{A}$. The bound in Altschuler et al. (2017) is originally stated as the complexity bound that $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1 \le \varepsilon$ in $t = \mathcal{O}\left(\varepsilon^{-2} \log\left(\frac{\sum_{ij} A_{ij}}{\min_{ij} A_{ij}}\right)\right)$ iterations, whereas the original result in Chakrabarty and Khanna (2021) is $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1 \le \varepsilon$ in $t = \mathcal{O}\left(\varepsilon^{-2} \log\left(\frac{\Delta \max_{ij} A_{ij}}{\min_{ij, A_{ij} > 0} A_{ij}}\right)\right)$ iterations, where $\Delta := \max_j |i \in [n] : A_{ij} > 0|$. The result in Léger (2021) applies more generally to couplings of probability distributions. In our asymptotic result, $\lambda_2(\tilde{A}^T \tilde{A})$ is the second-largest eigenvalue of $\tilde{A} := \mathcal{D}(1/\sqrt{\boldsymbol{p}}) \cdot \hat{A} \cdot \mathcal{D}(1/\sqrt{\boldsymbol{q}})$. In our global bound, $g^{(t)} = g(\boldsymbol{d}^{0(t)}, \boldsymbol{d}^{1(t)})$, whereas g^* is the minimum value of (16). $\lambda_{-2}(\mathcal{L})$ is the second-smallest eigenvalue of the Laplacian of the bipartite graph defined by A, $l = \min\{\max_j (A^T \mathbf{1}_n)_j, \max_i (A \mathbf{1}_m)_i\}$, $c_B = \exp(-4B)$, and B is a bound on the initial sublevel set, which is finite if and only if Assumption 3 holds.

compactness of (5), which guarantees the existence of a maximizer in the interior of the probability simplex. In fact, when an interior maximizer exists, it is also unique because Assumption 1 implies the following weaker condition, which guarantees the strict concavity of (5).

Assumption 2 (Connectivity). *In any partition of* $[m]$ *into two nonempty subsets* S *and* S^C, *some* $j \in S$ *and some* $k \in S^C$ *appear in the same choice set* S_i *for some* i.

The intuitions provided by Ford (1957) and Hunter (2004) are helpful for understanding Assumptions 1 and 2. If items from some $S \subsetneq [m]$ are never compared with those in S^C (i.e., never appeared together in any choice set S_i), it is impossible to rank across the two subsets. In this case, we can rescale the relative weights of S and S^C of an interior maximizer and obtain another maximizer. On the other hand, if items in S are always preferred to those in S^C, we can increase the likelihood by scaling s_j for items $j \in S^C$ toward zero, and no maximizer in the interior of the probability simplex exists. Nevertheless, a boundary solution can still exist. This case turns out to be important in the present work; in the equivalent matrix-balancing problem, it corresponds to the slow-down regime of Sinkhorn's algorithm, where scalings diverge but the scaled matrix converges (Section 5).

Assumption 2 also has a concise graph-theoretic interpretation. Define the weighted undirected comparison graph G_c on m vertices with adjacency matrix A^c given by

$$A^c_{jk} = \begin{cases} 0 & j = k \\ |\{i \in [n] | j, k \in S_i\}| & j \ne k. \end{cases} \tag{6}$$

In other words, there is an undirected edge between j and k if and only if they are both included in some choice set S_i, with the edge weight equal to the number of their co-occurrences, which could be zero. We can verify that Assumption 2 precisely requires G_c to be connected.

Remark 1 (Importance of Graph Connectivity). Under the standard Assumptions 1 and 2, previous works have studied the statistical efficiency of the MLE (Hajek et al. 2014, Shah et al. 2015, Seshadri et al. 2020) as well as the computational efficiency of the MM algorithm for computing the MLE (Vojnović et al. 2020). In both cases, the algebraic connectivity of G_c (Fiedler 1973), quantified by the second-smallest eigenvalue of the graph Laplacian of G_c, plays an important role. See Online Appendix A for more details. The importance of spectral properties for parameter learning in data with graph or matrix structures has appeared as early as Kendall and Smith (1940) and in the classic work of Keener (1993) on ranking sports teams as well as in works in economics (Abowd et al. 1999, Jochmans and Weidner 2019). These results, together with the connections that we establish in this paper between choice modeling and matrix balancing, inspire us to also quantify the convergence of Sinkhorn's algorithm using the algebraic connectivity of a bipartite graph defined in (7).

3.2. The Canonical Matrix-Balancing Problem

Matrix balancing is a classic problem that shows up in a wide range of disciplines. See Online Appendix D for a short survey on some applications. The underlying mathematical problem can be stated concisely in matrix form as follows.

Given positive vectors $\boldsymbol{p} \in \mathbb{R}^n_{++}, \boldsymbol{q} \in \mathbb{R}^m_{++}$ with $\sum_i \boldsymbol{p}_i = \sum_j \boldsymbol{q}_j = c > 0$, which without loss of generality, can be set to $c = 1$, and a nonnegative matrix $A \in \mathbb{R}^{n\times m}_+$, find positive diagonal matrices D^1, D^0 satisfying the conditions $D^1 A D^0 \cdot \mathbf{1}_m = \boldsymbol{p}$ and $D^0 A^T D^1 \cdot \mathbf{1}_n = \boldsymbol{q}$.

We, henceforth, refer to the above as the "canonical" matrix-balancing problem. Other variants of the problem replace the row and column sums (the 1-norm) with other norms (Bauer 1963, Ruiz 2001). Note that for any $c > 0$, $(D^0/c, cD^1)$ is also a solution whenever (D^0, D^1) is. A finite positive solution (D^0, D^1) to the canonical matrix-balancing problem is often called a direct scaling.

The structure of the matrix-balancing problem suggests a simple iterative scheme; starting from any initial positive diagonal D^0, invert $D^1 A D^0 \mathbf{1}_m = \boldsymbol{p}$ using $\boldsymbol{p}/(AD^0\mathbf{1}_m)$ to update D^1. Then, invert $D^0 A^T D^1 \mathbf{1}_n = \boldsymbol{q}$ using $\boldsymbol{q}/(A^T D^1 \mathbf{1}_n)$ to compute the new estimate of D^0, and repeat the procedure, leading to a solution if it converges. Here, divisions involving two vectors of the same length are entry wise. This simple iterative scheme is precisely Sinkhorn's algorithm, described in Algorithm 1, where vectors $\boldsymbol{d}^0, \boldsymbol{d}^1$ are the diagonal elements of D^0, D^1.

An important dichotomy occurs depending on whether the entries of A are strictly positive. If A contains no zero entries, then direct scalings and a unique scaled matrix $D^1 A D^0$ always exist (Sinkhorn 1964). Moreover, Sinkhorn's algorithm converges linearly (Franklin and Lorenz 1989). On the other hand, when A contains zero entries, the problem becomes more complicated. Additional conditions are needed to guarantee meaningful solutions, and the convergence behavior of Sinkhorn's algorithm is less clearly understood. Well posedness of the matrix-balancing problem has been studied by Brualdi (1968), Sinkhorn (1974), and Pukelsheim and Simeone (2009) among others, who characterize the following equivalent existence conditions.

Assumption 3 (Strong Existence). (a) *There exists a nonnegative matrix $A' \in \mathbb{R}^{n\times m}_+$ with the same zero patterns as A and with row and column sums $\boldsymbol{p}$ and $\boldsymbol{q}$. Or, equivalently,* (b) *for every pair of sets of indices $N \subsetneq [n]$ and $M \subsetneq [m]$ such that $A_{ij} = 0$ for $i \notin N$ and $j \in M$, $\sum_{i\in N} \boldsymbol{p}_i \geq \sum_{j\in M} \boldsymbol{q}_j$, with equality if and only if $A_{ij} = 0$ for all $i \in N$ and $j \notin M$ as well.*

It is well known in the matrix-balancing literature that the above two conditions are equivalent and that a positive finite solution (D^0, D^1) to the canonical problem exists if and only if they hold. See, for example, Pukelsheim and Simeone (2009, theorem 6). Assumption 3 also guarantees the convergence of Sinkhorn's algorithm. However, it is not a necessary condition. In other words, Sinkhorn's algorithm could converge even if the matrix-balancing problem does not admit a direct scaling. This phenomenon turns out to be important in characterizing the convergence rate, which we study in Section 5.

Clearly, Assumption 3(a) is the minimal necessary condition when a solution to the matrix-balancing problem exists and trivially holds when $A > 0$ (take, for example, A' as the Kronecker product of $\boldsymbol{p}, \boldsymbol{q}$). Assumption 3(b) is closely connected to conditions for perfect matchings in bipartite graphs (Hall 1935, Galichon and Salanié 2021). In flow networks (Ford and Fulkerson 1956, 1957; Gale 1957), it is a capacity constraint that guarantees that the maximum flow on a weighted bipartite graph is equal to $\sum_i \boldsymbol{p}_i = \sum_j \boldsymbol{q}_j$ and with positive flow on every edge (Idel 2016). The weighted bipartite graph, denoted by G_b, is important in this paper. Its adjacency matrix $A^b \in \mathbb{R}^{(n+m)\times(n+m)}$ can be represented concisely using A as

$$A^b := \begin{bmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{bmatrix}, \tag{7}$$

and A is sometimes called the biadjacency matrix of G_b. See Online Appendix A for more information. Just like in the choice setting, where the connectivity of the undirected comparison graph G_c plays an important role, the connectivity of G_b turns out to be important for the linear convergence rate of Sinkhorn's algorithm (see Section 5).

Lastly, the necessary and sufficient condition for the uniqueness of finite scalings essentially requires that A is not block diagonal and precisely guarantees that G_b is connected.

Assumption 4 (Uniqueness). *D^0 and D^1 are unique modulo normalization if and only if A is indecomposable* (*i.e., there does not exist permutation matrices P, Q such that PAQ is block diagonal*).

Algorithm 1 (Sinkhorn's Algorithm)

Input: $A, \boldsymbol{p}, \boldsymbol{q}, \epsilon_{\text{tol}}$.
initialize $\boldsymbol{d}^0 \in \mathbb{R}^m_{++}$
repeat
  $\boldsymbol{d}^1 \leftarrow \boldsymbol{p}/(A\boldsymbol{d}^0)$
  $\boldsymbol{d}^0 \leftarrow \boldsymbol{q}/(A^T\boldsymbol{d}^1)$
  $\epsilon \leftarrow$ maximal update in $(\boldsymbol{d}^0, \boldsymbol{d}^1)$ or distance between $D^1 A \boldsymbol{d}^0$ and $\boldsymbol{p}$
until $\epsilon < \epsilon_{\text{tol}}$

With a proper introduction to both problems, we are now ready to establish the equivalence between Luce choice model estimation and matrix balancing. In Section 5, we return to Sinkhorn's algorithm for the matrix-balancing problem and provide answers to open problems concerning its linear convergence for nonnegative A by leveraging the connections that we establish next.

