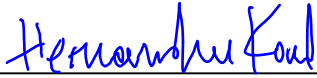


ALGORITHMS FOR DISCRETE DATA IN STATISTICS AND OPERATIONS RESEARCH

BY

WILLIAM K. SCHWARTZ

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy in Applied Mathematics  
in the Graduate College of the  
Illinois Institute of Technology

Approved   
Adviser

Approved   
Co-Adviser

Chicago, Illinois  
December 2021

© Copyright by  
WILLIAM K. SCHWARTZ  
December 2021

## ACKNOWLEDGMENTS

Thank you to my advisers, Hemanshu Kaul and Sonja Petrović, for years of teaching, guidance, feedback, and research-topic ideas. Thank you also to my other committee members, Ruoting Gong and Sanjiv Kapoor.

Chapter 2 includes joint work with Sonja and Hemanshu. We extracted an article, Schwartz et al. (2021), from the chapter. Many thanks also to Alessandro Rinaldo for his pivotal pointer to Kuchler and Sørensen (1997), as well as for his feedback when he sat on the committee that heard my dissertation-proposal in which I presented some of the results in chapter 2. Thanks also to Ruoting, who helped me catch up on the measure theory I needed for subsection 2.3.2. The Air Force Office of Scientific Research’s grant FA9550-14-1-0141 supported Sonja’s and my initial work on this project.

Chapter 3 includes joint work with Sonja, Debdeep Pati, and Vishesh Karwa: Karwa et al. (2021–present). The project grew out of a workshop Sonja hosted at IIT over May 8–11, 2017, where I met Debdeep and Vishesh, as well as Alessandro. Along with Elizabeth Gross, Nicolas Kim, Mateja Raič, Despina Stasi, and Dane Wilburne, we hashed out the skeleton of the ideas that the chapter analyzes. I am particularly grateful to Sonja for her feedback and insights on early drafts of this article, and to Dane for introducing me to **ERGMS** and algebraic statistics.

Chapter 4 benefited from many insightful discussions with James Bono, Allan T. Ingraham, Shreyas Ravi, Christopher T. Sojourner, Katherine Senseman, and Gabriel Perez-Putnam. Allan, Shreyas, and Christopher also worked with me on some combinatorial-auction ideas that did not make it into the chapter. However, my giving a presentation, Schwartz et al. (2019), on that joint work at the 2019 INFORMS Annual Meeting gave me the opportunity at the same conference to present a poster, Schwartz (2019), containing an early cut of the ideas that did make it into chapter 4. Hemanshu reviewed and gave me much helpful feedback on the poster as well as chapter 4 itself. I would also like to thank him for his help with Schwartz (2016), an earlier project on a different discrete optimization problem arising from the same type of auctions that chapter 4 discusses. Writing that paper helped prepare me to write chapter 4.

Thanks to Robert Kulick for his insightful correspondence with me about economic modeling, which helped me with chapter 1.

Most importantly thank you to Jillian Foley—my tragically underpaid graphic designer, editor, research librarian, wife, and mother of our daughter—without whom I would have written none of the chapters. She kept me alive these last six years in graduate school; taken on more than her fair share of childcare while I finished writing, even while she has been writing her own theses; and was a sounding board for more ideas that didn’t make it into this thesis than that are in here total.

## AUTHORSHIP STATEMENT

I am the sole author of all the writing in this thesis except as I have otherwise clearly indicated by a complete citation to the source material per standard academic practice.

In accordance with the norms of the academic community of applied mathematics (see Illinois Institute of Technology [IIT], n.d., app. S “Authorship”), this thesis reports the following collaborative research. Chapter 2 reports research I conducted partially in collaboration with Sonja Petrović and Hemanshu Kaul. The chapter formed the basis of an article manuscript by all three of us: Schwartz et al. (2021). The manuscript is submitted for publication. Chapter 3 reports research I conducted partially in collaboration with Sonja, Debdeep Pati, and Vishesh Karwa. The chapter forms the basis of an article manuscript by all four of us: Karwa et al. (2021–present). The manuscript is in preparation, and we intend to submit it for publication in the coming months. Additionally I am the sole author of the software that generated the figures in chapter 3. Chapter 4 reports research I conducted alone. The chapter forms the basis of an article manuscript by myself. The manuscript is in preparation, and I intend to submit it for publication in the coming months.

William K. Schwartz  
November 1, 2021  
Chicago

## TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS . . . . .	iii
AUTHORSHIP STATEMENT . . . . .	iv
TABLE OF CONTENTS . . . . .	v
LIST OF TABLES . . . . .	vi
LIST OF FIGURES . . . . .	vii
LIST OF SYMBOLS . . . . .	viii
ABSTRACT . . . . .	ix
 CHAPTER	
1. INTRODUCTION . . . . .	1
2. INTERPRETABLE DATA REDUCTION FOR NETWORK MARKOV CHAINS	8
2.1. Introduction . . . . .	8
2.2. Sufficiency and Parametric Markov Chains . . . . .	9
2.3. Conditional Exponential Families . . . . .	14
2.4. Permutation-Uniform Markov Chains . . . . .	42
2.5. Markov Chains of Graphs . . . . .	69
2.6. Conclusion . . . . .	86
3. HYPOTHESIS TESTS FOR MIXED MEMBERSHIP STOCHASTIC BLOCK MODELS . . . . .	88
3.1. Model . . . . .	88
3.2. Goodness of Fit . . . . .	100
4. LEXICOGRAPHIC WINNER DETERMINATION . . . . .	122
4.1. Model . . . . .	122
4.2. Pre-Solving . . . . .	142
4.3. Upper Bounds . . . . .	145
4.4. Dynamic Programming . . . . .	158
4.5. Price Determination . . . . .	173
BIBLIOGRAPHY . . . . .	177

## LIST OF TABLES

Table		Page
2.1	Graph theoretic operations on edge-indicator vectors . . . . .	82
3.1	The set $\mathcal{D}_n$ of allowed edges . . . . .	90
3.2	The set $\mathcal{G}_n$ of adjacency matrices . . . . .	91
3.3	The set $\mathcal{B}_{n,k}$ of sender- or receiver-block matrices . . . . .	92
3.4	The set $\mathcal{B}_{n,k}$ of block-assignments arrays . . . . .	92
4.1	Some recent combinatorial auctions for spectrum . . . . .	123
4.2	An illustrative bid stack . . . . .	129
4.3	How to add side constraints to a canonical WDP <sup>x</sup> . . . . .	135
4.4	Summary of lemma 4.4.12: how to read the recursion memo . . . . .	172

## LIST OF FIGURES

Figure		Page
3.1	Histogram of a simple-hypothesis p-values from a single, conditional PMF .	115
3.2	Convergence of p-value estimates . . . . .	120
3.3	Approximate lower bound on the number of iterations needed in Monte Carlo integration . . . . .	121

## LIST OF SYMBOLS

Symbol	Definition
$\mathbb{1}$	Indicator function for events: $\mathbb{1}(E) = 1$ if $E$ happens and zero if not.
$\mathcal{D}_n$	Dyads $ij \equiv (i, j)$ on vertex set $[n]$
$\mathbb{E}$	Expectation operator
$\mathbb{N}$	Natural numbers, i.e., $\{0, 1, \dots\}$
$n$	Chapters 2 and 3: number of nodes in network; chapter 4: number of bids total across all bidders
$[n]$	$\{1, \dots, n\}$ when $n \in \mathbb{N}$
$\mathbb{P}$	Probability measure
$\mathbb{R}$	Real numbers, i.e., $(-\infty, \infty)$



## ABSTRACT

This thesis develops mathematical background for the design of algorithms for discrete-data problems, two in statistics and one in operations research. Chapter 1 gives some background on what chapters 2 to 4 have in common. It also defines some basic terminology that the other chapters use.

Chapter 2 offers a general approach to modeling longitudinal network data, including exponential random graph models (ERGMs), that vary according to certain discrete-time Markov chains (The abstract of chapter 2 borrows heavily from the abstract of Schwartz et al., 2021). It connects conditional and Markovian exponential families, permutation-uniform Markov chains, various (temporal) ERGMs, and statistical considerations such as dyadic independence and exchangeability. Markovian exponential families are explored in depth to prove that they and only they have exponential family finite sample distributions with the same parameter as that of the transition probabilities. Many new statistical and algebraic properties of permutation-uniform Markov chains are derived. We introduce *exponential random  $t$ -multigraph models*, motivated by our result on replacing  $t$  observations of a permutation-uniform Markov chain of graphs with a single observation of a corresponding multigraph. Our approach simplifies analysis of some network and autoregressive models from the literature. Removing models' temporal dependence but not interpretability permitted us to offer closed-form expressions for maximum likelihood estimators that previously did not have closed-form expression available.

Chapter 3 designs novel, exact, conditional tests of statistical goodness-of-fit for *mixed membership stochastic block models* (MMSBMs) of networks, both directed and undirected. The tests employ a  $\chi^2$ -like statistic from which we define p-values for the general null hypothesis that the observed network's distribution is in the MMSBM as well as for the simple null hypothesis that the distribution is in the MMSBM with specified parameters. For both tests the alternative hypothesis is that the distribution is unconstrained, and they both assume we have observed the block assignments. As exact tests that avoid asymptotic arguments, they are suitable for both small and large networks. Further we provide and

analyze a Monte Carlo algorithm to compute the p-value for the simple null hypothesis. In addition to our rigorous results, simulations demonstrate the validity of the test and the convergence of the algorithm. As a conditional test, it requires the algorithm sample the fiber of a sufficient statistic. In contrast to the Markov chain Monte Carlo samplers common in the literature, our algorithm is an exact simulation, so it is faster, more accurate, and easier to implement. Computing the p-value for the general null hypothesis remains an open problem because it depends on an intractable optimization problem. We discuss the two schools of thought evident in the literature on how to deal with such problems, and we recommend a future research program to bridge the gap those two schools.

Chapter 4 investigates an auctioneer’s revenue maximization problem in combinatorial auctions. In combinatorial auctions bidders express demand for discrete packages of multiple units of multiple, indivisible goods. The auctioneer’s NP-complete *winner determination problem* (WDP) is to fit these packages together within the available supply to maximize the bids’ sum. To shorten the path practitioners traverse from from legalese auction rules to computer code, we offer a new WDP formalism to reflect how government auctioneers sell billions of dollars of radio-spectrum licenses in combinatorial auctions today. It models common tie-breaking rules by maximizing a sum of bid vectors lexicographically. After a novel pre-solving technique based on package bids’ marginal values, we develop an algorithm for the WDP. In developing the algorithm’s branch-and-bound part adapted to lexicographic maximization, we discover a partial explanation of why classical WDP has been successful in using the linear programming relaxation: it equals the Lagrangian dual. We adapt the relaxation to lexicographic maximization. The algorithm’s dynamic-programming part retrieves already computed partial solutions from a novel data structure suited specifically to our WDP formalism. Finally we show that the data structure can “warm start” a popular algorithm for solving for opportunity-cost prices.