4. Connecting Choice Modeling and Matrix Balancing

In this section, we formally establish the connections between choice modeling and matrix balancing. We show that maximizing the log-likelihood (5) is equivalent

to solving a canonical matrix-balancing problem. We also precisely describe the correspondence between the relevant conditions in the two problems. In view of this equivalence, we show that Sinkhorn's algorithm, when applied to estimate Luce choice models, is in fact a parallelized generalization of the classic iterative algorithm for choice models dating back to Zermelo (1929), Dykstra (1956), and Ford (1957) and also studied extensively by Hunter (2004) and Vojnović et al. (2020, 2023).

4.1. Maximum Likelihood Estimation of Luce Choice Models as Matrix Balancing

The optimality conditions for maximizing the log-likelihood (5) for each s_j are given by

$$\partial_{s_j}\ell(s) = \sum_{i\in[n]\,|\,(j,S_i)} \frac{1}{s_j} - \sum_{i\in[n]\,|\,j\in S_i} \frac{1}{\sum_{k\in S_i} s_k} = 0.$$

Multiplying by s_j and dividing by $1/n$, we have

$$\frac{W_j}{n} = \frac{1}{n} \sum_{i\in[n]\,|\,j\in S_i} \frac{s_j}{\sum_{k\in S_i} s_k}, \tag{8}$$

where $W_j := |\{i \in [n]\,|\,(j,S_i)\}|$ is the number of observations where j is selected.

Note that in the special case where $S_i \equiv [n]$ (i.e., every choice set contains all items), the MLE simply reduces to the familiar empirical frequencies $\hat{s}_j = W_j/n$. However, when the choice sets S_i vary, no closed-form solution to (8) exists, which is the primary motivation behind the long line of works on the algorithmic problem of solving (8). With varying S_i, we can interpret the optimality condition as requiring that the observed frequency of j being chosen (left-hand side) be equal to the average or expected probability of j being selected (right-hand side), which conditional on choice set S_i, is $\frac{s_j}{\sum_{k\in S_i} s_k}$ if $j \in S_i$ and zero otherwise. In addition, note that because the optimality condition in (8) only involves the frequency of selection, distinct data sets could yield the same optimality conditions and hence, the same MLEs. For example, suppose that two alternatives j and k both appear in choice sets S_i and $S_{i'}$, with j selected in S_i and k selected in $S_{i'}$. Then, switching the choices in S_i and $S_{i'}$ does not alter the likelihood and optimality conditions. This feature holds more generally with longer cycles of items and choice sets, and it can be viewed as a consequence of the context-independent nature of Luce's choice axiom (i.e., IIA). In some sense, it is also the underpinning of many works in economics that estimate choice models based on marginal sufficient statistics. A prominent example is Berry et al. (1995), which estimates consumer preferences using data on aggregate market shares of products.

Remark 2 (Reduction to Unique Choice Sets). In practice, the choice sets of many observations may be identical to each other (i.e., $S_i \equiv S_{i'}$ for some $i, i' \in [n]$). Because (8) only depends on the total "winning" counts of items, we may aggregate over observations with the same S_i:

$$\sum_{i\in[n]\,|\,j\in S_i} \frac{s_j}{\sum_{k\in S_i} s_k} = \sum_{i'\in[n^*]\,|\,j\in S^*_{i'}} R_{i'} \cdot \frac{s_j}{\sum_{k\in S^*_{i'}} s_k},$$

where each $S^*_{i'}$ is a unique choice set that appears in $R_{i'} \geq 1$ observations for $i' = 1, \ldots, n^* \leq n$. By construction, $\sum_{i'=1}^{n^*} R_{i'} = n$. Note, however, that the selected item could vary across different appearances of S^*_i, yet the optimality conditions only involve each item's winning count W_j. From now on, we will assume this reduction and drop the * superscript. In other words, without loss of generality, we assume that we observe n unique choice sets, and choice set S_i has multiplicity R_i (Shah et al. 2015). The resulting maximum likelihood problem has optimality conditions

$$W_j = \sum_{i\in[n]\,|\,j\in S_i} R_i \cdot \frac{s_j}{\sum_{k\in S_i} s_k}. \tag{9}$$

We are now ready to reformulate (9) as a canonical matrix-balancing problem. Define $\boldsymbol{p} \in \mathbb{R}^n$ as $\boldsymbol{p}_i = R_i$ (i.e., the number of times that choice set S_i appears in the data). Define $\boldsymbol{q} \in \mathbb{R}^m$ as $\boldsymbol{q}_j = W_j$ (i.e., the number of times that item j was selected in the data). By construction, we have $\sum_i \boldsymbol{p}_i = \sum_j \boldsymbol{q}_j$, and $\boldsymbol{p}_i, \boldsymbol{q}_j > 0$ whenever Assumption 1 holds.

Now, define the $n \times m$ binary matrix A by $A_{ij} = 1\{j \in S_i\}$ so that the ith row of A is the indicator of which items appear in the (unique) choice set S_i and the jth column of A is the indicator of which choice sets item j appears in. We refer to this A constructed from a choice data set as the participation matrix. By construction, A has distinct rows but may still have identical columns. We may also remove repeated columns by "merging" items and their win counts. Their estimated scores can be computed from the score of the merged item proportional to their respective win counts. We do not require this reduction in our results. Figure 1 provides an illustration of the matrix-balancing representation of the Luce choice-modeling problem with $(A, \boldsymbol{p}, \boldsymbol{q})$ defined as above.

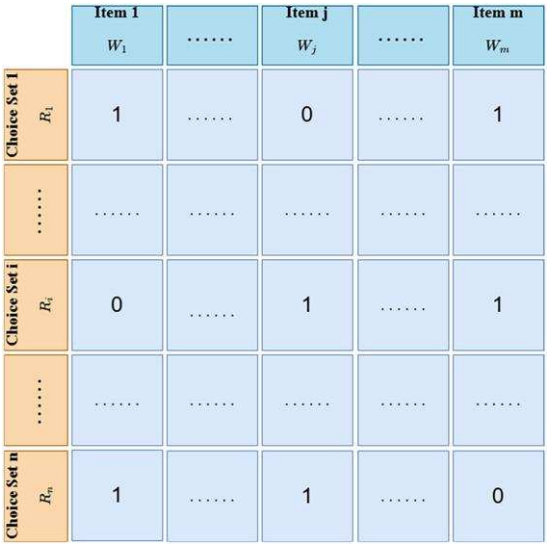
Let $D^0 \in \mathbb{R}^{m\times m}$ be the diagonal matrix with $D^0_j = s_j$ and $D^1 \in \mathbb{R}^{n\times n}$ be the diagonal matrix with $D^1_i = R_i/\sum_{k\in S_i} s_k$, and define the scaled matrix

$$\hat{A} := D^1 A D^0. \tag{10}$$

The matrices D^1 and D^0 are scalings of rows and columns of A, respectively, and

$$\hat{A}_{ij} = \frac{R_i}{\sum_{k\in S_i} s_k} \cdot 1\{j \in S_i\} \cdot s_j.$$

Figure 1. (Color online) Representation of Luce Choice Data with Participation Matrix A, R_i as the Frequency of Appearances of Choice Set i, and W_j as the Frequency of Choices of Item j as a Matrix-Balancing Problem $(A, \boldsymbol{p}, \boldsymbol{q})$ with Target Marginals $\boldsymbol{p}_i = R_i$ and $\boldsymbol{q}_j = W_j$



The key observation is that the optimality conditions (9) can be rewritten as

$$\hat{A}^T \mathbf{1}_n = \boldsymbol{q}. \tag{11}$$

Moreover, by construction, $\hat{A}$ also satisfies

$$\hat{A}\mathbf{1}_m = \boldsymbol{p}. \tag{12}$$

Therefore, if s_j's satisfy the optimality conditions for maximizing (5), then D^0, D^1 defined above solve the matrix-balancing problem in Equations (10)–(12). Moreover, the converse is also true, and we thus establish the equivalence between choice maximum likelihood estimation and matrix balancing. All omitted proofs appear in Online Appendix F.

Theorem 1 (Equivalence of Problems). *Let $(A, \boldsymbol{p}, \boldsymbol{q})$ be constructed from a choice data set as follows: $\boldsymbol{p} \in \mathbb{R}^n$ with $\boldsymbol{p}_i = R_i$, the multiplicity of choice set S_i; $\boldsymbol{q} \in \mathbb{R}^m$ with $\boldsymbol{q}_j = W_j$, the total number of times that item j is chosen in the choice data set; and $A \in \mathbb{R}^{n\times m}$ with $A_{ij} = 1\{j \in S_i\}$ (i.e., the ith row of A is a one-hot encoding of the choice set S_i).*

Then, $D^0, D^1 > 0$ with $\sum_j D^0_j = 1$ solves the matrix-balancing problem

$$\begin{aligned} D^1 A D^0 &= \hat{A} \\ \hat{A}\mathbf{1}_m &= \boldsymbol{p} \\ \hat{A}^T \mathbf{1}_n &= \boldsymbol{q} \end{aligned} \tag{13}$$

if and only if $s \in \Delta_m$ with $s_j = D^0_j$ satisfies the optimality conditions (9) of the maximum likelihood estimation problem of a Luce choice model given the choice data set.

Theorem 1 implies that (5) has a unique maximizer s in the interior of the probability simplex if and only if (13) has a unique normalized solution D^0 as well. The next question, naturally, is then how Assumption 1 and Assumption 2 for choice modeling are connected to Assumption 3 and Assumption 4 for matrix balancing.

Proposition 1 (Equivalence of Assumptions). *Let $(A, \boldsymbol{p}, \boldsymbol{q})$ be constructed from the choice data set as in Theorem 1, with $\boldsymbol{p}, \boldsymbol{q} > 0$. Assumption 2 is equivalent to Assumption 4. Furthermore, Assumption 1 holds if and only if $(A, \boldsymbol{p}, \boldsymbol{q})$ satisfy Assumption 3 and A satisfies Assumption 4.*

Thus, when the choice maximum likelihood estimation problem is cast as a matrix-balancing problem, Assumption 3 exactly characterizes the gap between Assumption 2 and Assumption 1. We provide some intuition for Proposition 1. When we construct a triplet $(A, \boldsymbol{p}, \boldsymbol{q})$ from a choice data set, with $\boldsymbol{p}$ the numbers of appearances of unique choice sets and $\boldsymbol{q}$ the winning counts of each item, Assumption 4 precludes the possibility of partitioning the items into two subsets that never get compared with each other (i.e., Assumption 2). Then, Assumption 3 requires that whenever a strict subset $M \subsetneq [m]$ of objects only appears in a strict subset $N \subsetneq [n]$ of the observations, their total winning counts are strictly smaller than the total number of these observations (i.e., there is some object $j \notin M$ that is selected in S_i for some $i \in N$, which is required by Assumption 1).

Interestingly, although Assumption 1 requires the directed comparison graph, defined by the $m \times m$ matrix of counts of item j being chosen over item k, to be strongly connected, the corresponding conditions for the equivalent matrix-balancing problem concern the $n \times m$ participation matrix A and positive vectors $\boldsymbol{p}, \boldsymbol{q}$, which do not explicitly encode the specific choice of each observation. This apparent discrepancy is because of the fact that $(A, \boldsymbol{p}, \boldsymbol{q})$ form the sufficient statistics of the Luce choice model. In other words, there can be more than one choice data set with the same optimality condition (9) and $(A, \boldsymbol{p}, \boldsymbol{q})$ defining the equivalent matrix-balancing problem (see Figure 1).

Remark 3 (Aggregate Data as Sufficient Statistics). The feature where "marginal" or aggregate quantities constitute the sufficient statistics of a parametric model is an important characteristic that underlies many works in economics and statistics (Stone 1962, Birch 1963, Good 1963, Theil 1967, Fienberg 1970, Berry et al. 1995, Kullback 1997, Fofana et al. 2002, Maystre and Grossglauser 2017, Chang et al. 2024). It makes the task of estimating a joint model from marginal quantities feasible and is very useful because in many applications, only marginal data are available because of high sampling cost or privacy-preserving considerations.

Having formulated a particular matrix-balancing problem from the estimation problem given choice

data, we may ask how one can go in the other direction. In other words, when/how can we construct a "choice data set" whose sufficient statistics are a given triplet $(A, \boldsymbol{p}, \boldsymbol{q})$? First off, for $(A, \boldsymbol{p}, \boldsymbol{q})$ to be valid sufficient statistics of a Luce choice model, $\boldsymbol{p}, \boldsymbol{q}$ need to be positive integers. Moreover, A must be a binary matrix with unique rows, each containing at least two nonzero elements (valid choice sets have at least two items). Given such an $(A, \boldsymbol{p}, \boldsymbol{q})$ satisfying Assumptions 3 and 4, a choice data set can be constructed efficiently. Such a procedure is described, for example, in Kumar et al. (2015), where A is motivated by random walks on a graph instead of matrix balancing (Online Appendix B). Their construction relies on finding the max flow on the bipartite graph G_b. For rational $\boldsymbol{p}, \boldsymbol{q}$, this max flow can be found efficiently in polynomial time (Balakrishnan et al. 2004, Idel 2016). Moreover, the maximum flow implies a matrix A' satisfying Assumption 3(a), thus providing a feasibility certificate for the matrix-balancing problem as well.

We have thus closed the loop and fully established the equivalence of the maximum likelihood estimation of Luce choice models and the canonical matrix-balancing problem.

Corollary 1. *There is a one-to-one correspondence between classes of maximum likelihood estimation problems with the same optimality conditions* (9) *and canonical matrix-balancing problems with* $(A, \boldsymbol{p}, \boldsymbol{q})$, *where* A *is a valid binary participation matrix and* $\boldsymbol{p}, \boldsymbol{q} > 0$ *have integer entries.*

Connections to discrete choice modeling have also been established for the related problem of regularized semidiscrete optimal transport (Taşkesen et al. 2023), although the problem setting and results are distinct from the ones studied in this paper. We next turn our attention to the algorithmic connections between choice modeling and matrix balancing.

4.2. Algorithmic Connections Between Matrix Balancing and Choice Modeling

Given the equivalence between matrix balancing and choice modeling, we can naturally consider applying Sinkhorn's algorithm to maximize (5). In this case, one can verify that the updates in each full iteration of Algorithm 1 reduce algebraically to

$$s_j^{(t+1)} = W_j \Big/ \sum_{i \in [n] \mid j \in S_i} \frac{R_i}{\sum_{k \in S_i} s_k^{(t)}} \tag{14}$$

in the tth iteration. Comparing (14) with the optimality condition in (9), which we recall is given by

$$W_j = \sum_{i \in [n] \mid j \in S_i} R_i \frac{s_j}{\sum_{k \in S_i} s_k} = s_j \cdot \sum_{i \in [n] \mid j \in S_i} \frac{R_i}{\sum_{k \in S_i} s_k},$$

we can, therefore, interpret the iterations as simply dividing the winning count W_j by the coefficient of s_j on the right repeatedly in the hope of converging to a fixed point. A similar intuition was given by Ford (1957) in the special case of pair-wise comparisons. Indeed, the algorithm proposed in that paper is a cyclic variant of (14) applied to pair-wise comparisons. However, this connection is mainly algebraic as the optimality condition in Ford (1957) does not admit a reformulation as the matrix-balancing problem in (13).

In Online Appendix B, we provide further discussions on the connections of Sinkhorn's algorithm to existing frameworks and algorithms in the choice-modeling literature and the optimization literature. We demonstrate that many existing algorithms for Luce choice model estimation are in fact special cases or analogs of Sinkhorn's algorithm. These connections also illustrate the many interpretations of Sinkhorn's algorithm (e.g., as a distributed optimization algorithm as well as a minorization-maximization or MM algorithm) (Lange 2016). However, compared with most algorithms for choice modeling discussed in this work, Sinkhorn's algorithm is more general as it applies to nonbinary A and noninteger $\boldsymbol{p}, \boldsymbol{q}$, and it has the additional advantage of being parallelized and distributed and hence, more efficient in practice. To highlight the ubiquity of Sinkhorn's algorithm in the choice setting, we summarize these algorithmic connections below.

Theorem 2 (Equivalence of Algorithms). *Sinkhorn's algorithm, when applied to matrix-balancing formulations of various choice-modeling problems, is equivalent to the following algorithms*:

- *the iterative algorithms of Zermelo* (1929), *Dykstra* (1956), *and Ford* (1957) *for the BTL model of pair-wise comparison data*;
- *the MM algorithm of Hunter* (2004) *for the Plackett–Luce model of ranking data*;
- *the unregularized version of the ChoiceRank algorithm of Maystre and Grossglauser* (2017) *for their proposed network choice model*; *and*
- *the BLP algorithm of Berry et al.* (1995) *in a logit random utility model with only intercepts*.

Moreover, Sinkhorn's algorithm can be viewed as a general MM algorithm as well as a message-passing algorithm, and in the latter case, it is a variant of the accelerated spectral-ranking algorithm of Agarwal et al. (2018) *for Luce choice models based on a different moment condition*.

The mathematical and algorithmic connections between matrix balancing and choice modeling that we establish in this paper allow for the transmission of ideas in both directions. For example, inspired by regularized maximum likelihood estimation (Maystre and Grossglauser 2017), we propose a regularized version of Sinkhorn's algorithm in Online Appendix C, which is guaranteed to converge even when the original Sinkhorn's algorithm does not converge. Leveraging optimization connections between maximum likelihood

estimation of choice models and matrix balancing, Chang et al. (2024) propose a statistical framework for network traffic that justifies the popular use of Sinkhorn's algorithm to infer detailed dynamic networks from aggregate node-level activities. In the rest of the main text, we focus on resolving some interesting open problems on the convergence of Sinkhorn's algorithm motivated by results in choice modeling on the importance of algebraic connectivity in quantifying statistical and computational efficiencies.

5. Linear Convergence of Sinkhorn's Algorithm for Nonnegative Matrices

In this section, we turn our attention on matrix balancing and study the global and asymptotic linear convergence of Sinkhorn's algorithm for general nonnegative matrices $A \geq 0$ and positive marginals $\boldsymbol{p}, \boldsymbol{q} > 0$. We first present the relevant optimization principles behind matrix balancing and discuss some existing results in Section 5.1, and then, we present the convergence results in Sections 5.2–5.4. Throughout, we use superscript (t) to denote quantities after t iterations of Sinkhorn's algorithm with normalized columns, described in Algorithm 1.

5.1. Preliminaries

We start with the optimization principles associated with matrix balancing and Sinkhorn's algorithm. Given a matrix-balancing problem with $A \geq 0, \sum_{ij} A_{ij} = 1$ and target marginals $\boldsymbol{p}, \boldsymbol{q}$ with $\sum_i \boldsymbol{p}_i = \sum_j \boldsymbol{q}_j = 1$, consider the following KL divergence (relative entropy) minimization problem:

$$\begin{aligned} &\min_{\hat{A} \in \mathbb{R}_+^{n \times m}} D_{\mathrm{KL}}(\hat{A} \| A) \\ &\hat{A}\mathbf{1}_m = \boldsymbol{p} \\ &\hat{A}^T \mathbf{1}_n = \boldsymbol{q}. \end{aligned} \tag{15}$$

It is well known that when scalings D^0, D^1 solve the matrix-balancing problem with $(A, \boldsymbol{p}, \boldsymbol{q})$, the scaled matrix $\hat{A} = D^1 A D^0$ is the unique minimizer of (15) (Bregman 1967a, Ireland and Kullback 1968). Moreover, vector representations $\boldsymbol{d}^0, \boldsymbol{d}^1$ of the optimal scalings D^0, D^1 precisely minimize the following (negative) dual objective of (15):

$$g(\boldsymbol{d}^0, \boldsymbol{d}^1) := (\boldsymbol{d}^1)^T A \boldsymbol{d}^0 - \sum_{i=1}^n \boldsymbol{p}_i \log \boldsymbol{d}_i^1 - \sum_{j=1}^m \boldsymbol{q}_j \log \boldsymbol{d}_j^0, \tag{16}$$

and Sinkhorn's algorithm is a block coordinate descent-type algorithm (Tseng 2001) applied to minimize (16). Luo and Tseng (1992) study the linear convergence of block coordinate descent algorithms for a general class of objectives that include (16). In particular, their result implies that the convergence of Sinkhorn's algorithm, measured in terms of the optimality gap of g, is linear with some unknown rate as long as finite positive scalings D^0, D^1 exist that satisfy (2). The function g, sometimes referred to as the potential function of Sinkhorn's algorithm, also turns out to be crucial in quantifying the global linear convergence rate in the present work.

Remark 4 (Optimization Connections). Interestingly, minimizing (16) is in fact equivalent to maximizing the log-likelihood function $\ell(s)$ in (5) for valid $(A, \boldsymbol{p}, \boldsymbol{q})$ because $\min_{d^1} g(\boldsymbol{d}^0, \boldsymbol{d}^1) = -\ell(\boldsymbol{d}^0) + c$ for some $c > 0$. Moreover, the optimality condition of minimizing g with respect to $\boldsymbol{d}^0$ reduces to the optimality condition (9). A detailed discussion can be found in Online Appendix B.4. This connection relates choice modeling and matrix balancing from an optimization perspective.

Although convergence results on Sinkhorn's algorithm are abundant, they are often stated in different forms, are developed under different assumptions, and apply to settings of varying degrees of generality. Next, we briefly discuss some existing works to clarify their connections and distinctions, which also help motivate the technical results in this paper. They are summarized in Table 1.