## Chapter 1

## INTRODUCTION

Over the next three chapters, I present a mosaic of models ranging from the wholly abstract to those describing specific, real-world events. They include statistical models of longitudinal data and of samples of size one. They include optimization problems that I know how to solve and optimization problems that I do not. Each chapter is mostly self contained, defining its own notation and terminology, because none discusses the models of the others.<sup>1</sup>

What they do share, however, stems from my motivation in choosing the topics and how I have pursued them. My interest in specifically *applied* mathematics arose from first studying economics and later working as an economist, specializing in industrial organization (antitrust) and auctions, before I matriculated at MIT. From this background I have developed a predisposition toward models of social phenomena that are *discrete*, as in indivisible and not continuous. Economics models often propose that some collection of people make choices that optimize some social welfare or private payoff function, at least in some average sense. These models can be discrete in two senses: in the set of actors or in the actors' sets of available choices. My impression is that the more "micro" the microeconomics—examining smaller numbers of actors in more controlled environments, such as auctions—the more quantitatively predictive the model. In part this is because such models drop the continuous approximations that models of larger markets make. When millions of people participate in a market, we can pretend that adding one more hungry mouth or one more productive farmer is an infinitesimal change, ignoring such details as which farmer sells food to which family. By contrast markets with, say, dozens of participants force us to forego the calculus that has made continuous models more accessible, at least to economists, than discrete models have been. As models add more

---

<sup>1</sup>↑I repeat a few definitions across chapters because notation differs. In any case, if you read the thesis straight through, the repetitions are sufficiently widely separated as to be useful reminders. If you skim the thesis looking for topics of interest, the repetitions are even more useful.

realistic constraints, like who sells to whom, the choice sets become discrete and require mixed or integer optimization to solve.

The difference in age of continuous versus discrete economic models reflects this dynamic. Modern mathematical economics began in earnest in the so-called “marginal revolution” of the second half of the nineteenth century with the continuous supply-and-demand models of, e.g., Léon Walras, Carl Menger, or Alfred Marshall.<sup>2</sup> Discrete models in economics are much younger. A major source is game theory, the modern form of which dates only to 1944 when von Neumann and Morgenstern published *Theory of Games and Economic Behavior* (Fudenberg & Tirole, 1991, p. xviii). Elsewhere in social science, the literature on social networks also dates from the mid-twentieth century (see subsection 2.5.1). Given this attention gap, I have a hunch that more low hanging fruit remains in the study of social phenomena that are discrete than continuous.

And so all three chapters draw inspiration from social science literature or directly address social phenomena. Major inspirations for chapters 2 and 3 were the articles Airoidi et al. (2008) and Hanneke et al. (2010) on social networks. Chapter 4 addresses a class of auctions for discrete goods. Moreover all the chapters’ models are all finite and discrete. Chapter 4 considers the sale by an auctioneer of discrete items to a number of distinct bidders. Chapters 2 and 3 both address the statistics of finite samples of networks.

A *network* or *graph* is a (finite) set of *nodes* or *vertices* representing anything indivisible and mutually distinct, say, people. In both chapters 2 and 3 we assign consecutive numbers one through  $n$ , a positive integer, to the nodes of interest, and then just operate on networks whose node set is  $[n] := \{1, \dots, n\}$ . *Edges* between nodes represent bilateral relationships between them. Edges may be either *directed* (Alice pays Bob) or *undirected* (Alice and Bob transact). We always consider all of the edges in a graph as directed or all of them as undirected. In the former case, edges are ordered pairs of nodes, and we read the edge as going from the first to the second node in the pair. In the latter case we can form undirected edges by including the directed edge in both directions. However, if

---

<sup>2</sup>↑Carl Menger was the father both of the so-called Austrian school of economics and of MIT’s own Karl Menger.

we also assume, as chapter 2 does, that networks of interest are *simple*, meaning no node forms an edge with itself, called a *self-loop*, then we identify edges with *dyads*, or sets of two distinct nodes. Chapter 3 in contrast permits networks to be simple or non-simple, directed or undirected. Regardless of the exact definition that these assumptions imply, both chapters use  $\mathcal{D}_n$  to denote the set of all possible edges among nodes in  $[n]$ . Chapter 3 uses  $\mathcal{G}_n$  to denote the set of graphs on the node set  $[n]$ . However, chapter 2 uses  $\mathcal{G}_{n,1}$  because the chapter also considers the possibility of more than one copy of each edge. More generally, for a positive integer  $t$ , a *multigraph* in  $\mathcal{G}_{n,t}$  can have up to  $t$  copies of each allowed edge in  $\mathcal{D}_n$ . The utility of multigraphs to the chapter is the connection it describes between multigraphs and *temporal networks*, sequences of snapshots of edges over time longitudinally for a fixed set of nodes.

Another effect of my background as a working economist on my research priorities is that all three chapters develop mathematics that has the potential to solve data-driven problems. The chapters describe algorithms using formulas, sequences, recursive definitions, and the occasional table or pseudo-code listing. Chapter 2 discusses a class of statistical models that generate temporal network data. If a data set tracks relationships among a fixed set of nodes, the chapter says how to read in the data and rewrite it as a multigraph susceptible to other network analytic techniques that already exist for single snapshots of networks but have not been extended to time series data. Then chapter 2's theorems say how to translate conclusions about the multigraph back into conclusions about the temporal network (section 2.6 summarizes the process). Chapter 3 offers a Monte Carlo algorithm 3.2.2 for testing the goodness of fit between a single observation of a network and a popular statistical model of such networks. That algorithm's subroutine, algorithm 3.2.1, produces random draws from the subset of networks on the same set of nodes as the input data that would have produced the same estimates for the model's parameters. Such a subroutine is useful in its own right for simulations. Figures 3.1 and 3.2 report simulations I implemented of the goodness-of-fit test itself. Chapter 4 details an algorithm for determining who wins

what in a type of auction governments use to lease multiple radio frequencies at a time.<sup>3</sup> Bidders express demand in terms of packages that the auctioneer isn't allowed to split apart, like "one block of cheese and one bottle of wine" or "no cheese and two bottles of wine". The input to the algorithm is a spreadsheet like the one in table 4.2. I formulated the problem specifically to model how auctioneers actually implement the auctions. While I have implemented several parts of those algorithms in working code, which contributed to Bono et al. (2019) and Schwartz et al. (2019), building a fully functional prototype awaits future research.<sup>4</sup>

Both chapters 2 and 3 began with the goal of applying algebraic statistical techniques to new models of *random graphs*, networks on a fixed set of nodes whose edge exist randomly according to some probability distribution. The set  $\mathcal{G}_n$  of networks on nodes named one through  $n$  is a finite set, and we can pick the probability distribution of edges' existence by assigning each graph  $g \in \mathcal{G}_n$  a probability  $\mu(g)$ . By making that probability  $\mu_\theta(g)$  depend on a real-valued vector *parameter*  $\theta := (\theta_1, \dots, \theta_d)$  that we pick from a fixed *parameter space*  $\Theta$ , we create a *statistical model*, or just *model*,  $\mathcal{M} := \{\mu_\theta \mid \theta \in \Theta\}$ . Models that fit certain criteria, which chapter 2 explains in great detail and chapter 3 revisits, are *exponential families*, and *exponential random graph models* (ERGM) are statistical models of random graphs that are also exponential families. In some statistical tests of data potentially coming from a distribution in an ERGM, it is helpful to simulate uniformly random draws from a subset of  $\mathcal{G}_n$  that depend on the particular ERGM. To do so, we can use what I will lump together as *ERGM algebraic statistics techniques* (EASTS) (for an overview of EASTS [but that didn't use the term], see Petrović, 2015; Diaconis & Sturmfels, 1998, stated and proved what has come to be known as the *fundamental theorem of algebraic statistics*). Both chapters 2

---

<sup>3</sup>Chapter 4's theorems 4.3.7 and 4.4.4, lemma 4.4.12, and corollary 4.5.2 are that chapter's main algorithmic content. Accompanying the equations scattered about are summaries of how to put them together in steps: tables 4.3 and 4.4 and remarks 4.3.6, 4.4.1 and 4.4.11.

<sup>4</sup>One challenge has been robustly handling rounding errors emanating from optimization software packages such as `GLPK`, `Gurobi`, `lp_solve`, `CPLEX`. Surprisingly little appears in numerical analysis and computer science textbooks on how to compare floating-point values within an error tolerance, much less how to compare vectors of them lexicographically while handling infinities and NaNs. My manuscript in preparation describing the literature and the algorithms I landed on didn't fit in with the other chapters here.

and 3 were originally meant to be “simple” applications of EASTS.

Sonja Petrović, one of my advisors, originally approached me with the project that turned into chapter 2 to apply EASTS to the models of Markov chain of random networks in Hanneke et al. (2010). A (discrete time, homogeneous) *Markov chain* is a sequence of random variables  $X_0, X_1, X_2, \dots$  together with a joint probability distribution  $\mathbb{P}$  under which the probability of the next random variable  $X_{t+1}$  depends only on the state  $x$  that the current random variable  $X_t$  is in, which is to say that the sequence and  $\mathbb{P}$  have the *Markov property*  $\mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, \dots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t)$  for all nonnegative integers  $t$ . Hanneke et al. (2010)’s Markov chains took values in  $\mathcal{G}_n$  and  $\mathbb{P}_\theta(X_{t+1} = x_{t+1} \mid X_t = x_t)$ , properly parameterized, formed an exponential family. The critical first step in EASTS is identifying that the statistical model of interest is in fact an exponential family when parameterized appropriately for the application. That requires looking at the finite-sample *joint* distribution  $\mathbb{P}_\theta(X_{t+1} = x_{t+1}, X_0 = x_0, \dots, X_t = x_t)$ , not the *transition* distributions  $\mathbb{P}_\theta(X_{t+1} = x_{t+1} \mid X_t = x_t)$ . As subsection 2.5.4 concludes, only some of Hanneke et al.’s models are exponential families with the right parameterization. So we could not blanket apply EASTS to those models. Instead theorem 2.3.11 gives necessary and sufficient conditions for when a Markov chain’s finite-sample joint distribution is an exponential family with the same parameter as the transition probabilities are an exponential family. Further section 2.4 gives a sufficient condition in the form of a class of Markov chain models that have that property.

Sonja also brought me in on the project that so far has turned into chapter 3. Hers and our coauthors’ goal was to apply EASTS to a specific class of ERGMS called *mixed membership stochastic block models* (MMSBMS), which subsections 3.1.2 and 3.1.3 define. The value of EASTS is they provide tools for deriving and proving the correctness of an algorithm for sampling from subsets of  $\mathcal{G}_n$  depending on the ERGM of interest. Plug an ERGM into EASTS, get an algorithm out. The algorithm is itself a Markov chain of graphs in the appropriate subset of  $\mathcal{G}_n$ , and that Markov chain’s marginal distribution  $\mathbb{P}(X_t = x)$  converges to uniform as  $t \rightarrow \infty$ . But if you already have an algorithm, then you don’t need EASTS, especially if your algorithm, like algorithm 3.2.1, which which I stumbled upon while trying to apply EASTS

to MMSBMS, samples directly and exactly uniformly from the right subset  $\mathcal{G}_n$ . In particular algorithm 3.2.1 is faster, more accurate, and easier to implement than the Markov chain that EASTS would provide. The upshot is that both chapters 2 and 3 were supposed to be about EASTS and neither is.

The type of auction chapter 4 focuses on is called a *combinatorial auction* because bidders express demand for multiple combinations of discrete, indivisible goods—say, a “barrel of water” rather than “water”. The auctioneer can fit together multiple bidders’ package bids, not all of which are for directly comparable combinations of products. Thus a simplistic rule like “highest bidder wins” loses. This raises all sorts of fascinating economic questions (see Cramton et al., 2006b, pt. 1), but chapter 4 focuses on the mathematical and algorithmic facets, particularly the *winner determination problem*. That problem is to maximize the auctioneer’s revenue while awarding bidders only those packages they bid for and constrained by the auctioneer’s actual supply of those products. We model it as a *mathematical programming* problem: optimizing an objective function on a real vector space subject to some finite number of constraints on which vectors qualify (Minoux, 1983/1986, p. 1). In particular we write it as a zero-one integer program. The chapter presupposes the reader is familiar with integer and linear programs, but let’s briefly review the definitions. Fix an  $m \times n$  real-valued matrix  $A$ , an  $m$ -vector  $b$ , and an  $n$ -vector  $c$ :

<i>linear program</i>	<i>zero-one integer program</i>
$\underset{x}{\text{maximize}} \ c^\top x$	$\underset{x}{\text{maximize}} \ c^\top x$
$\text{subject to } Ax = b,$	$\text{subject to } Ax = b,$
$x \geq 0.$	$x \in \{0, 1\}^n.$

Both problem types permit minimizing instead of maximizing and replacing the = in the constraints  $Ax = b$  with either  $\leq$  or  $\geq$ . Bertsimas and Tsitsiklis (1997, chaps. 1–4, 10–11) is a readable but rigorous introduction to linear and integer programming.

I have followed some typographical and notational conventions across all three chapters, some of which I have already used. The expression  $a := b$  means that I am defining  $a$  to be  $b$ . Likewise this *introduction font* indicates the first time I fully define a term. (A partial definition of the same term in italics may foreshadow that full definition.)



If you come across a symbol or term you don't know, skim backward until you see the symbol next to a  $:=$  or the term in the introduction font. Latin italic capital letters are generally matrices or random variables. Bold symbols like  $\mathbf{X}$  or  $\alpha$  denote matrices or vectors. Sets are capitalized either in Greek ( $\Theta$ ), Latin calligraphic ( $\mathcal{D}_n, \mathcal{G}_n$ ), or Latin script ( $\mathcal{P}$ ) fonts.  $\mathbb{N}$  is the set of nonnegative integers.  $\mathbb{R}$  is the set of real numbers. If  $\mathbb{P}$  is a probability distribution, then  $\mathbb{E}$  is its expected value operator. However, while  $\mathbb{P}$  is always some probability distribution, I give its exact meaning in each context.

## Chapter 2

INTERPRETABLE DATA REDUCTION FOR NETWORK MARKOV CHAINS<sup>5</sup>**2.1 Introduction**

An assumption of popular models of network creation is the existence of a low dimensional sufficient statistic for often very high dimensional networks. They suppose that a large network is generated by a random process entirely determined by these sufficient statistics, allowing the estimation of similarly low dimensional parameters that are the actual goal of the researcher. Assuming a fixed set of vertices  $[n] = \{1, \dots, n\}$ , these exponential random graph models (ERGMs) are parameterized distributions on the set of graphs on  $[n]$ . Hanneke et al. (2010), Hanneke and Xing (2006), Krivitsky and Handcock (2013), and Robins and Pattison (2001) extended the models to encompass time series of graphs via Markov chains. They do so largely from the bottom up: proposing sufficient statistics of the current and next graphs in the chain and specifying transition probabilities drawn from exponential families of distributions with these sufficient statistics.

Our goal is a top-down analysis of parametric Markov chains, focusing on those discrete-time Markov chains on discrete state spaces that bear sufficient statistics of low dimension. A *sufficient statistic* for a parameter  $\theta$  of the distribution of a set  $X = \{X_0, X_1, \dots, X_t\}$  of random variables is a random variable  $\tau(X)$  such that the distribution of  $X$  conditional on  $\tau(X)$  does not depend on  $\theta$ . The *sufficiency principle* asserts that any inference about  $\theta$  should be the same whether we observe  $X = x$  or  $X = y$  as long as  $\tau(x) = \tau(y)$  (Casella & Berger, 2002, § 6.2). *Darmois-Koopman-Pitman theorems* state that, under regularity conditions, the only parameterized families of probability distributions whose sufficient statistics do not grow in dimension with the size of the sample space are the exponential families (E. L. Lehmann & Casella, 1998, § 1.6). We are concerned with situations in which the dimension of  $X$  is high and the dimension of  $\tau(X)$  is low and does

---

<sup>5</sup>This chapter includes joint work with Sonja Petrović and Hemanshu Kaul. The Air Force Office of Scientific Research's grant FA9550-14-1-0141 supported Sonja's and my initial work on this project. Schwartz et al. (2021) is an article manuscript derived from this chapter. We have submitted it for publication.

not depend on  $t$ . Section 2.2 discuss several theorems in this vein for discrete spaces.

The Darrois-Koopman-Pitman theorems suggest that restricting Markov chains to having exponential family distributions can make estimation of their parameters easier by limiting the amount of data we need to observe. This is particularly important when the state space is finite but large, such as the set of all networks on  $n$  vertices, which has  $\mathcal{O}(2^{n^2})$  networks in it. Section 2.3 classifies the transition matrices giving rise to exponential families of likelihood functions based on ideas from Feigin (1981) and Küchler and Sørensen (1997, 1998). Further, extending ideas from Gani (1955), the section discusses their algebraic structure.

Exponential-family transition probabilities' form is sufficiently constrained that, under certain circumstances, we can identify the Markov chain with an independently and identically distributed (IID) sequence on the same state space. This is the goal of section 2.4. Extending ideas first proposed in Rosenblatt (1959), we characterize all Markov chains that can be identified with an IID sequence in this way. This identification reduces analysis of the autocorrelated Markov chain to analysis of an IID sequence. The function to compute the IID sequence from the Markov chain (and its inverse, which necessarily exists) is often easy to derive in practical cases. We will conclude by looking at applications to Markov chains of networks in section 2.5.

## 2.2 Sufficiency and Parametric Markov Chains

In this section we introduce notation and terminology in preparation for reviewing the Darrois-Koopman-Pitman-type theorems for discrete-space, discrete-time Markov chains. While similar theorems exist for continuous time or space, we feel that the measure theory needed to state the theorems in their full generality would obscure their meaning and usefulness to researchers focused on discrete applications such as the network applications in section 2.5.

For a gentle introduction to Markov chains, see Hoel et al. (1972). Levin and Peres (2017) provides a modern perspective focused on mixing times.

Let  $\mathcal{S}$  be a discrete *state space*, either finite or countable. Let  $X = \{X_t\}_{t \in \mathbb{N}}$  be a time-homogeneous<sup>6</sup> Markov chain with transition matrix  $P_\theta \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$  for any  $\theta$  in some

non-empty *parameter space*  $\Theta \subseteq \mathbb{R}^d$ ,  $d \geq 1$ :

$$\mathbb{P}(X_{t+1} = b \mid X_t = a) = P_{\theta}(a, b).$$

Associate with  $X$  the process  $N = \{N_t\}_{t \in \mathbb{N}}$  of  $\mathbb{N}^{\mathcal{S} \times \mathcal{S}}$  *transition-count matrices* whose  $a, b$  entry counts the number of times  $X$  transitioned from state  $a$  to state  $b$  by time  $t$ , i.e., if  $\mathbb{1}$  is the indicator function for  $A \subseteq \mathcal{S}$ , then (cf. Gani, 1955, Eqs. 8–9; Stefanov, 1995, Eq. 7)

$$\begin{aligned} N_t(a, b) &:= \sum_{i=0}^{t-1} \mathbb{1}(X_i = a) \mathbb{1}(X_{i+1} = b), \\ N_t(+, b) &:= \sum_{a \in \mathcal{S}} N_t(a, b) = \sum_{i=1}^t \mathbb{1}(X_i = a), \\ N_t(a, +) &:= \sum_{b \in \mathcal{S}} N_t(a, b) = N_t(+, a) + \mathbb{1}(X_0 = a) - \mathbb{1}(X_t = a). \end{aligned} \tag{2.1}$$

$N_t(+, b)$  is the  $b$ th entry of the *contingency table*  $N(+, \cdot)$ , giving the number of times that  $X$  has visited state  $b$  during times  $1, \dots, t$ , inclusive. Likewise,  $N_t(a, +)$  gives the number of times that  $X$  has visited state  $a$  during times  $0, \dots, t-1$ , inclusive. We can then write the probability mass function (PMF) of  $X_1, \dots, X_t$  under  $P_{\theta}$  under the *initial condition*  $X_0 = x_0 \in \mathcal{S}$  as (Küchler & Sørensen, 1997, Eq. 3.2.5)

$$L_{\theta, x_0}^t(x_1, \dots, x_t) = \exp\left(\sum_{a, b \in \mathcal{S}} N_t(a, b) \log P_{\theta}(a, b)\right) \tag{2.2}$$

(as long as we take  $0 \log 0 = 0$ ).