First and foremost, how to define and measure the convergence of Sinkhorn's algorithm is not entirely trivial. Because of the indeterminacy of scalings under the transformation $(D^0, D^1) \to (D^0/c, c \cdot D^1)$, most works define convergence using quantities that are invariant under this transformation. Let $D^{0(t)}, D^{1(t)}$ be the scalings obtained after t iterations of Sinkhorn's algorithm, and let $A^{(t)} := D^{1(t)} A D^{0(t)}$ be the scaled matrix based on these scalings. Because $A^{(t)}$ is invariant under the transformation, some earlier works, such as Franklin and Lorenz (1989) and Soules (1991), measure convergence in terms of $A^{(t)}$ to the optimally scaled matrix $\hat{A} = D^1 A D^0$ that has the target row and column sums $\boldsymbol{p}$, $\boldsymbol{q}$. Most later works focus instead on the convergence of the marginal quantities

$$\boldsymbol{r}^{(t)} := A^{(t)} \mathbf{1}_m; \quad \boldsymbol{c}^{(t)} := A^{(t)T} \mathbf{1}_n$$

to the target row and column sums $\boldsymbol{p}$, $\boldsymbol{q}$. Because after each iteration in Algorithm 1, the column constraint is always satisfied $(A^{(t)} \mathbf{1}_n = \boldsymbol{q})$, it suffices to focus on the convergence of $\boldsymbol{r}^{(t)}$ to $\boldsymbol{p}$. For example, Léger (2021) uses the KL divergence $D_{\mathrm{KL}}(\boldsymbol{r}^{(t)} \| \boldsymbol{p})$, whereas Altschuler et al. (2017) and Chakrabarty and Khanna (2021) use the ℓ^1 distance $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1$, which is upper bounded by the KL divergence via Pinsker's inequality. Given the entropy optimization perspective of matrix balancing and Sinkhorn's algorithm in (15) and (16), it is also possible to measure convergence in terms of the dual optimality gap $g(\boldsymbol{d}^{0(t)}, \boldsymbol{d}^{1(t)}) - g(\boldsymbol{d}^0, \boldsymbol{d}^1)$, which in turn, bounds KL

divergences via

$$g(\boldsymbol{d}^{0(t)}, \boldsymbol{d}^{1(t)}) - g(\boldsymbol{d}^0, \boldsymbol{d}^1) = D_{\mathrm{KL}}(D^1 A D^0 \| A^{(t)}) = \sum_{s=t}^{\infty} D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{r}^{(s)}) + D_{\mathrm{KL}}(\boldsymbol{q}\|\boldsymbol{c}^{(s)}). \tag{17}$$

Luo and Tseng (1992) show that this dual optimality gap converges linearly, but how the convergence rate depends on problem structure has been an open question. Our global linear convergence result in Theorem 3 is the first to characterize this rate.

Next, we note the distinction between global and local (particularly asymptotic) convergence results. Global results hold for all iterations $t > 0$. For example, Léger (2021) shows $D_{\mathrm{KL}}(\boldsymbol{r}^{(t)}\|\boldsymbol{p}) \le D^*_{\mathrm{KL}}/t$ for all t, where D^*_{KL} is the optimal value of the relative entropy minimization Problem (15). On the other hand, local convergence results pertain to the behavior of an algorithm in a neighborhood of the optimal solution, whereas asymptotic results only hold in the limit as $t \to \infty$. For example, Knight (2008) characterizes the asymptotic rate $\lim_{t\to\infty} \|\boldsymbol{r}^{(t+1)} - \boldsymbol{p}\|_*/\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_*$ of linear convergence for Sinkhorn's algorithm, where $\|\cdot\|_*$ is an implicitly defined norm. When $A \ge 0$, Knight (2008) is the only work on the exact asymptotic convergence rate for square A and uniform $\boldsymbol{p}, \boldsymbol{q}$. Theorem 4 is the first asymptotic result for general $A \ge 0$ and nonuniform $\boldsymbol{p}, \boldsymbol{q}$, with an explicit norm ($\|\cdot\|_2$). Although asymptotic results provide a more precise description of an algorithm's behavior near the optimal solution, global results are useful for obtaining complexity bounds on the number of iterations required to obtain ε accuracy solutions. In fact, global results are often stated directly as complexity bounds. For example, the result in Altschuler et al. (2017) is that for $A > 0$, $t = 4(\varepsilon)^{-2}\log(\sum_{ij} A_{ij}/\min_{ij} A_{ij})$ iterations of Sinkhorn's algorithm guarantee $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1 \le \varepsilon$.

Lastly, we note the distinction between (global) linear and sublinear convergence results. Linear convergence is often understood as successive improvements of the convergence metric by a constant factor. For example, Franklin and Lorenz (1989) show that for $A > 0$ and the Hilbert metric $\boldsymbol{d}$, $d(\boldsymbol{r}^{(t+1)}, \boldsymbol{p}) \le \lambda \cdot d(\boldsymbol{r}^{(t)}, \boldsymbol{p})$ for all $t > 0$ for some $\lambda \in (0,1)$. As a result, $d(\boldsymbol{r}^{(t)}, \boldsymbol{p}) \le \lambda^t \cdot d(\boldsymbol{r}^{(0)}, \boldsymbol{p})$ decreases exponentially in t so that $d(\boldsymbol{r}^{(t)}, \boldsymbol{p}) \le \varepsilon$ in $\mathcal{O}(\log(1/\varepsilon))$ iterations. In contrast, in sublinear results, such as Léger (2021), the convergence metric $D_{\mathrm{KL}}(\boldsymbol{r}^{(t)}\|\boldsymbol{p})$ only decreases polynomially in t, requiring $\mathcal{O}(1/\varepsilon)$ iterations to guarantee $D_{\mathrm{KL}}(\boldsymbol{r}^{(t)}\|\boldsymbol{p}) \le \varepsilon$. Although sublinear complexity bounds have worse (polynomial) dependence on $1/\varepsilon$, they often focus on optimizing the dependence on problem size and dimension. Our main focus in this paper is on understanding the linear convergence behavior of Sinkhorn's algorithm when $A \ge 0$ (i.e., $\mathcal{O}(\log(1/\varepsilon))$ iteration complexity). Nevertheless, we also provide refined complexity bounds in Proposition 2 that optimize dependence on problem constants.

As discussed before, when $A \ge 0$, only sublinear convergence results with explicit rates are known (Kalantari et al. 2008, Chakrabarty and Khanna 2021, Léger 2021), whereas Luo and Tseng (1992) implies global linear convergence with an unknown rate. We now characterize this global linear rate of convergence in terms of the algebraic connectivity of the bipartite graph defined in (7).

5.2. Global Linear Convergence

Our analysis starts with the following change of variables to transform the potential function (16):

$$\boldsymbol{u} := \log \boldsymbol{d}^0, \quad \boldsymbol{v} := -\log \boldsymbol{d}^1, \tag{18}$$

resulting in the reparameterized potential function $g(u, v)$ of (16):

$$g(\boldsymbol{u}, \boldsymbol{v}) := \sum_{ij} A_{ij} e^{-v_i + u_j} + \sum_{i=1}^{n} \boldsymbol{p}_i v_i - \sum_{j=1}^{m} \boldsymbol{q}_j u_j. \tag{19}$$

Note first that $g(\boldsymbol{u}, \boldsymbol{v}) = g(\boldsymbol{u} + a, \boldsymbol{v} + a)$ for any constant $a \in \mathbb{R}$. We can verify that Sinkhorn's algorithm is equivalent to the alternating minimization algorithm (Bertsekas 1997) for (19), which alternates between minimizing with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$, holding the other block fixed:

$$\boldsymbol{u}^{(t)} \leftarrow \arg\min_{u} g(\boldsymbol{u}, \boldsymbol{v}^{(t-1)}), \quad \boldsymbol{v}^{(t)} \leftarrow \arg\min_{v} g(\boldsymbol{u}^{(t)}, \boldsymbol{v}) \tag{20}$$

or written more explicitly element wise,

$$\boldsymbol{u}_j^{(t)} \leftarrow \log \frac{\boldsymbol{q}_j}{\sum_i A_{ij} e^{-v_i^{(t-1)}}}, \quad \boldsymbol{v}_i^{(t)} \leftarrow \log \frac{\boldsymbol{p}_i}{\sum_j A_{ij} e^{\boldsymbol{u}_j^{(t)}}}. \tag{21}$$

A main reason to focus on (19) instead of the log-barrier form (16) is that (19) has a Hessian with desirable properties for proving linear convergence. The Hessian of $g(\boldsymbol{u}, \boldsymbol{v})$ is

$$\nabla^2 g(\boldsymbol{u}, \boldsymbol{v}) = \begin{bmatrix} \mathcal{D}(\hat{A}\mathbf{1}_m) & -\hat{A} \\ -\hat{A}^T & \mathcal{D}(\hat{A}^T \mathbf{1}_n) \end{bmatrix}, \tag{22}$$

where $\mathcal{D}$ converts a vector to a diagonal matrix and $\hat{A} = \mathcal{D}(\boldsymbol{d}^1) A \mathcal{D}(\boldsymbol{d}^0) = \mathcal{D}(\exp(-\boldsymbol{v})) A \mathcal{D}(\exp(\boldsymbol{u}))$ is the matrix scaled by $\boldsymbol{u}$, $\boldsymbol{v}$. Note that the Hessian $\nabla^2 g(\boldsymbol{u}, \boldsymbol{v})$ always has $\mathbf{1}_{m+n}$ in its null space. On the surface, it may seem that standard linear convergence results for first-order methods, which require strong convexity (or the Polyak–Łojasiewicz condition) of the objective function, do not apply to $g(\boldsymbol{u}, \boldsymbol{v})$. However, we will show that whenever the matrix-balancing problem has finite scaling solutions, $g(\boldsymbol{u}, \boldsymbol{v})$ is in fact strongly convex when restricted to bounded subsets of the subspace

$$\mathbf{1}^{\perp}_{m+n} := \{\boldsymbol{u} \in \mathbb{R}^m, \boldsymbol{v} \in \mathbb{R}^n : (\boldsymbol{u}, \boldsymbol{v})^T \mathbf{1}_{m+n} = 0\}. \tag{23}$$

Moreover, the invariance of $g(\boldsymbol{u}, \boldsymbol{v})$ and its gradient and Hessian under constant translations of $(\boldsymbol{u}, \boldsymbol{v})$ by $\mathbf{1}_{m+n}$

guarantees that the strong convexity constant of $g(\boldsymbol{u},\boldsymbol{v})$ on $\mathbf{1}_{m+n}^{\perp}$ in fact quantifies the linear convergence of Sinkhorn's algorithm, even if the iterates $\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)}$ are not in $\mathbf{1}_{m+n}$. Similar types of "restricted strong convexity" properties have been studied by, for example, Agarwal et al. (2010). It also shares similarities with the exp-concavity property popular in online learning (Hazan 2016, Orabona 2019), which implies the strong convexity of a function in the direction of the gradient evaluated at any point. In contrast, $g(\boldsymbol{u},\boldsymbol{v})$ is strongly convex along any direction orthogonal to $\mathbf{1}_{m+n}$, but its gradient evaluated at any $(\boldsymbol{u},\boldsymbol{v})$ is not necessarily orthogonal to $\mathbf{1}_{m+n}$. However, the key is that along the trajectory of iterates $(\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)})$ obtained by running Sinkhorn's algorithm, the gradients of g evaluated at $(\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)})$ are indeed orthogonal to $\mathbf{1}_{m+n}$, which is sufficient to guarantee the linear convergence of Sinkhorn's algorithm.

We now introduce the key quantities and definitions used in our result. Let Sinkhorn's algorithm initialize with a $\boldsymbol{u}^{(0)}$, with $\boldsymbol{v}^{(0)}$ given by (21). Define the constant B as

$$\begin{aligned} B := \sup_{(\boldsymbol{u},\boldsymbol{v})} & \|(\boldsymbol{u},\boldsymbol{v})\|_\infty \\ \text{subject to } & (\boldsymbol{u},\boldsymbol{v})^T\mathbf{1}_{m+n}=0, \\ & g(\boldsymbol{u},\boldsymbol{v}) \le g(\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)}). \end{aligned} \tag{24}$$

In other words, B is the diameter of the initial normalized sublevel set. We will show that B is finite and that it bounds normalized Sinkhorn iterates $\|(\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)})\|_\infty$ because under Assumption 3, the function $g(\boldsymbol{u},\boldsymbol{v})$ is coercive on the subspace $\mathbf{1}_{m+n}^{\perp}$. Coercivity is an important property, and we define it below following Bertsekas (2016).