Equation (2.2) has the form of an exponential family, which we now define. A set  $\mu = \{\mu_{\theta} \mid \theta \in \Theta\}$  is called a *model on*  $\mathcal{S}$  if  $\mu_{\theta}$  is a probability distribution on  $\mathcal{S}$  for each  $\theta \in \Theta$  (Petrović, 2015, § 2). If  $X$  is distributed according to  $\mu_{\theta}$  and  $\Theta$  contains distinct points  $\theta_1 \neq \theta_2$  for which  $\mu_{\theta_1} = \mu_{\theta_2}$ , then we say that  $\theta$  *unidentifiable on the basis of*  $X$  (E. L. Lehmann & Casella, 1998, § 1.5, Def. 5.2, p. 24).

The model  $\mu$  is called an *exponential family* if there exist  $\ell \in \mathbb{N}_{>0}$ , functions  $\kappa: \mathcal{S} \rightarrow [0, \infty)$ ,  $\eta: \Theta \rightarrow \mathbb{R}^{\ell}$ ,  $\tau: \mathcal{S} \rightarrow \mathbb{R}^{\ell}$ , and, for all states  $a \in \mathcal{S}$  and all parameters  $\theta \in \Theta$ ,

$$\mu_{\theta}(a) = \frac{\kappa(a) \exp(\eta(\theta) \cdot \tau(a))}{\sum_{b \in \mathcal{S}} \kappa(b) \exp(\eta(\theta) \cdot \tau(b))}. \tag{2.3}$$

---

<sup>6</sup>Throughout this paper, when we say *Markov chain*, we mean a *time-homogeneous* Markov chain, one whose transition probabilities are not functions of time.

Equation (2.3) actually defines the probability density or mass function  $\mu_\theta$  of the distribution that we are also calling  $\mu_\theta$ . Normally this ambiguity is no bother, but when it matters in subsection 2.3.2, eq. (2.3) will be understood to define densities with respect to some dominating,  $\sigma$ -finite measure on  $\mathcal{S}$ , typically the counting measure (E. L. Lehmann & Casella, 1998, pp. 23–24). In fact, every measure on a discrete space is absolutely continuous with respect to the counting measure, so we may always assume, without loss of generality, that the dominating measure is the counting measure and eq. (2.3) defines a classical probability mass function. See lemma 2.3.2.

Instead of saying that  $\mu$  is an exponential family, we might say that  $\mu_\theta$  has the *exponential family representation* in eq. (2.3) for all  $\theta \in \Theta$  or that  $\mu_\theta$  is *drawn from the exponential family* in eq. (2.3) for all  $\theta \in \Theta$ .

We call  $\kappa$  the *carrier measure*,  $\eta$  the *parameter function*.  $\tau$  is a sufficient statistic for  $\theta$  by the factorization theorem (Casella & Berger, 2002, Thms. 6.2.6 and 6.2.10).

Define  $\zeta: \Theta \rightarrow \mathbb{R} \cup \{\infty\}$  by

$$\zeta(\theta) = \log \left( \sum_{b \in \mathcal{S}} \kappa(b) e^{\eta(\theta) \cdot \tau(b)} \right), \quad (2.4)$$

so we can rewrite

$$\mu_\theta(a) = \kappa(a) e^{\eta(\theta) \cdot \tau(a) - \zeta(\theta)}. \quad (2.5)$$

We call  $\zeta$  the *log-partition function* or *log normalizer* (And thus might call  $e^\zeta$  the *normalizer* or *partition function*. See Wainwright & Jordan, 2008, Eq. 3.6; Nielsen & Garcia, 2011, § 1.2).

If  $d = \ell$  and  $\eta(\theta) = \theta$  for all  $\theta \in \Theta$ , we say that  $\mu$  is *naturally parameterized* with *natural parameter space*  $\mathcal{H} := \{\gamma \in \mathbb{R}^\ell \mid \sum_{b \in \mathcal{S}} \kappa(b) e^{\gamma \cdot \tau(b)} < \infty\}$  (Casella & Berger, 2002, § 3.4, p. 114; though some authors call parameterization with the natural parameter the *canonical form* and that the natural parameter space of eq. (2.3) is  $\eta^{-1}(\mathcal{H})$ . See, e.g., E. L. Lehmann & Casella, 1998, § 1.5, pp. 23–24; Wainwright & Jordan, 2008, § 3.2, p. 39). We always stipulate that  $\eta(\Theta)$  lies in the natural parameter space (Casella & Berger, 2002, p. 114); hence  $|\zeta(\theta)| < \infty$  for all  $\theta \in \Theta$ .  $\mu$  is called a *log-linear model* if it is naturally parameterized

and  $\kappa$  is constant (In this case, the elements of  $\mathcal{S}$  are called *cells* and  $\tau$  is called the *design matrix*. See Fienberg & Rinaldo, 2012a, § 2). Log linear families have contingency tables with exponential family distributions, and maximum likelihood estimation (MLE) for this class of models is well understood; see, e.g., Fienberg and Rinaldo (2012a).

The size and shape of  $\Theta$ ,  $\eta(\Theta)$ , and  $\tau(\mathcal{S})$  play an important role in exponential family theory. Since the codomain of  $\eta$  and  $\tau$  is  $\mathbb{R}^\ell$ , we say that  $\mu$  is  $\ell$  *dimensional* (E. L. Lehmann & Casella, 1998, § 1.5, p. 23). If  $d < \ell$  then  $\mu$  is *curved*; if  $d = \ell$  then  $\mu$  is *full* (Casella & Berger, 2002, § 3.4, Def. 3.4.7, p. 115). An exponential family is *regular* if its natural parameter space is open. Equation (2.3) is a *minimal representation* of  $\mu$  if the entries of  $\tau$  and of  $\eta$  are *affinely independent*, meaning  $\gamma \cdot \tau(a) = g$  for all  $a \in \mathcal{S}$  with  $\kappa(a) \neq 0$  implies that  $\gamma = \mathbf{0}$  and  $g = 0$  and  $\delta \cdot \eta(\theta) = h$  for all  $\theta \in \Theta$  implies  $\delta = \mathbf{0}$  and  $h = 0$  (Barndorff-Nielsen, 1978, Cor. 8.1, p. 113; Küchler & Sørensen, 1997, p. 38; Wainwright & Jordan, 2008, p. 40). If  $\eta$  and  $\tau$  are affinely independent, then  $\theta$  is identifiable (Wainwright & Jordan, 2008, p. 40). If, in addition to affine independence,  $\mathcal{H}$  contains an open,  $\ell$ -dimensional rectangle, then we say that  $\mu$  is *full rank* (E. L. Lehmann & Casella, 1998, § 1.5, p. 24). A curved exponential family is not full rank (E. L. Lehmann & Casella, 1998, § 1.5, p. 26). If eq. (2.3) is not a minimal representation, then it is an *overcomplete representation* (Wainwright & Jordan, 2008, § 3.2, p. 40).

For a rigorous introduction to exponential families, see Barndorff-Nielsen (1978, chap. 8). Küchler and Sørensen (1997) rigorously treats exponential families for stochastic processes, including Markov chains.

From eq. (2.2) we see that the joint probability of the Markov chain  $X$  through time  $t$  conditional on the initial state  $X_0 = x_0$  is drawn from an exponential family. Supposing that  $P_\theta = \theta =: P$  in eq. (2.2), we can take the parameter space  $\Theta$  to be the set of  $\mathcal{S} \times \mathcal{S}$  stochastic matrices. Thus  $N_t$  is a sufficient statistic for the transition matrix  $P$ . Moreover, the (consistent, asymptotically normal) MLE  $\hat{P}$  for  $P$  is (Asymptotic normality stands even when  $P_\theta$  is not simply  $\theta$ , but does require some additional hypotheses. See Stefanov, 1995,

§ 2; see also Al-Eideh et al., 1988; Gani, 1955)

$$\hat{P}_{ab} = \frac{N_t(a, b)}{N_t(a, +)}. \quad (2.6)$$

Because of the constraint that  $\sum_{b \in \mathcal{S}} P_{ab} = 1$  for all  $a \in \mathcal{S}$  in any stochastic matrix  $p$ , the minimal representation of eq. (2.2) actually requires that  $\Theta$  have dimension  $|\mathcal{S}|^2 - |\mathcal{S}| - k$ , where  $k \in \{0, 1, \dots, |\mathcal{S}|^2 - |\mathcal{S}|\}$  is the number of entries of  $p$  that equal zero (Stefanov, 1991, § 2; however, the dimension can be reduced by another multiple of  $|\mathcal{S}|$  if we take  $t$  to be a random stopping time in terms of  $N_t$ . See Stefanov, 1991; and for more on this topic, see Stefanov, 1995, Eq. 8, but we will not address stopping times further in this article). (We have to be careful about zero entries because we take logarithms in eq. (2.2).) The upshot is that the dimension of the sufficient statistic  $\mathcal{O}(|\mathcal{S}|^2)$ .

We review the Darmois-Koopman-Pitman theorems available for discrete state spaces.

**Theorem 2.2.1** (Darmois-Koopman-Pitman on Discrete State Spaces). *Let  $\mu = \{\mu_\theta \mid \theta \in \Theta\}$  be a family of probability distributions on some discrete space  $\mathcal{S}$  and the elements of  $\Theta$  have dimension  $d$ .*

**Diaconis and Freedman (1981)** *Suppose  $\mathcal{S}$  is the integers and the distributions of  $\mu$  have common support. If, for each  $t \in \mathbb{N}_{>0}$ , the sum of  $t$  IID random variables distributed according to  $\mu_\theta$  is sufficient for  $\theta$ , then  $\mu$  is an exponential family.*

**Denny (1972)** *Suppose  $\mathcal{S}$  is countably infinite. If  $\mu$  is not an exponential family,  $\mathcal{S}$  has an infinite subset  $A$  such that*

$$\tau(x_1, \dots, x_t) = \tau(y_1, \dots, y_t) \implies (x_1, \dots, x_t) = (y_{\pi(1)}, \dots, y_{\pi(t)})$$

*for any  $(x_1, \dots, x_t), (y_1, \dots, y_t) \in A$  and for at least one permutation  $\pi$  of  $[t]$ .*

**Andersen (1970)** *Suppose  $\mathcal{S} = \{1, \dots, s\}$ ,  $\mu_\theta(a) > 0$  for all  $\theta \in \Theta$  and all  $a \in \mathcal{S}$ , and  $d = 1$ . Further, suppose that if  $j = 1, 2$ ,  $q = \tau(x_1, \dots, x_t)$ ,  $q_1 = \tau(x_1, \dots, x_{j-1}, x'_j, x_{j+1}, \dots, x_t)$ , and  $q_2 = \tau(x_1, \dots, x_{j-1}, x''_j, x_{j+1}, \dots, x_t)$ , then  $q_1 < q < q_2$  implies  $x_j$  exists with either  $x'_j \leq x_j \leq x''_j$  or  $x''_j \leq x_j \leq x'_j$ . Then  $\mu$  is an exponential family with sufficient statistic  $\tau$  of dimension one.*

**Gani (1955)** Suppose  $\mathcal{S}$  is finite and  $d = 1$ . If  $\tau$  is real valued and is formally differentiable<sup>7</sup> with respect to the entries of the contingency table  $\mathbf{N}_t(+, \cdot)$ , then  $\mu$  is an exponential family whose sufficient statistic is a function of  $\tau$ .<sup>8</sup>

### 2.3 Conditional Exponential Families

A *conditional exponential family* (Feigin, 1981, Def. A) (CEF) is a family of transition matrices  $P := \{P_\theta \mid \theta \in \Theta\}$  such that each row of  $P_\theta$  is an exponential family with the same parameter function  $\eta$  and same parameter space  $\Theta$ , i.e.,

$$P_\theta(a, b) = \kappa(a, b)e^{\eta(\theta) \cdot \tau(a, b) - \zeta(a, \theta)}. \quad (2.7)$$

The *natural parameter space* for this CEF is the set of all  $\gamma \in \mathbb{R}^\ell$  such that

$$\sum_{b \in \mathcal{S}} \kappa(a, b)e^{\gamma \cdot \tau(a, b)}$$

is finite for all  $a \in \mathcal{S}$ . We always stipulate that  $\eta(\Theta)$  lies in the natural parameter space (Feigin, 1981, p. 598); hence  $0 < e^{\zeta(a, \theta)} < \infty$  for all  $\theta \in \Theta$  and all  $a \in \mathcal{S}$ .

The CEF is a *conditionally additive exponential family* (CAEF) if  $\zeta(a, \theta) = \psi(a)\phi(\theta)$  for some functions  $\phi: \Theta \rightarrow \mathbb{R}$  and  $\psi: \mathcal{S} \rightarrow \mathbb{R}$  (The original definition required that  $\psi$  were such that the range of  $\psi$  contains either one or an interval  $(0, \delta)$  for some  $\delta > 0$ . In our context,  $\mathcal{S}$  is a discrete space, so  $\psi(\mathcal{S})$  cannot contain an interval. We can make  $1 \in \psi(\mathcal{S})$  by rescaling  $\phi$ , so we have not included these constraints in our definition. The original definition appears at Feigin, 1981, Def. B). Finally, we say that a CAEF is a *Markovian exponential family* (MEF) if  $\psi(a)$  is a non-zero constant for all  $a \in \mathcal{S}$ , i.e.,  $\zeta(a, \theta) = \zeta(b, \theta) =: \zeta(\theta)$  so that (cf. eq. (2.9) and Hudson, 1982, Eq. 2.1)

$$P_\theta(a, b) = \kappa(a, b)e^{\eta(\theta) \cdot \tau(a, b) - \zeta(\theta)}. \quad (2.8)$$

**2.3.1 Literature Review.** Suppose  $X = \{X_t\}_{t \in \mathbb{N}}$  is a Markov chain whose transition matrix is drawn from a family  $P = \{P_\theta \mid \theta \in \Theta\}$ . Gani (1955, 1956) showed that when  $\mathcal{S}$  is finite,

<sup>7</sup>↑Gani (1955) concludes using this assumption that  $\tau$  is a function of a linear combination of the entries of the contingency table.

<sup>8</sup>↑Gani (1955) additionally requires that  $\mu_\theta$  be differentiable with respect to  $\theta$ . This assumption can be dropped by applying the factorization theorem to  $dL_t(\theta)/dL_t(\theta_0)$  for some fixed  $\theta_0 \in \Theta$  rather than to  $L_t(\theta)$ .



$\Theta$  is a scalar set, and every transition matrix of  $P$  has a stationary distribution<sup>9</sup> and is differentiable with respect to  $\theta$ ,<sup>10</sup> for  $L_{\theta, x_0}^t(x_1, \dots, x_t)$  (defined in eq. (2.2)) to have a one-dimensional sufficient statistic for  $\theta$  required that  $P$  be an MEF. This is also true of chains with a single absorbing state and a random stopping time. Bhat and Gani (1960) extends this result to the case where the initial state  $X_0$  is also random with a distribution depending on the same parameter. Bofinger (1965) gives an analogous result for time-inhomogeneous Markov chains (but the result is closer to CEFS than MEFS(Bofinger, 1965, Eq. 7)). Adke and Swamy (1979) and Mitrofanova (1968/1971) extend the time-homogeneous case to continuous state spaces.

Apparently unaware of Gani (1955), Heyde and Feigin (1974) proposed CAEFS (which they called conditional exponential families (Feigin, 1981, p. 599)) when trying to define asymptotic efficiency for estimation when observing from dependent sequences of random variables. Their definition was in terms of the score function induced by the transition densities. Feigin (1981) introduced the more general class of CEFS with vector parameters over  $\mathbb{R}^p$  and studied their statistical properties. The functional form of MEFS with a scalar parameter first appears in Gani (1955). Adke and Swamy (1979), Bhat (1988), Bhat and Gani (1960), Bofinger (1965), Gani (1956), and Mitrofanova (1968/1971) all cite Gani (1955), but Feigin (1981), Heyde and Feigin (1974), and K uchler and S orensen (1997, 1998) do not.<sup>11</sup> Hudson (1982) introduced the name *Markovian exponential families*(Hudson, 1982, p. 88) and vector parameters, but with the particular form (We take the liberty of replacing some of the letters with those matching eq. (2.8) compared to Hudson, 1982, Eq. 2.1)

$$\kappa(x, y) \exp(\alpha(\theta) \cdot m(x, y) - \gamma(\theta)T(x) - \zeta(\theta)) \quad (2.9)$$

for some nonnegative function  $T$ . The author concludes that the presence of the  $T$  term

---

<sup>9</sup>↑Gani (1956) assumed that the Markov chain had a single, irreducible, closed subset of states. A Markov chain on a finite state space with a single, irreducible closed subset of states has a unique stationary distribution (Hoel et al., 1972, chap. 2).

<sup>10</sup>↑As in footnote 8, differentiability with respect to  $\theta$  is not essential.

<sup>11</sup>↑Though K uchler and S orensen (1997, p. 78) and K uchler and S orensen (1998, p. 4) cite each of Bhat (1988), Feigin (1981), and Heyde and Feigin (1974). On February 19, 2018, Google Scholar said that Gani (1955) has been cited 27 times.

prevents the joint distribution of  $t$  observations from being an exponential family. According to our theorem 2.3.11, this was a hasty conclusion: Setting  $\eta(\theta)$  equal to the vector  $(\alpha(\theta), -\gamma(\theta))$  and  $\tau(x, y)$  equal to the vector  $(m(x, y), T(x))$  would turn eq. (2.9) into an MEF, thereby satisfying theorem 2.3.11.

Our interest here is in the structure of CEFS and MEFS more so than in the statistical properties they have. Partially this is because our interest in this problem arose in model selection for Markov chains of networks; a better understanding of the structure of MEFS allows us in section 2.5 to pick models of Markov chains of networks that are interpretable and have sufficient statistics of small, constant dimension à la theorem 2.2.1.

Another reason to focus on structure over statistics is that the literature so far has mostly focused on statistics. Gani (1955) proves many results about MLE for general discrete time and space Markov chains. Bhat and Gani (1960) discuss the unbiasedness and minimum variance properties in estimating a one-parameter MEF with random initial state. Adke and Swamy (1979) and Mitrofanova (1968/1971) provide statistical results for MEFS defined on continuous state spaces. Heyde and Feigin (1974) proposes efficiency for stochastic processes with a scalar parameter, and shows that MLE for CAEFS (which Heyde and Feigin call *conditional exponential families*) is efficient and strongly consistent under certain conditions. The authors also derive the Fisher information for single-parameter CAEFS.

Feigin (1981) extend facts about the Fisher information matrix from exponential families to CEFS, showing, for example, that the Fisher information matrix for the likelihood function of  $t$  observations from a CEF is the Hessian matrix of  $\sum_{i=0}^{t-1} \zeta(\theta, X_i)$ , and that this forms a zero-mean,  $L^2(p(\theta))$  martingale (Feigin, 1981, Thm. 1). In the scalar case,  $\prod_{i=0}^{t-1} \tau(X_i, X_{i+1})/\zeta'(\theta, X_i)$  is also a martingale, which converges almost surely to some function of  $\theta$  (Feigin, 1981, Eqs. 2.4–5). In the CAEF (rather than CEF) case, the authors derive an explicit expression for the score function and use it to give conditions for the existence of the MLE. Further, when  $\tau$  is invertible in the second slot, under certain regularity conditions, MLE for CAEFS is strongly consistent and asymptotically normal (Feigin, 1981, Thm. 4). Hudson (1982) proves that MEFS are locally asymptotically mixed normal, including a central

limit theorem for the score function an asymptotic for the likelihood ratio statistic under regularity conditions. Bhat (1988), which provides a concise summary of major results about exponential families in general. Further, the author shows that the statistics of naturally scalar parameterized MEFS (he does not use this terminology) on  $\mathcal{S}$  are the same as those of IID sequences on  $\mathcal{S}^2$  (the transitions  $(X_i, X_{i+1})$  with an exponential family distribution. He does this by comparing the characteristic function  $\exp[n(\zeta(\theta + iu) - \zeta(\theta))]$  of a naturally scalar parameterized exponential family with the same sufficient statistic over  $\mathcal{S}$  to the characteristic function of eq. (2.17). The author points out that  $\sum_{i=1}^n \tau(X_i)/n$  is the uniformly minimum variance unbiased estimator of  $\mathbb{E}(\tau(X))$ , and discusses some sequential tests. Hwang and Basawa (1994) proves asymptotic results for CEFS, including local asymptotic normality, that establish the optimality of some classical statistical tests for hypotheses about the parameters. Stefanov (1984, 1995) discuss sequential estimation of general, discrete Markov chains. Lindsey (2004, § 8.2) merely describes the real-valued CEFS and gives a couple of examples of autoregressive generalized linear models that are CEFS. Sharia (2007, 2010) introduce (and Zhong (2015, § 3.2.2) also reports) a recursive estimator of naturally parameterized CEFS, but the algorithm requires the entire sequence of observations; the corresponding recursive algorithm for CAEFS requires the entire sequence of observed sufficient statistics  $\{(\tau(X_i, X_{i+1}), \psi(X_i)) \mid i \leq t\}$  (Sharia, 2010, Eq. 4.7). Under regularity conditions, the estimator is strongly consistent (Sharia, 2010, Prop. 4.3) and asymptotically linear (Sharia, 2010, Cor. A1, app. A). Al-Eideh et al. (1988) proves consistency of MLE for discrete Markov chains assuming only twice continuous differentiability in the parameter and bounded third derivatives.