Definition 1 (Coercivity). A function $f(x):\mathbb{R}^d\to\mathbb{R}$ is coercive on a subspace $V\subseteq\mathbb{R}^d$ if

$$f(x^{(t)})\to+\infty \quad \text{whenever } x^{(t)}\in V \text{ and } \|x^{(t)}\|_\infty\to+\infty. \tag{25}$$

Next, define the Laplacian matrix $\mathcal{L}:=\mathcal{L}(A)$ of the bipartite graph G_b (see (7)) as

$$\mathcal{L} := \begin{bmatrix} \mathcal{D}(A\mathbf{1}_m) & -A \\ -A^T & \mathcal{D}(A^T\mathbf{1}_n) \end{bmatrix}, \tag{26}$$

and refer to the second-smallest eigenvalue $\lambda_{-2}(\mathcal{L})$ as the Fiedler eigenvalue. $\lambda_{-2}(\mathcal{L})>0$ if and only if Assumption 4 holds and it quantifies "connectivity" of the data structure (Spielman 2007). Although an important quantity in the choice-modeling literature, algebraic connectivity has not been previously used in the analysis of Sinkhorn's algorithm. For details on the graph Laplacian and the Fiedler eigenvalue, see Online Appendix A. Finally, define the smoothness parameters

$$l_0 := \max_j (A^T\mathbf{1}_n)_j, \quad l_1 := \max_i (A\mathbf{1}_m)_i, \tag{27}$$

which are used to quantify the smoothness of $g(\boldsymbol{u},\boldsymbol{v})$.

We can now state one of our main contributions to the study of Sinkhorn's algorithm.

Theorem 3 (Global Linear Convergence). *Suppose Assumption 3 and Assumption 4 hold. Let $\mathcal{L}$ be the bipartite graph Laplacian defined in (26) and $\lambda_{-2}(\mathcal{L})$ be its second-smallest eigenvalue. Let l_0,l_1 be the smoothness parameters defined in (27). Let $(\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)})$ be Sinkhorn iterates at iteration t defined in (21) and B be the bound on $\|(\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)})\|_\infty$ defined in (24). Define $g^*:=\inf_{\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{u},\boldsymbol{v})$. For all $t>0$, the optimality gap of the dual objective $g(\boldsymbol{u},\boldsymbol{v})$ defined in (19) satisfies*

$$g(\boldsymbol{u}^{(t+1)},\boldsymbol{v}^{(t+1)})-g^* \le \left(1-e^{-4B}\frac{\lambda_{-2}(\mathcal{L})}{\min\{l_0,l_1\}}\right)(g(\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)})-g^*). \tag{28}$$

The ratio $\min\{l_0,l_1\}/\lambda_{-2}(\mathcal{L})$ can be interpreted as a condition number of $\mathcal{L}$.

The linear convergence rate of Sinkhorn's algorithm is, therefore, quantified by $\lambda_{-2}(\mathcal{L})/\min\{l_0,l_1\}$, which is invariant under rescalings of $A\to c\cdot A$. Although the corresponding bipartite graph G_b with biadjacency matrix A is a natural object to consider in the study matrix-balancing problems, to our knowledge, Theorem 3 is the first to highlight the precise role of its spectral property, described by $\lambda_{-2}(\mathcal{L})$, in the linear convergence of Sinkhorn's algorithm. It fills the gap left by Luo and Tseng (1992), who establish linear convergence with an implicit rate, and it allows us to compute its dependence on problem parameters by applying established bounds on $\lambda_{-2}(\mathcal{L})$ from spectral graph theory (Spielman 2007).

Remark 5 (Importance of Assumptions for Linear Convergence). The importance of Assumptions 3 and 4 is clearly reflected in the bound (28). First, note that the Fiedler eigenvalue $\lambda_{-2}(\mathcal{L})>0$ if and only if Assumption 4 holds (see Online Appendix A). On the other hand, Assumption 3 guarantees the coercivity of g on $\mathbf{1}_{m+n}^{\perp}$ (see (25)). This property ensures that B defined in (24) satisfies $B<\infty$ and consequently, that normalized iterates stay bounded by B. That Assumption 3 guarantees $g(\boldsymbol{u},\boldsymbol{v})$ is coercive should be compared with the observation by Hunter (2004) that Assumption 1 guarantees the upper compactness (a closely related concept) of the log-likelihood function (5). When Assumption 3 fails, B may become infinite, and $\|(\boldsymbol{u}^{(t)},\boldsymbol{v}^{(t)})\|_\infty\to\infty$.

Remark 6 (Self-Normalizing Property of Sinkhorn). The ability of Sinkhorn's algorithm to exploit the (subspace) strong convexity of $g(\boldsymbol{u},\boldsymbol{v})$ on $\mathbf{1}_{m+n}^{\perp}$ to achieve linear convergence relies critically on the invariance of

the scaled matrix $\hat{A} = D^1 A D^0$ and $g(\boldsymbol{u}, \boldsymbol{v})$ under the transformation $(D^0, D^1) \to (D^0/c, c \cdot D^1)$. This is an intrinsic feature of the matrix-balancing problem that has been well known but not fully exploited in the convergence analysis so far. It guarantees that the translation $(\boldsymbol{u}, \boldsymbol{v}) \to (\boldsymbol{u} - \log c, \boldsymbol{v} - \log c)$ does not alter $g(\boldsymbol{u}, \boldsymbol{v})$ and its derivatives in (19). We can, therefore, impose the auxiliary normalization $(\boldsymbol{u}, \boldsymbol{v})^T \mathbf{1}_{m+n} = 0$ or equivalently, $\prod_j \boldsymbol{d}_j^0 = \prod_i \boldsymbol{d}_i^1$, which is easily achieved by requiring that after every update in Algorithm 1, a normalization $(\boldsymbol{d}^0/c, c\boldsymbol{d}^1)$ is performed using the normalizing constant

$$c = \sqrt{\prod_j \boldsymbol{d}_j^0 / \prod_i \boldsymbol{d}_i^1}. \tag{29}$$

See Algorithm 3 for the normalized Sinkhorn's algorithm, which given (29), results in a virtual sequence of $\boldsymbol{u}^{(t)}, \boldsymbol{v}^{(t)}$ satisfying $(\boldsymbol{u}^{(t)}, \boldsymbol{v}^{(t)})^T \mathbf{1}_{m+n} = 0$. Moreover, the values of $g(\boldsymbol{u}, \boldsymbol{v})$ on this virtual sequence are identical to those on the standard Sinkhorn iterates. As a result, the convergence result (28) applies to the standard Sinkhorn's algorithm without normalization (or with any other normalization) because of the invariance of $g(\boldsymbol{u}, \boldsymbol{v})$. Normalization of Sinkhorn's algorithm is also considered in the analyses in Carlier et al. (2023), although they use the asymmetric condition $\boldsymbol{u}_0 = 0$, which does not guarantee that normalized Sinkhorn iterates stay in $\mathbf{1}_{m+n}^{\perp}$.

With this auxiliary normalization procedure, the proof of Theorem 3 then relies on the observation that the Hessian of $g(\boldsymbol{u}, \boldsymbol{v})$ is precisely the graph Laplacian $\mathcal{L}(\boldsymbol{u}, \boldsymbol{v})$ of the bipartite graph with biadjacency matrix $\hat{A} = \mathcal{D}(\exp(-\boldsymbol{v}))A\mathcal{D}(\exp(\boldsymbol{u}))$. As $(\boldsymbol{u}, \boldsymbol{v})$ are bounded on normalized Sinkhorn iterates thanks to the coercivity of g, the Fiedler eigenvalue of $\mathcal{L} = \mathcal{L}(0, 0)$ quantifies the strong convexity on $\mathbf{1}_{m+n}^{\perp}$. Linear convergence then follows from results on block coordinate descent and alternating minimization methods for strongly convex and smooth functions (Beck and Tetruashvili 2013). Typically, the leading eigenvalue of the Hessian quantifies the smoothness, which is bounded by $2\max\{l_0, l_1\}$ for $\mathcal{L}$. For alternating minimization methods, the better smoothness constant $\min\{l_0, l_1\}$ is available. Thus, the quantity $\min\{l_0, l_1\}/\lambda_{-2}(\mathcal{L})$ in (28) can be interpreted as a type of "condition number" of the graph Laplacian $\mathcal{L}$. When A is positive (not just nonnegative), then the strong existence and uniqueness conditions are trivially satisfied, and our results continue to hold with the rate quantified by $\min\{l_0, l_1\}/\lambda_{-2}(\mathcal{L})$. In this case, both $\min\{l_0, l_1\}$ and $\lambda_{-2}(\mathcal{L})$ are $\Theta(n)$, where n is problem dimension, so $\min\{l_0, l_1\}/\lambda_{-2}(\mathcal{L})$ does not increase with problem dimension.

Remark 7 (Significance of Theorem 3). A main innovation in our paper is in introducing the concept of algebraic connectivity when quantifying the global convergence of Sinkhorn's algorithm for nonnegative matrices. In this respect, the significance of Theorem 3 is more conceptual than technical because once we identify the right quantity (algebraic connectivity) and utilize the self-normalizing property, the convergence result can be obtained using standard matrix analysis and applying the theory of Beck and Tetruashvili (2013) for block coordinate descent algorithms. Nevertheless, we feel that the role of algebraic connectivity in the study of matrix-balancing problems holds general significance and likely will lead to more results in related areas. See, for example, Chang et al. (2024), which highlights its importance for the statistical efficiency of a network traffic model based on matrix balancing.

Although Theorem 3 implies an $\mathcal{O}(\log(1/\varepsilon))$ iteration complexity, the complexity bound's dependence on problem parameters can be further improved. In particular, the constant B, which bounds $\|(\boldsymbol{u}^{(t)}, \boldsymbol{v}^{(t)})\|_\infty$, can be hard to compute for some problems. We next establish an iteration complexity bound that does not depend exponentially on the implicit constant B defined in (24).