Some of the literature has discussed the structure of MEFS and CEFS, starting with discussions in Bhat and Gani (1960) and Gani (1955) about the range of possible values of  $\tau$  and  $\kappa$ . Bhat (1988) provided an example from that theory. We will discuss it in more depth below. Küchler (1982) and Küchler and Sørensen (1997, 1998) discuss structure of MEFS' joint distributions with a focus on continuous time and continuous space. In particular, Küchler and Sørensen (1997) is a book-length treatment of stochastic processes whose likelihood functions are drawn from exponential families. (Küchler and Sørensen (1998) is largely

an article length republication of chapter six from Küchler and Sørensen (1997), which is specifically about Markov chains.) Nagaoka (2005) Considers the information geometry of CAEFS and gives a theorem characterizing the set of all MEFS in terms of the affine geometry of the exponents  $\theta \cdot \tau_a(b)$ . Sections 4 and 5 translate basic concepts from usual exponential families to MEFS. Finally it explores the information geometry of MEFS. This paper in turn inspired Hayashi and Watanabe (2016) to derive MEFS (That the topic is the same functional form as eq. (2.8) is opaque, but see Hayashi & Watanabe, 2016, Eq. 4.5) from an information geometry point of view, provide several information geometrical theorems characterizing MEFS, and demonstrate asymptotic efficiency of the MLE.

**2.3.2 Characterizing Exponential Families of Markov Chains.** The main theorem of this subsection, theorem 2.3.11 on page 29, connects CEFS, MEFS, and exponential families: a Markov chain's probability mass function has an exponential family representation for every length of finite sample if and only if the chain is an MEF. In this sense, we might say that a CEF is an exponential family if and only if it is an MEF. From a slightly different perspective, conditions under which a stochastic process whose joint distribution is an exponential family also has marginal distributions from an exponential family have been studied for continuous-time processes (Ycart, 1988, 1989, 1992a, 1992b). Küchler and Sørensen (1997, § 12.1) summarizes these results succinctly. Our goal, however, is to find the class of Markov chains whose transitions are exponential families that also have joint probabilities from exponential families.

The first part of the theorem relies on Küchler and Sørensen (1997, Cor. 6.3.4) and Küchler and Sørensen (1998, Cor. 4.10), which requires much more measure-theoretically technical definitions than we have had so far. Even though theorem 2.3.11 is the main theorem of section 2.3, we have relegated it to this subsection because we will not need all the measure theory afterward. Much of the notation and many of the definitions below follow Küchler and Sørensen (1997, chaps. 3, 6) very closely. Küchler and Sørensen (1998), which is nearly but not exactly the same, also has many of these definitions and similar notation. We will cite only the former except where the latter does something slightly different that is more convenient. We assume the reader is familiar with  $\sigma$ -algebras, measurability, and the

Radon-Nikodým theorem. See Shiryaev (2016, chap. 2, § 6.7, p. 233) for a detailed review of the Radon-Nikodým theorem and Radon-Nikodým densities.

**2.3.2.1 Preparatory Definitions.** Represent time with the set  $\mathcal{T}$ , which is either  $\mathbb{N}$  or  $[0, \infty)$ . Let  $(\Omega, \mathcal{F})$  be a measurable space. Let the parameter space  $\Theta \subseteq \mathbb{R}^d$  be non-empty. Let  $(\mathcal{S}, \mathcal{S})$  be any measurable state space. When  $\mathcal{S}$  is countable, we will always assume it is endowed with the *discrete  $\sigma$ -algebra*  $\mathcal{S} = 2^{\mathcal{S}}$ , and likewise  $\mathcal{S}^t$  will be endowed with its discrete  $\sigma$ -algebra  $2^{\mathcal{S}^t}$ , which is also the product  $\sigma$ -algebra for  $t$  copies of  $\mathcal{S}$  (Tao, 2011, Exercise 1.7.19(viii), p. 162). When we require measurability of a function whose codomain is a subset of any topological space, we always assume the appropriate Borel  $\sigma$ -algebra.

Let  $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$  be a *filtration*, which is a family of  $\sigma$ -algebras satisfying  $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$  for all  $s \leq t \in \mathcal{T}$  (Øksendal, 2003, Def. 3.2.2, p. 31). If  $\kappa$  is any nonnegative measure on  $(\Omega, \mathcal{F})$ , define  $\kappa^t$ , for each  $t \in \mathcal{T}$ , to be the *restriction* of  $\kappa$  to the  $\sigma$ -algebra  $\mathcal{F}_t$ , i.e.,  $\kappa^t: \mathcal{F}_t \rightarrow [0, \infty]$  such that  $\kappa^t(A) := \kappa(A)$  for any  $A \in \mathcal{F}_t$ . If  $P = \{P_\theta\}_{\theta \in \Theta}$  is a set of probability measures on  $(\Omega, \mathcal{F})$ , then let  $P^t := \{P_\theta^t\}_{\theta \in \Theta}$ .

Suppose  $\mu$  and  $\nu$  are  $\sigma$ -finite measures on the same space  $(\Omega, \mathcal{F})$ . We say that  $\mu$  is *absolutely continuous* with respect to  $\nu$ , written  $\mu \ll \nu$ , if  $E \in \mathcal{F}$  and  $\nu(E) = 0$  implies  $\mu(E) = 0$  (Jacod & Protter, 2004, Def. 28.1, p. 244; Shiryaev, 2016, chap. 2, § 6.7, p. 232). If additionally  $\nu \ll \mu$  then we say that  $\mu$  and  $\nu$  are *equivalent* and write  $\mu \sim \nu$  (Jacod & Protter, 2004, Exercise 28.1, p. 247). Absolute continuity is transitive.

An  $\mathcal{S}$ -valued *stochastic process*  $X$  is a function  $X: \mathcal{T} \times \Omega \rightarrow \mathcal{S}$  such that  $X_t$  is measurable with respect to  $\mathcal{F}$  and  $\mathcal{S}$  for each  $t \in \mathcal{T}$ .  $X$  is *adapted* to  $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$  if  $X_t$  is measurable with respect to  $\mathcal{F}_t$  and  $\mathcal{S}$  for each  $t \in \mathcal{T}$  (Øksendal, 2003, Def. 3.1.3, p. 25). For each  $t \in \mathcal{T}$ ,  $\sigma(X_t)$  denotes the smallest  $\sigma$ -algebra containing all the subsets of  $\mathcal{S}$  that are elements of  $X_t^{-1}(\mathcal{S}) := \{X_t^{-1}(A) \mid A \in \mathcal{S}\}$ , which is always a sub- $\sigma$  algebra of  $\mathcal{F}$  (Jacod & Protter, 2004, chap. 10, p. 65). If  $\mathcal{F}_t = \sigma(X_s \mid s \leq t)$  for all  $t \in \mathcal{T}$  then we say that  $X$  *generates* the filtration  $\{\mathcal{F}_t\}_t$ , which in turn is the *filtration of the process*  $X$  (Øksendal, 2003, chap. 3, Exercise 3.3, p. 38).

We will rely on lemma 2.3.1 to connect measurability and functional relationships. A set is *Polish* if it's a complete, separable metric space (Shiryaev, 2016, p. 183). A  $\sigma$ -algebra  $\mathcal{A}$

*separates points* if  $\mathbb{1}(\omega \in A) = \mathbb{1}(\xi \in A)$  for all  $A \in \mathcal{A}$  implies  $\omega = \xi$  (Hoffmann-Jørgensen, 1994, chap. 6, § 2, p. 442).

**Lemma 2.3.1** (Doob-Dynkin (Hoffmann-Jørgensen, 1994, chap. 6, § 4, pp. 443–444; the name comes from the  $\mathbb{R}^n$ -valued case. See Øksendal, 2003, Lem. 2.1.2)). *Let  $(A, \mathcal{A})$  and  $(B, \mathcal{B})$  be measurable spaces,  $Y: \Omega \rightarrow (A, \mathcal{A})$  and  $Z: \Omega \rightarrow (B, \mathcal{B})$ . Suppose either that  $A$  is Polish and  $\mathcal{A}$  is Borel  $\sigma$ -algebra of  $A$ , or that  $Z(\Omega) \in \mathcal{B}$  and  $\mathcal{A}$  separates points. Then  $\sigma(Y) \subseteq \sigma(Z)$  (i.e.,  $Y$  is measurable with respect to  $\sigma(Z)$ ) if and only if there exists a measurable function  $f: (B, \mathcal{B}) \rightarrow (A, \mathcal{A})$  such that  $Y = f \circ Z$ .*

Lemma 2.3.1 is true for  $A = \mathbb{R}^n$  with the Borel  $\sigma$ -algebra (Jacod & Protter, 2004, Thm. 23.2). The lemma is also true whenever  $(A, \mathcal{A}) = (C, 2^C)$  for any at-most countable set  $C$ . An at-most countable set  $C$  is Polish when endowed with the *discrete topology*  $2^C$  in which every set is open (J. K. Hunter & Nachtergaele, 2005, chap. 4, p. 82), which it can be via the *discrete metric*  $d(x, y) = 1 \iff x \neq y$  (Reimann, 2011, p. 2).  $2^C$  is also the  $\sigma$ -algebra generated by the singletons of  $C$  (Chung, 1968/2000, chap. 2, § 1, Exercise 4, p. 20). This is compatible with our assumption of using the discrete  $\sigma$ -algebra on countable state spaces because the Borel  $\sigma$ -algebra for the discrete topology is the discrete  $\sigma$ -algebra.

The measurable spaces  $(A, \mathcal{A})$  for which lemma 2.3.1 is true regardless of  $(B, \mathcal{B})$  have been completely characterized (Pratelli, 1990, Thm. 3.1). All measurable Lusin spaces satisfy the conditions (Pratelli, 1990, Thm. 3.2). All Polish spaces are Lusin spaces (Takesaki, 1979, p. 379).  $T_1$  topological spaces also make the lemma true (Taraldsen, 2018, Lem. 2).

The following lemma is well known but rarely written down explicitly (We could not find it in Chung, 1968/2000; Jacod & Protter, 2004; Shiryaev, 2016; Tao, 2011).

**Lemma 2.3.2.** *A Radon-Nikodým density on a discrete state space is the simple ratio of the probabilities. More concretely, if  $\mu$  and  $\nu$  are  $\sigma$ -finite measures on an at-most countable space  $\mathcal{S}$  and  $\mu \ll \nu$ , then  $\frac{d\mu}{d\nu}(a) = \mu(\{a\})/\nu(\{a\})$  for all  $a \in \mathcal{S}$  such that  $\nu(\{a\}) \neq 0$ . In particular, if  $\nu$  is the counting measure, then  $\frac{d\mu}{d\nu}(a) = \mu(\{a\})$ .*

*Proof.* By Radon-Nikodým theorem (Shiryaev, 2016, chap. 2, § 6.7, p. 233),  $d\mu/d\nu$  is the  $\nu$ -almost-surely unique,  $[0, \infty]$ -valued random variable on  $\mathcal{S}$  such that  $\mu(A) = \int_A \frac{d\mu}{d\nu}(a)\nu(da)$

for all  $A \in \mathcal{S}$ . Since  $\mathcal{S}$  is at most countable,  $\mathcal{S} = 2^{\mathcal{S}}$ . Since  $\mu \ll \nu$ , for any  $A \subseteq \mathcal{S}$  and any  $a \in A$ , we have exactly one of

1.  $\mu(\{a\}) \neq 0$  and  $\nu(\{a\}) \neq 0$ , or
2.  $\mu(\{a\}) = 0$  and  $\nu(\{a\}) \neq 0$ , or
3.  $\mu(\{a\}) = 0$  and  $\nu(\{a\}) = 0$ .

Applying this trichotomy to get the last equality in the first row, we have, for any  $A \subseteq \mathcal{S}$ ,

$$\begin{aligned} \int_A \frac{d\mu}{d\nu}(a) \nu(da) &= \mu(A) = \sum_{a \in A} \mu(\{a\}) = \sum_{\substack{a \in A \\ \mu(\{a\}) \neq 0}} \mu(\{a\}) = \sum_{\substack{a \in A \\ \nu(\{a\}) \neq 0}} \mu(\{a\}) \\ &= \sum_{\substack{a \in A \\ \nu(\{a\}) \neq 0}} \frac{\mu(\{a\})}{\nu(\{a\})} \nu(\{a\}) = \int_A \frac{\mu(\{a\})}{\nu(\{a\})} \nu(da). \end{aligned}$$

We conclude from the density's  $\nu$ -almost-sure uniqueness that  $\frac{d\mu}{d\nu}(a) = \mu(\{a\})/\nu(\{a\})$  for all  $a \in \mathcal{S}$ .

When  $\nu$  is the counting measure,  $\nu(\{a\}) = 1$  for all  $a \in \mathcal{S}$ , so the last statement of the lemma follows from the previous one.  $\square$

**2.3.2.2 Exponential Families of a Stochastic Process.** Let  $P = \{P_\theta\}_{\theta \in \Theta}$  be a set of probability measures on  $(\Omega, \mathcal{F})$ . Following Küchler and Sørensen (1997, chap. 3) closely, we say that  $P$  is an *exponential family with respect to the filtration* (Küchler & Sørensen, 1997, § 3.1, p. 19)  $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$  if there exist

- a  $\sigma$ -finite measure  $\mu$  on  $(\Omega, \mathcal{F})$  such that for all  $\theta \in \Theta$  and  $t \in \mathcal{T}$ ,  $P_\theta^t \ll \mu^t$ ;
- non-random functions  $\eta: \Theta \rightarrow \mathbb{R}^\ell$  and  $a_t: \Theta \rightarrow (0, \infty)$  for each  $t \in \mathcal{T}$ ;<sup>12</sup>
- a  $[0, \infty)$ -valued stochastic process  $q$  and an  $\mathbb{R}^\ell$ -valued process  $B$ , called a *canonical process*, both of which are  $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$ -adapted;<sup>13</sup>

<sup>12</sup>↑ While Küchler and Sørensen (1997) requires that  $a$  have right continuity and left limits, Küchler and Sørensen (1998) does not require the same of  $-\log a$ . We stick with the simpler definition here. That  $\eta$  does not depend on time  $t$  renders  $P$  *time homogeneous*.

<sup>13</sup>↑ While Küchler and Sørensen (1997) requires that  $q$  and  $B$  have right continuity and left limits, Küchler and Sørensen (1998) does not. We stick with the simpler definition here.

such that for all  $t \in \mathcal{T}$  and  $\theta \in \Theta$ , the Radon-Nikodým density of  $P_\theta^t$  with respect to  $\mu^t$  is

$$\frac{dP_\theta^t}{d\mu^t} = a_t(\theta)q_t \exp(\eta(\theta) \cdot B_t). \quad (2.10)$$

If  $X = \{X_t\}_{t \in \mathcal{T}}$  is a stochastic process and generates  $\{\mathcal{F}_t\}_t$ , then and  $P$  is an *exponential family of the stochastic process*  $X$  (Küchler & Sørensen, 1997, § 3.1, p. 21, requires that the filtration be right continuous by replacing  $\mathcal{F}_t$  with  $\mathcal{F}_{t+} := \bigcap_{s>t} \mathcal{F}_s$  for each  $t \in \mathcal{T}$ ; however Küchler & Sørensen, 1998, Def. 2.1, p. 6, does not. We stick with simpler definition here. ).

Our goal for the remainder of this sub-subsection is to prove lemma 2.3.5 on the next page, which characterizes the relationship between exponential families and exponential families for a stochastic process in discrete time and space. For the remainder of sub-subsection 2.3.2.2 we assume the following.

**Assumption 2.3.3.** Suppose that  $\mathcal{S}$  is at most countable and  $\mathcal{T} = \mathbb{N}$ . Let  $X = \{X_t\}_{t \in \mathbb{N}}$  be an  $\mathcal{S}$ -valued stochastic process on  $(\Omega, \mathcal{F})$  generating the filtration  $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$ . For any  $t \in \mathbb{N}$ , define  $\mathbf{X}_t := (X_0, \dots, X_t)$ , an  $\mathcal{S}^{t+1}$ -valued random variable. Assume  $\mathcal{F} = \sigma(X)$ .

Keep in mind throughout that because  $\mathcal{S}^t$  is at most countable for each  $t \in \mathbb{N}$ , its  $\sigma$ -algebra is  $2^{\mathcal{S}^t}$ . This means that, for all  $t \in \mathbb{N}$ ,  $\mathbf{X}_t$  generates the  $\sigma$ -algebra

$$\mathcal{F}_t = \sigma(\mathbf{X}_t) = \mathbf{X}_t^{-1}(2^{\mathcal{S}^{t+1}}) = \{\mathbf{X}_t^{-1}(A) \mid A \subseteq \mathcal{S}^{t+1}\}$$

where, for all  $A \subseteq \mathcal{S}^{s+1}$ ,  $A$  is at most countable, so

$$\mathbf{X}_s^{-1}(A) = \{\mathbf{X}_t \in A\} = \bigcup_{\mathbf{x} \in A} \mathbf{X}_t^{-1}(\mathbf{x}) = \bigcup_{\mathbf{x} \in A} \{\mathbf{X}_t = \mathbf{x}\}.$$

Since the smallest  $\sigma$ -algebra containing each event of the form  $\{\mathbf{X}_t = \mathbf{x}\}$ ,  $\mathbf{x} \in \mathcal{S}^{t+1}$ , contains all countable unions of them and they partition  $\Omega$ , they generate  $\mathcal{F}_t$ . To summarize, we have the following lemma.

**Lemma 2.3.4.** *Under assumption 2.3.3,*

$$\mathcal{F}_t = \left\{ \bigcup_{\mathbf{x} \in A} \{\mathbf{X}_t = \mathbf{x}\} \mid A \subseteq \mathcal{S}^{t+1} \right\} = \sigma\left(\{\{\mathbf{X}_t = \mathbf{x}\} \mid \mathbf{x} \in \mathcal{S}^{t+1}\}\right).$$



By lemma 2.3.4, the class of events  $\{\{\mathbf{X}_t = \mathbf{x}\} \mid \mathbf{x} \in \mathcal{S}^{t+1}\} \cup \{\emptyset\}$  generates  $\mathcal{F}_t$ . That class is closed under finite intersections, all such intersections just being the empty set. By Jacod and Protter (2004, Cor. 6.1, p. 36) and the fact that the measure of  $\emptyset$  is always zero, if  $Q^t$  and  $R^t$  are two probability measures on  $(\Omega, \mathcal{F}_t)$  such that  $Q^t(\mathbf{X}_t = \mathbf{x}) = R^t(\mathbf{X}_t = \mathbf{x})$ , then  $Q^t = R^t$ . In other words, the singletons on  $\mathcal{S}^{t+1}$  uniquely characterize  $Q^t$ .

We may think of the stochastic process  $X$  with state space  $\mathcal{S}$  as a random variable with state space  $\mathcal{S}^{\mathbb{N}}$ , the set of all  $\mathcal{S}$ -valued sequences (Øksendal, 2003, p. 11). The  $\sigma$ -algebra  $\mathcal{S}_{\mathbb{N}} = \bigotimes_{i=0}^{\infty} 2^{\mathcal{S}}$  of  $\mathcal{S}^{\mathbb{N}}$  is generated by sets of the form  $\{\{x_i\}_{i=0}^{\infty} \in \mathcal{S}^{\mathbb{N}} \mid x_{t_1} \in A_1, \dots, x_{t_k} \in A_k\}$  for  $k, t_1, \dots, t_k \in \mathbb{N}$  and  $A_1, \dots, A_k \subseteq \mathcal{S}$  (Øksendal, 2003, p. 11; Jacod & Protter, 2004, p. 70; Shiryaev, 2016, chap. 2, § 2.8, pp. 182–183).  $\mathcal{S}_{\mathbb{N}}$  is the smallest  $\sigma$ -algebra containing  $\bigcup_{t=0}^{\infty} 2^{\mathcal{S}^t}$ . Likewise,  $\sigma(X)$  is generated by sets of the form

$$\{\omega \in \Omega \mid X_{t_1}(\omega) \in A_1, \dots, X_{t_k}(\omega) \in A_k\}$$

for  $k, t_1, \dots, t_k \in \mathbb{N}$  and  $A_1, \dots, A_k \subseteq \mathcal{S}$  (Øksendal, 2003, p. 11; Jacod & Protter, 2004, p. 70).  $\sigma(X)$  is the smallest  $\sigma$ -algebra containing  $\bigcup_{t=0}^{\infty} \mathcal{F}_t$ .