Proposition 2 (Iteration Complexity). *Under Assumption 3 and Assumption 4, let $\boldsymbol{d}_*^0, \boldsymbol{d}_*^1$ be a pair of optimal scalings. Define $\|v\|_{-\infty} := \min_i |v_i|$, and let $C =: \max\left\{\frac{\|\boldsymbol{d}_*^0\|_\infty}{\|\boldsymbol{d}_*^0\|_{-\infty}}, \frac{1}{\|\boldsymbol{d}_*^0\|_{-\infty}\|\boldsymbol{d}_*^1\|_{-\infty}}, \|\boldsymbol{d}_*^0\|_\infty \|\boldsymbol{d}_*^1\|_\infty\right\}$. Suppose Sinkhorn's algorithm initializes with $\boldsymbol{u}^{(0)} = \mathbf{1}_m$. Then, $\|(\boldsymbol{e}^{u^{(t)}}, \boldsymbol{e}^{v^{(t)}})\|_\infty \le C$ for all $t > 0$. Moreover, for any $\varepsilon \le \frac{1}{2}\min\{\|\boldsymbol{p}\|_{-\infty}, \|\boldsymbol{q}\|_{-\infty}\}$, after*

$$\mathcal{O}\left(C^2 \cdot \frac{\min\{\|\boldsymbol{p}\|_\infty, \|\boldsymbol{q}\|_\infty\}}{\lambda_{-2}(\mathcal{L})} \cdot (\log(1/\varepsilon) + \log\log C)\right) \tag{30}$$

iterations of Sinkhorn's algorithm, the optimality gap and the ℓ^1 distance $\|\boldsymbol{r}^{(t)} - \boldsymbol{p}\|_1 \le \varepsilon$.

Proposition 2 relies on two technical innovations. First, we bound $e^B \le C$, where C is explicitly constructed from any optimal solution pair and is invariant under rescalings. Second, we improve the dependence from e^{4B} to e^{2B} using the target marginals $\boldsymbol{p}, \boldsymbol{q}$ to quantify the smoothness of $g(\boldsymbol{u}, \boldsymbol{v})$ instead. Our message here is that the convergence behavior of Sinkhorn's algorithm has two phases. Initially, we can apply a sublinear complexity bound with $\mathcal{O}(1)$ iterations to obtain Sinkhorn iterates sufficiently close to the optimal solution. Afterwards, the convergence can better be captured by a linear convergence with rate depending on the optimal solution and target marginals $\boldsymbol{p}, \boldsymbol{q}$. The dependence of C on problem dimension is problem specific. In the worst case, it can be exponential (Kalantari and Khachiyan 1993). In Online Appendix E.2, we plot (30) as a function of problem dimension on randomly generated data, and we find that the dependence is quadratic. In contrast, sublinear bounds, such as Altschuler et al. (2017) and Chakrabarty and Khanna (2021), have logarithmic

dependence on problem dimension. It remains an interesting question to improve the dependence on problem dimension in (30) and to study trade-offs with the dependence on ε.

5.3. Strong vs. Weak Convergence of Sinkhorn's Algorithm

We now discuss the two different convergence regimes of Sinkhorn's algorithm when $A \geq 0$. As mentioned in Sections 3.2 and 5.1, when $A \geq 0$, the canonical matrix-balancing problem with target marginals $\boldsymbol{p}, \boldsymbol{q}$ has a finite positive solution pair D^0, D^1 if and only if Assumption 3 holds (which trivially holds when $A > 0$). In this case, Sinkhorn's algorithm converges to $D^1 A D^0$, which also solves the KL minimization Problem (15). We call this case the strong convergence of Sinkhorn.

However, even if Assumption 3 fails and no positive finite scalings D^0, D^1 exist that solve the matrix-balancing problem, the sequence of scaled matrices $A^{(t)} = D^{1(t)} A D^{0(t)}$ based on Sinkhorn's algorithm can still converge entry wise to the solution of (15) whenever it has a finite solution. This apparent discrepancy is explained by the fact that the solution of Problem (15) requires a weaker condition than Assumption 3 for the matrix-balancing problem. It can be stated in the following equivalent forms.

Assumption 5 (Weak Existence). (a) *There exists a nonnegative matrix $A' \in \mathbb{R}_+^{n \times m}$ that inherits all zeros of A and has row and column sums $\boldsymbol{p}$ and $\boldsymbol{q}$. Or, equivalently,* (b) *for every pair of sets of indices $N \subsetneq [n]$ and $M \subsetneq [m]$ such that $A_{ij} = 0$ for $i \notin N$ and $j \in M$, $\sum_{i \in N} \boldsymbol{p}_i \geq \sum_{j \in M} \boldsymbol{q}_j$.*

The equivalence of the two conditions above follows from Pukelsheim and Simeone (2009, theorem 4), which also shows that they are the minimal requirements for the convergence of Sinkhorn's algorithm. Assumption 5(a) precisely guarantees that the constrained KL minimization Problem (15) is feasible and bounded. It relaxes Assumption 3(a) by allowing additional zeros in the matrix A'. Similarly, Assumption 5(b) relaxes Assumption 3(b) by allowing equality between $\sum_{i \in N} \boldsymbol{p}_i$ and $\sum_{j \in M} \boldsymbol{q}_j$, even when M, N do not correspond to a block-diagonal structure.

The distinction between Assumption 3 and Assumption 5 is important for the matrix-balancing problem and Sinkhorn's algorithm. Assumption 3 guarantees that the solutions of (2) and (15) coincide and have exactly the same zero pattern as A. If Assumption 5 holds but Assumption 3 fails, then the solution $\hat{A}$ of (15) has additional zeros relative to A, and no direct (finite and positive) scaling (D^0, D^1) exists such that $\hat{A} = D^1 A D^0$. However, the sequence of scaled matrices $\hat{A}^{(t)}$ still converges to $\hat{A}$. We call this case the weak convergence of Sinkhorn. In this case, the matrix-balancing problem is said to have a limit scaling, where some entries of D^0, D^1 in Sinkhorn iterations approach zero or ∞, resulting in additional zeros in $\hat{A}$. Below, we give an example adapted from Pukelsheim and Simeone (2009), where $\boldsymbol{p}, \boldsymbol{q} = (3, 3)$ and the scaled matrices $\hat{A}^{(t)}$ converge but no direct scaling exists:

$$A^{(t)} = D^{1(t)} A D^{0(t)} = D^{1(t)} \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}$$

$$D^{0(t)} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{3t}{2} \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{t} \end{bmatrix} \to \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}. \tag{31}$$

Under Assumption 5, Léger (2021) proves the sublinear convergence of Sinkhorn's algorithm, whereas it is known since at least Soules (1991) and Luo and Tseng (1992) that the convergence is linear under Assumption 3. It is, therefore, important to clarify the convergence behaviors of Sinkhorn's algorithm in the two settings. We next show that if Assumption 5 holds but Assumption 3 fails, then there exists an entry of $A^{(t)}$ that converges at a lower-bound rate $\Omega(1/t)$ (i.e., sublinear). Together with existing and new results in this paper, Proposition 3 fully characterizes the following convergence behavior of Sinkhorn's algorithm; whenever a direct scaling exists for the matrix-balancing problem, Sinkhorn's algorithm converges linearly. If only a limit scaling exists, then convergence deteriorates to sublinear. This generalizes the observations made by Sinkhorn and Knopp (1967) and Achilles (1993) for square matrices and uniform marginals.

Proposition 3 (Linear Versus Sublinear Convergence of Sinkhorn). *For general nonnegative matrices, Sinkhorn's algorithm converges linearly if and only if $(A, \boldsymbol{p}, \boldsymbol{q})$ satisfy Assumption 3 and Assumption 4. The convergence is sublinear if and only if the weak existence condition Assumption 5 holds but Assumption 3 fails.*

The regime of sublinear convergence also has an interpretation in the choice-modeling framework. The weak existence condition Assumption 5, when applied to $(A, \boldsymbol{p}, \boldsymbol{q})$ constructed from a choice data set, allows the case where some subset S of items is always preferred over S^C, which implies, as observed already by the early work of Ford (1957), that the log-likelihood function (5) is only maximized at the boundary of the probability simplex by shrinking s_j for $j \in S^C$ toward zero (i.e., $D_j^0 \to 0$). Incidentally, Bacharach (1965) also refers to the corresponding regime in matrix balancing as "boundary solutions."

5.4. Sharp Asymptotic Rate

Having established the global convergence and iteration complexity of Sinkhorn's algorithm when finite scalings exist, we now turn to the problem of characterizing the sharp (i.e., best-possible) asymptotic linear convergence rate as $t \to \infty$ for general nonnegative A and

nonuniform marginals $(\boldsymbol{p},\boldsymbol{q})$. Knight (2008) computed this rate for uniform $(\boldsymbol{p},\boldsymbol{q})$ under an implicit metric. Our analysis is distinct from the analysis of Knight (2008) and relies on an intrinsic orthogonality structure of Sinkhorn's algorithm, which is also different from the auxiliary normalization in our global linear convergence analysis. Note that unlike the global rate, which depends on the initial problem data A and $(\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)})$, the asymptotic rate now depends on the optimal solution $\hat{A} = D^1AD^0$ as expected.

Theorem 4 (Sharp Asymptotic Rate). *Suppose $(A,\boldsymbol{p},\boldsymbol{q})$ satisfy Assumption 3 and Assumption 4. Let $\hat{A}$ be the unique scaled matrix with target marginals $\boldsymbol{p}$, $\boldsymbol{q}$ defined in (10). Then, marginals $\boldsymbol{r}^{(t)} = A^{(t)}\mathbf{1}$, where $A^{(t)}$ is the scaled matrix after t iterations of Sinkhorn's algorithm, satisfy*

$$\lim_{t\to\infty}\frac{\|\boldsymbol{r}^{(t+1)}/\sqrt{\boldsymbol{p}}-\sqrt{\boldsymbol{p}}\|_2}{\|\boldsymbol{r}^{(t)}/\sqrt{\boldsymbol{p}}-\sqrt{\boldsymbol{p}}\|_2}=\lambda_\infty, \tag{32}$$

where the asymptotic linear rate of convergence λ_∞ is given by

$$\lambda_\infty := \lambda_2(\tilde{A}\tilde{A}^T) = \lambda_2(\tilde{A}^T\tilde{A})$$
$$\tilde{A} := \mathcal{D}(1/\sqrt{\boldsymbol{p}})\cdot\hat{A}\cdot\mathcal{D}(1/\sqrt{\boldsymbol{q}})$$

and $\lambda_2(\cdot)$ denotes the second-largest eigenvalue.

In the special case of $m = n$ and $\boldsymbol{p} = \boldsymbol{q} = \mathbf{1}$, the asymptotic rate in Theorem 4 reduces to that in Knight (2008). Note, however, that the convergence metric is different; we use the ℓ^2 norm $\|\boldsymbol{r}^{(t)}/\sqrt{\boldsymbol{p}}-\sqrt{\boldsymbol{p}}\|_2$, whereas Knight (2008) uses $\|\boldsymbol{r}^{(t)}-\boldsymbol{p}\|_*$ with an implicit norm $\|\cdot\|_*$ on $\mathbb{R}^n$. Moreover, one cannot directly extend results for square matrices, such as those in Knight (2008), to nonsquare matrices by padding them with zeros, as doing so results in target marginals that are not strictly positive. See, however, Knight (2008) for a symmetrization proposal.

The proof of Theorem 4 relies on a sequence of novel data-dependent mappings associated with Sinkhorn's algorithm. Intuitively, the dependence of the asymptotic linear rate on the second-largest eigenvalue of $\tilde{A}^T\tilde{A}$ (and $\tilde{A}\tilde{A}^T$) is because of the fact that near the fixed point $\sqrt{\boldsymbol{p}}$ of the mapping associated with Sinkhorn iterations, $\tilde{A}\tilde{A}^T$ (which is the Jacobian at $\sqrt{\boldsymbol{p}}$) approximates the first-order change in $\boldsymbol{r}^{(t)}/\sqrt{\boldsymbol{p}}$. Normally, the leading eigenvalue quantifies this change. The unique leading eigenvalue of $\tilde{A}\tilde{A}^T$ is equal to one with eigenvector $\sqrt{\boldsymbol{p}}$, which does not imply contraction. Fortunately, using the quantity $\boldsymbol{r}^{(t)}/\sqrt{\boldsymbol{p}}$ allows us to exploit the following orthogonality structure:

$$(\boldsymbol{r}^{(t)}/\sqrt{\boldsymbol{p}}-\sqrt{\boldsymbol{p}})^T\sqrt{\boldsymbol{p}}=\sum_i(\boldsymbol{r}_i^{(t)}-\boldsymbol{p}_i)=0$$

by virtue of Sinkhorn's algorithm, preserving the quantities $\boldsymbol{r}^{(t)T}\mathbf{1}_n$ for all t. Thus, the residual $\boldsymbol{r}^{(t)}/\sqrt{\boldsymbol{p}}-\sqrt{\boldsymbol{p}}$ is always orthogonal to $\sqrt{\boldsymbol{p}}$, which is both the leading eigenvector and the fixed point of the iteration. The convergence is then controlled by the second-largest eigenvalue of $\tilde{A}\tilde{A}^T$. This proof approach echoes that of the global linear convergence result in Theorem 3, where we also exploit an orthogonality condition to obtain a meaningful bound. In Theorem 3, the bound depends on the second-smallest eigenvalue of a Hessian matrix, whereas in Theorem 4, the bound depends on the second-largest eigenvalue of a Jacobian matrix.

Lastly, we note that the asymptotic rate λ_∞ is itself a Fiedler eigenvalue associated with the Laplacian that is the Schur complement of the scaled graph Laplacian

$$\begin{bmatrix}\mathcal{D}(1/\sqrt{\boldsymbol{p}}) & 0\\ 0 & \mathcal{D}(1/\sqrt{\boldsymbol{q}})\end{bmatrix}\begin{bmatrix}\mathcal{D}(\hat{A}\mathbf{1}_m) & -\hat{A}\\ -\hat{A}^T & \mathcal{D}(\hat{A}^T\mathbf{1}_n)\end{bmatrix}\begin{bmatrix}\mathcal{D}(1/\sqrt{\boldsymbol{p}}) & 0\\ 0 & \mathcal{D}(1/\sqrt{\boldsymbol{q}})\end{bmatrix}.$$

6. Conclusion

In this paper, we develop extensive connections between matrix balancing and choice modeling. We show that the maximum likelihood estimation of choice models based on the Luce axioms of choice is an instance of the canonical matrix-balancing problem. Moreover, many algorithms in choice modeling can be viewed as special cases or analogs of Sinkhorn's algorithm for matrix balancing. These connections can potentially benefit multiple disciplines. For choice modeling, they open the door to tools and insights from well-studied topics in optimization and numerical linear algebra. For matrix balancing, the connections enable us to resolve some interesting open problems on the linear convergence of Sinkhorn's algorithm for nonnegative matrices. We establish the first quantitative global linear convergence result for Sinkhorn's algorithm applied to general nonnegative matrices. Our analysis reveals the importance of algebraic connectivity for matrix balancing. We also provide the first characterization of the exact asymptotic linear rate of convergence for general nonnegative matrix and nonuniform target marginals. Lastly, we clarify the linear and sublinear convergence behaviors of Sinkhorn's algorithm under the strong and weak existence assumptions for matrix balancing. Overall, we believe that the connections established in this paper are useful for researchers from different domains and can lead to further interesting results.

Acknowledgments

The authors are grateful for insightful comments and suggestions from Kurt Anstreicher, Dimitris Bertsimas, Stéphane Bonhomme, Agostino Capponi, Serina Chang, Patrick Ding,

Yanqin Fan, Robert Freund, Han Hong, Yuchen Hu, Guido Imbens, Süleyman Kerimov, Yongchan Kwon, Frederic Koehler, Flavien Léger, Gary Qian, Pavel Shibayev, Ruoxuan Xiong, and Yinyu Ye. The authors also thank the area editor, the associate editor, and the three anonymous referees for helping to improve the manuscript significantly.

References

Abowd JM, Kramarz F, Margolis DN (1999) Computing person and firm effects using linked longitudinal employer-employee data. *Econometrica* 67(6):1411–1453.

Achilles E (1993) Implications of convergence rates in sinkhorn balancing. *Linear Algebra Its Appl.* 187:109–112.

Agarwal A, Negahban S, Wainwright MJ (2010) Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Adv. Neural Inform. Processing Systems*, vol. 23 (Curran Associates Inc., Red Hook, NY).

Agarwal A, Patil P, Agarwal S (2018) Accelerated spectral ranking. *Internat. Conf. Machine Learn*. (PMLR, New York), 70–79.

Allen-Zhu Z, Li Y, Oliveira R, Wigderson A (2017) Much faster algorithms for matrix scaling. *2017 IEEE 58th Annual Sympos. Foundations Comput. Sci. (FOCS)* (IEEE, Piscataway, NJ), 890–901.

Altschuler J, Niles-Weed J, Rigollet P (2017) Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Adv. Neural Inform. Processing Systems*, vol. 30 (Curran Associates Inc., Red Hook, NY).

Bacharach M (1965) Estimating nonnegative matrices from marginal data. *Internat. Econom. Rev.* 6(3):294–310.

Bacharach M (1970) *Biproportional Matrices and Input-Output Change*, vol. 16 (CUP Archive, Cambridge, UK).

Balakrishnan H, Hwang I, Tomlin CJ (2004) Polynomial approximation algorithms for belief matrix maintenance in identity management. *2004 43rd IEEE Conf. Decision Control (CDC)*, IEEE catalog no. 04ch37601, vol. 5 (IEEE, Piscataway, NJ), 4874–4879.

Batsell RR, Polking JC (1985) A new class of market share models. *Marketing Sci.* 4(3):177–198.

Bauer FL (1963) Optimally scaled matrices. *Numerische Mathematik* 5(1):73–87.

Beck A, Tetruashvili L (2013) On the convergence of block coordinate descent type methods. *SIAM J. Optim.* 23(4):2037–2060.

Berkson J (1944) Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.* 39(227):357–365.

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica* 63(4):841–890.

Bertsekas DP (1997) Nonlinear programming. *J. Oper. Res. Soc.* 48(3):334–334.

Bertsekas D (2016) *Nonlinear Programming*, vol. 4 (Athena Scientific, Nashua, NH).

Birch M (1963) Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B Methodological* 25(1):220–233.

Bishop CM, Nasrabadi NM (2006) *Pattern Recognition and Machine Learning*, vol. 4 (Springer, New York).

Blanchet J, Chen L, Zhou XY (2022) Distributionally robust mean-variance portfolio selection with Wasserstein distances. *Management Sci.* 68(9):6382–6410.

Blanchet J, Gallego G, Goyal V (2016) A Markov chain approximation to choice modeling. *Oper. Res.* 64(4):886–905.

Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* 56(3):830–857.

Bong H, Rinaldo A (2022) Generalized results for the existence and consistency of the MLE in the Bradley–Terry–Luce model. *Internat. Conf. Machine Learn.* (PMLR, New York), 2160–2177.

Boyd S, Ghosh A, Prabhakar B, Shah D (2006) Randomized gossip algorithms. *IEEE Trans. Inform. Theory* 52(6):2508–2530.

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39(3/4):324–345.

Bregman LM (1967a) Proof of the convergence of Sheleikhovskii's method for a problem with transportation constraints. *USSR Comput. Math. Math. Phys.* 7(1):191–204.

Bregman LM (1967b) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* 7(3):200–217.

Brualdi RA (1968) Convex sets of non-negative matrices. *Canadian J. Math.* 20:144–157.

Bushell PJ (1973) Hilbert's metric and positive contraction mappings in a Banach space. *Arch. Rational Mech. Anal.* 52:330–338.

Carlier G (2022) On the linear convergence of the multimarginal Sinkhorn algorithm. *SIAM J. Optim.* 32(2):786–794.

Carlier G, Dupuy A, Galichon A, Sun Y (2023) SISTA: Learning optimal transport costs under sparsity constraints. *Comm. Pure Appl. Math.* 76(9):1659–1677.

Chakrabarty D, Khanna S (2021) Better and simpler error analysis of the Sinkhorn–Knopp algorithm for matrix scaling. *Math. Programming* 188(1):395–407.

Chang S, Koehler F, Qu Z, Leskovec J, Ugander J (2024) Inferring dynamic networks from marginals with iterative proportional fitting. *Forty-First Internat. Conf. Machine Learn.* (PMLR, New York).

Chen L, Kyng R, Liu YP, Peng R, Gutenberg MP, Sachdeva S (2022) Maximum flow and minimum-cost flow in almost-linear time. *2022 IEEE 63rd Annual Sympos. Foundations Comput. Sci. (FOCS)* (IEEE, Piscataway, NJ), 612–623.

Cohen MB, Madry A, Tsipras D, Vladu A (2017) Matrix scaling and balancing via box constrained newton's method and interior point methods. *2017 IEEE 58th Annual Sympos. Foundations Comput. Sci. (FOCS)* (IEEE, Piscataway, NJ), 902–913.

Cottle RW, Duvall SG, Zikan K (1986) A Lagrangean relaxation algorithm for the constrained matrix problem. *Naval Res. Logist. Quart.* 33(1):55–76.

Critchlow DE, Fligner MA, Verducci JS (1991) Probability models on rankings. *J. Math. Psych.* 35(3):294–318.

Cuturi M (2013) Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inform. Processing Systems*, vol. 26 (Curran Associates Inc., Red Hook, NY).

De Paula A (2017) Econometrics of network models. *Adv. Econom. Econometrics Theory Appl. Eleventh World Congress* (Cambridge University Press, Cambridge, UK), 268–323.

Deming WE, Stephan FF (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* 11(4):427–444.

Di Marino S, Gerolin A (2020) An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *J. Sci. Comput.* 85(2):27.

Djoković D (1970) Note on nonnegative matrices. *Proc. Amer. Math. Soc.* 25(1):80–82.

Dvurechensky P, Gasnikov A, Kroshnin A (2018) Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. *Internat. Conf. Machine Learn.* (PMLR, New York), 1367–1376.

Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the web. *Proc. 10th Internat. Conf. World Wide Web* (ACM, New York), 613–622.

Dykstra O (1956) A note on the rank analysis of incomplete block designs—Applications beyond the scope of existing tables. *Biometrics* 12(3):301–306.

Elo AE (1978) *The Rating of Chessplayers, Past and Present* (Arco Publishing, New York).

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1–2):115–166.

Fiedler M (1973) Algebraic connectivity of graphs. *Czechoslovak Math. J.* 23(2):298–305.
Fienberg SE (1970) An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* 41(3):907–917.
Fofana I, Lemelin A, Cockburn J (2002) Balancing a social accounting matrix. Working paper, CREFA-Université Laval, Quebec City.
Ford LR (1957) Solution of a ranking problem from binary comparisons. *Amer. Math. Monthly* 64(8P2):28–33.
Ford LR, Fulkerson DR (1956) Maximal flow through a network. *Canadian J. Math.* 8:399–404.
Ford LR, Fulkerson DR (1957) A simple algorithm for finding maximal network flows and an application to the Hitchcock problem. *Canadian J. Math.* 9:210–218.
Franklin J, Lorenz J (1989) On the scaling of multidimensional matrices. *Linear Algebra Its Appl.* 114:717–735.
Gale D (1957) A theorem on flows in networks. *Pacific J. Math.* 7(2):1073–1082.
Galichon A, Salanié B (2021) Matching with trade-offs: Revealed preferences over competing characteristics. Preprint, submitted February 25, https://arxiv.org/abs/2102.12811.
Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* 48(2):603–655.
Ghosal P, Nutz M (2025) On the convergence rate of Sinkhorn's algorithm. *Math. Oper. Res.*, ePub ahead of print May 8, https://doi.org/10.1287/moor.2024.0427.
Good IJ (1963) Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* 34(3):911–934.
Hajek B, Oh S, Xu J (2014) Minimax-optimal inference from partial rankings. *Adv. Neural Inform. Processing Systems* (MIT Press, Cambridge, MA), 1475–1483.
Hall P (1935) On representatives of subsets. *J. London Math. Soc.* 1(1):26–30.
Hausman JA, Ruud PA (1987) Specifying and testing econometric models for rank-ordered data. *J. Econometrics* 34(1–2):83–104.
Hazan E (2016) Introduction to online convex optimization. *Foundations Trends Optim.* 2(3–4):157–325.
Hendrickx J, Olshevsky A, Saligrama V (2020) Minimax rate for learning from pairwise comparisons in the BTL model. *Internat. Conf. Machine Learn.* (PMLR, New York), 4193–4202.
Hunter DR (2004) MM algorithms for generalized Bradley–Terry models. *Ann. Statist.* 32(1):384–406.
Idel M (2016) A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. Preprint, submitted September 20, https://arxiv.org/abs/1609.06349.
Ireland CT, Kullback S (1968) Contingency tables with given marginals. *Biometrika* 55(1):179–188.
Jochmans K, Weidner M (2019) Fixed-effect regressions on network data. *Econometrica* 87(5):1543–1560.
Kalantari B, Khachiyan L (1993) On the rate of convergence of deterministic and randomized RAS matrix scaling algorithms. *Oper. Res. Lett.* 14(5):237–244.
Kalantari B, Khachiyan L (1996) On the complexity of nonnegative-matrix scaling. *Linear Algebra Its Appl.* 240:87–103.
Kalantari B, Lari I, Ricca F, Simeone B (2008) On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. *Math. Programming* 112(2):371–401.
Keener JP (1993) The Perron–Frobenius theorem and the ranking of football teams. *SIAM Rev.* 35(1):80–93.
Kendall MG, Smith BB (1940) On the method of paired comparisons. *Biometrika* 31(3/4):324–345.
Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. *J. ACM* 46(5):604–632.
Knight PA (2008) The Sinkhorn–Knopp algorithm: Convergence and applications. *SIAM J. Matrix Anal. Appl.* 30(1):261–275.
Kruithof J (1937) Telefoonverkeersrekening. *De Ingenieur* 52:15–25.
Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Oper Res. Management Sci. Age Analytics* 2019(October):130–166.
Kullback S (1997) *Information Theory and Statistics* (Dover, Mineola, NY).
Kumar R, Tomkins A, Vassilvitskii S, Vee E (2015) Inverting a steady-state. *Proc. Eighth ACM Internat. Conf. Web Search Data Mining* (ACM, New York), 359–368.
Lamond B, Stewart NF (1981) Bregman's balancing method. *Transportation Res. Part B Methodological* 15(4):239–248.
Landau E (1895) Zur relativen wertbemessung der turnierresultate. *Deutsches Wochenschach* 11:366–369.
Lange K (2016) *MM Optimization Algorithms* (SIAM, Philadelphia).
Léger F (2021) A gradient descent perspective on Sinkhorn. *Appl. Math. Optim.* 84(2):1843–1855.
Linial N, Samorodnitsky A, Wigderson A (1998) A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. *Proc. Thirtieth Annual ACM Sympos. Theory Comput.* (Association for Computing Machinery, New York), 644–652.
Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (Wiley, New York).
Luo ZQ, Tseng P (1992) On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.* 72(1):7–35.
Maystre L, Grossglauser M (2015) Fast and accurate inference of Plackett–Luce models. *Adv. Neural Inform. Processing Systems* (MIT Press, Cambridge, MA), 172–180.
Maystre L, Grossglauser M (2017) Choicerank: Identifying preferences from node traffic in networks. *Internat. Conf. Machine Learn.* (PMLR, New York), 2354–2362.
McFadden D (1973) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.
McFadden D, Train K (2000) Mixed MNL models for discrete response. *J. Appl. Econom.* 15(5):447–470.
Menon M (1968) Matrix links, an extremization problem, and the reduction of a non-negative matrix to one with prescribed row and column sums. *Canadian J. Math.* 20:225–232.
Negahban S, Oh S, Shah D (2012) Iterative ranking from pairwise comparisons. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 2474–2482.
Negahban S, Oh S, Shah D (2017) Rank centrality: Ranking from pairwise comparisons. *Oper. Res.* 65(1):266–287.
Nemirovski A, Rothblum U (1999) On complexity of matrix scaling. *Linear Algebra Its Appl.* 302:435–460.
Newman M (2023) Efficient computation of rankings from pairwise comparisons. *J. Machine Learn. Res.* 24(238):1–25.
Noothigattu R, Peters D, Procaccia AD (2020) Axioms for learning from pairwise comparisons. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 17745–17754.
Orabona F (2019) A modern introduction to online learning. Preprint, submitted December 31, https://arxiv.org/abs/1912.13213.
Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, Stanford, CA.
Peyré G, Cuturi M (2019) Computational optimal transport: With applications to data science. *Foundations Trends Machine Learn.* 11(5–6):355–607.
Plackett RL (1975) The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* 24(2):193–202.
Pukelsheim F, Simeone B (2009) On the iterative proportional fitting procedure: Structure of accumulation points and l1-error analysis. Working paper, Institute fur Mathematik, Universitatsstrasse, Augsburg, Germany.

Ragain S, Ugander J (2016) Pairwise choice Markov chains. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 3198–3206.

Rajkumar A, Agarwal S (2014) A statistical convergence perspective of algorithms for rank aggregation from pairwise data. *Internat. Conf. Machine Learn.* (PMLR, New York), 118–126.

Ruiz D (2001) A scaling algorithm to equilibrate both rows and columns norms in matrices. Technical Report No. CM-P00040415.

Ruschendorf L (1995) Convergence of the iterative proportional fitting procedure. *Ann. Statist.* 23(4):1160–1174.

Schneider MH, Zenios SA (1990) A comparative study of algorithms for matrix balancing. *Oper. Res.* 38(3):439–455.

Seshadri A, Ragain S, Ugander J (2020) Learning rich rankings. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates Inc., Red Hook, NY), 9435–9446.

Shah N, Balakrishnan S, Bradley J, Parekh A, Ramchandran K, Wainwright M (2015) Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Artificial Intelligence Statist.* (PMLR, New York), 856–865.

Sinkhorn R (1964) A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.* 35(2):876–879.

Sinkhorn R (1974) Diagonal equivalence to matrices with prescribed row and column sums. II. *Proc. Amer. Math. Soc.* 45(2):195–198.

Sinkhorn R, Knopp P (1967) Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* 21(2):343–348.

Soufiani HA, Chen W, Parkes DC, Xia L (2013) Generalized method-of-moments for rank aggregation. *Adv. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY), 2706–2714.

Soules GW (1991) The rate of convergence of sinkhorn balancing. *Linear Algebra Its Appl.* 150:3–40.

Spielman DA (2007) Spectral graph theory and its applications. *48th Annual IEEE Sympos. Foundations Comput. Sci. (FOCS'07)* (IEEE, Piscataway, NJ), 29–38.

Stone R (1962) Multiple classifications in social accounting. *Bull. de L'Institut Internat. de Statistique* 39(3):215–233.

Taşkesen B, Shafieezadeh-Abadeh S, Kuhn D (2023) Semi-discrete optimal transport: Hardness, regularization and numerical solution. *Math. Programming* 199(1):1033–1106.

Theil H (1967) *Economics and Information Theory* (North-Holland Publishing Company, Amsterdam).

Thionet P (1964) Note sur le remplissage d'un tableau à double entrée. *J. de la Société Française de Statistique* 105:228–247.

Thurstone LL (1927) The method of paired comparisons for social values. *J. Abnormal Soc. Psych.* 21(4):384–400.

Tomlin JA (2003) A new paradigm for ranking pages on the World Wide Web. *Proc. 12th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 350–355.

Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* 109: 475–494.

Tseng P, Bertsekas DP (1987) Relaxation methods for problems with strictly convex separable costs and linear constraints. *Math. Programming* 38(3):303–321.

Tversky A (1972) Elimination by aspects: A theory of choice. *Psych. Rev.* 79(4):281–299.

Vojnović M, Yun S (2016) Parameter estimation for generalized thurstone choice models. *Internat. Conf. Machine Learn.* (PMLR, New York), 498–506.

Vojnović M, Yun SY, Zhou K (2020) Convergence rates of gradient descent and MM algorithms for Bradley–Terry models. *Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 1254–1264.

Vojnović M, Yun S-Y, Zhou K (2023) Accelerated MM algorithms for inference of ranking scores from comparison data. *Oper. Res.* 71(4):1318–1342.

Xiao L, Boyd S, Kim SJ (2007) Distributed average consensus with least-mean-square deviation. *J. Parallel Distributed Comput.* 67(1): 33–46.

Yule GU (1912) On the methods of measuring association between two attributes. *J. Roy. Statist. Soc.* 75(6):579–652.

Zermelo E (1929) Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 29(1):436–460.

Zhaonan Qu is a postdoctoral scholar at Columbia University's Data Science Institute and an incoming faculty member at the Martin Tuchman School of Management at the New Jersey Institute of Technology. His research interests include choice modeling, causal inference, optimization, and machine learning, especially the interactions between them in theoretical, empirical, and policy questions.

Alfred Galichon is a professor of economics (Arts & Science) and mathematics (Courant Institute) at New York University. His research interests span widely across economics and include econometrics, microeconomic theory, and data science. He is one of the pioneers of the use of optimal transport theory in econometrics and the author of a monograph on the topic, *Optimal Transport Methods in Economics*. His works have appeared in journals such as *Annals of Statistics*, *Journal of Political Economy*, *Econometrica*, and *Review of Economic Studies*.

Wenzhi Gao is a PhD student at the Institute for Computational and Mathematical Engineering, Stanford University. His research interests lie in numerical optimization algorithms, online learning, and stochastic optimization.

Johan Ugander is an associate professor of management science and engineering at Stanford University. His research develops algorithmic and statistical frameworks for analyzing social networks, social systems, and other large-scale social and behavioral data. His awards include an NSF CAREER Award, a Young Investigator Award from the Army Research Office, several best paper awards, and the 2016 Eugene L. Grant Undergraduate Teaching Award from the Department of Management Science & Engineering.