**Lemma 2.3.5.** *Under assumption 2.3.3, define  $L_{\theta}^t$  to be the law of  $\mathbf{X}_t$  under  $P_{\theta}^t$  for all  $t \in \mathbb{N}$  and  $\theta \in \Theta$ .  $P$  is an exponential family of the stochastic process  $X$  if and only if, for each  $t \in \mathbb{N}$ ,  $L^t := \{L_{\theta}^t\}_{\theta \in \Theta}$  is an exponential family on  $\mathcal{S}^{t+1}$  with the same parameter function  $\eta: \Theta \rightarrow \mathbb{R}^{\ell}$  for all  $t \in \mathbb{N}$  and obeying the **consistency condition** that  $L_{\theta}^{t+1}(A \times \mathcal{S}) = L_{\theta}^t(A)$  for all  $\theta \in \Theta$  and  $A \subseteq \mathcal{S}^{t+1}$ .*

*Proof.* Let  $\nu^t$  be the counting measure on  $\mathcal{S}^{t+1}$  for all  $t \in \mathbb{N}$ . For any  $t \in \mathbb{N}$ , the Radon-Nikodým theorem (Shiryaev, 2016, chap. 2, § 6.7, p. 233) and the definition of *law* of a random variable gives us that, for all  $A \subseteq \mathcal{S}^{t+1}$ ,

$$P_{\theta}^t(\mathbf{X}_t \in A) = L_{\theta}^t(A) = \int_A \frac{dL_{\theta}^t}{d\nu^t}(x) \nu^t(dx) = \sum_{\mathbf{x} \in A} \frac{dL_{\theta}^t}{d\nu^t}(\mathbf{x}),$$

and in particular,  $P_{\theta}^t(\mathbf{X}_t = \mathbf{x}) = L_{\theta}^t(\{\mathbf{x}\}) = (dL_{\theta}^t/d\nu^t)(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{S}^{t+1}$ . According to the discussion after lemma 2.3.4, the forms of  $L_{\theta}^t(\{\mathbf{x}\})$ ,  $(dL_{\theta}^t/d\nu^t)(\mathbf{x})$ , and  $P_{\theta}^t(\mathbf{X}_t = \mathbf{x})$  each uniquely determine all of  $L_{\theta}^t$ ,  $dL_{\theta}^t/d\nu^t$ , and  $P_{\theta}^t$ .

( $\implies$ ) (This direction of the lemma appears without proof in Küchler & Sørensen, 1997, p. 37). Suppose  $P$  is an exponential family of the stochastic process  $X$  such that eq. (2.10) is true. Fix an arbitrary time  $t \in \mathbb{N}$ .

Since  $q$  and  $B$  are adapted to the filtration  $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$  generated by  $X$ , lemma 2.3.1 creates functions  $k_t: \mathcal{S}^{t+1} \rightarrow R$  and  $\tau_t: \mathcal{S}^{t+1} \rightarrow \mathbb{R}^\ell$  such that  $q_t = k_t(\mathbf{X}_t)$  and  $B_t = \tau_t(\mathbf{X}_t)$ . For each  $\theta \in \Theta$ , let  $\zeta(\theta) := -\log a_t(\theta)$ . Applying the Radon-Nikodým theorem (Shiryaev, 2016, chap. 2, § 6.7, p. 233) to eq. (2.10) yields

$$\begin{aligned} P_\theta^t(\mathbf{X}_t = \mathbf{x}) &= \int_{\{\omega \in \Omega \mid \mathbf{X}_t(\omega) = \mathbf{x}\}} \frac{dP_\theta^t}{d\mu^t}(\omega) \mu^t(d\omega) \\ &= \int_{\{\omega \in \Omega \mid \mathbf{X}_t(\omega) = \mathbf{x}\}} a_t(\theta) q_t(\omega) \exp(\boldsymbol{\eta}(\theta) \cdot B_t(\omega)) \mu^t(d\omega) \\ &= \int_{\{\omega \in \Omega \mid \mathbf{X}_t(\omega) = \mathbf{x}\}} a_t(\theta) k_t(\mathbf{X}_t(\omega)) \exp(\boldsymbol{\eta}(\theta) \cdot \tau_t(\mathbf{X}_t(\omega))) \mu^t(d\omega) \\ &= a_t(\theta) k_t(\mathbf{x}) \exp(\boldsymbol{\eta}(\theta) \cdot \tau_t(\mathbf{x})) \mu^t(\mathbf{X}_t = \mathbf{x}) = \kappa(\mathbf{x}) \exp(\boldsymbol{\eta}(\theta) \cdot \tau_t(\mathbf{x}) - \zeta(\theta)), \end{aligned}$$

where we have put  $\kappa_t(\mathbf{x}) = k_t(\mathbf{x}) \mu^t(\mathbf{X}_t = \mathbf{x})$ . This satisfies eq. (2.5) when the state space is  $\mathcal{S}^{t+1}$  and the parameter function is  $\boldsymbol{\eta}$  regardless of  $t$ . The parameter space does not depend on  $t$  because  $P_\theta^t$  is defined for all  $\theta \in \Theta$ . The consistency condition follows from Shiryaev (2016, chap. 2, § 3.4, Eq. 16 and Remark 4, pp. 197–199).

( $\impliedby$ ) (For a discussion of the complications to this direction of the lemma that arise in more general settings, see Küchler & Sørensen, 1997, pp. 81–82). Suppose that for each  $t \in \mathbb{N}$ , there exist functions  $\tau_t: \mathcal{S}^{t+1} \rightarrow \mathbb{R}^\ell$ ,  $\kappa_t: \mathcal{S}^{t+1} \rightarrow [0, \infty)$ , and  $\zeta_t: \Theta \rightarrow \mathbb{R}$  such that

$$L_\theta^t(\{\mathbf{x}\}) = \kappa_t(\mathbf{x}) \exp(\boldsymbol{\eta}(\theta) \cdot \tau_t(\mathbf{x}) - \zeta_t(\theta))$$

for any  $\mathbf{x} = (x_0, \dots, x_t) \in \mathcal{S}^{t+1}$  and any  $\theta \in \Theta$ , and that the consistency condition holds. By lemma 2.3.2 this defines the probability mass function with respect to the counting measure.

We want to extend  $L_\theta^t$  on length  $t + 1$  vectors to a measure  $L_\theta$  on sequences. For each  $\theta \in \Theta$  we can do this because, for each  $t \in \mathbb{N}$ ,  $L_\theta^t$  is a probability measure on  $(\mathcal{S}^{t+1}, 2^{\mathcal{S}^{t+1}})$  that obeys the consistency condition. As we discussed in sub-subsection 2.3.2.1 and after lemma 2.3.4,  $\mathcal{S}^{t+1}$  is a complete, separable metric space under the discrete metric, and  $2^{\mathcal{S}^{t+1}}$  is the Borel  $\sigma$ -algebra for that metric. Thus we can apply to  $\{L_\theta^t\}_{t \in \mathbb{N}}$  a generalization of

Kolmogorov's extension theorem, which says that there is a probability measure  $L_\theta$  on  $(\mathcal{S}^{\mathbb{N}}, \mathcal{S}_{\mathbb{N}})$  such that  $L_\theta(\{x \in \mathcal{S}^{\mathbb{N}} \mid (x_0, \dots, x_t) \in A\}) = L_\theta^t(A)$  for all  $A \subseteq \mathcal{S}^{t+1}$  for all  $t \in \mathbb{N}$  (Shiryaev, 2016, chap. 2, § 3.4, Thm. 3 and Remark 4, pp. 196–199).

The following definition, which entails the ensuing equalities, then extends  $P_\theta^t$  to  $P_\theta$ :

$$P_\theta(\{\omega \in \Omega \mid \mathbf{X}_t(\omega) \in A\}) := L_\theta(\{x \in \mathcal{S}^{\mathbb{N}} \mid (x_0, \dots, x_t) \in A\}) = L_\theta^t(A) = P_\theta^t(\mathbf{X}_t \in A)$$

for all  $A \subseteq \mathcal{S}^{t+1}$  and all  $t \in \mathbb{N}$ . Since sets of the form  $\{\omega \in \Omega \mid \mathbf{X}_t(\omega) \in A\}$  generate  $\sigma(X)$  and are closed under finite intersections,  $L_\theta$  uniquely determines  $P_\theta$  (Jacod & Protter, 2004, Cor. 6.1, p. 36). Thus  $P_\theta^t$  is the restriction of  $P_\theta$  to  $\mathcal{F}_t$ .

To prove eq. (2.10), we need a common dominating measure, which  $P$  will provide for itself. Fix  $t \in \mathbb{N}$ . If  $\kappa_t(\mathbf{x}) = 0$  for some  $\mathbf{x} \in \mathcal{S}^{t+1}$ , then  $L_\theta^t(\{\mathbf{x}\}) = 0$  for all  $\theta \in \Theta$ . Hence  $L_\theta^t \sim L_{\theta_0}^t$  for all  $\theta, \theta_0 \in \Theta$ . Fix some  $\theta_0 \in \Theta$ . Moreover, for each  $\mathbf{x} \in \mathcal{S}^{t+1}$  and  $\theta \in \Theta$ , we have  $P_\theta^t(\mathbf{X}_t = \mathbf{x}) = L_\theta^t(\{\mathbf{x}\})$ , so  $P_\theta^t \sim P_{\theta_0}^t$  for all  $\theta \in \Theta$ . Notice that  $P_{\theta_0}^t$  dominates  $P_\theta^t$ , and, being finite,  $P_{\theta_0}$  is  $\sigma$ -finite.

We have, for all  $\mathbf{x} \in \mathcal{S}^{t+1}$  such that  $\kappa_t(\mathbf{x}) \neq 0$ ,

$$\frac{P_\theta^t(\mathbf{X}_t = \mathbf{x})}{P_{\theta_0}^t(\mathbf{X}_t = \mathbf{x})} = \frac{L_\theta^t(\{\mathbf{x}\})}{L_{\theta_0}^t(\{\mathbf{x}\})} = \exp[(\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta_0)) \cdot \boldsymbol{\tau}_t(\mathbf{x}) - (\zeta_t(\theta) - \zeta_t(\theta_0))]. \quad (2.11)$$

Therefore,

$$\begin{aligned} P_\theta^t(\mathbf{X}_t = \mathbf{x}) &= \exp[(\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta_0)) \cdot \boldsymbol{\tau}_t(\mathbf{x}) - (\zeta_t(\theta) - \zeta_t(\theta_0))] P_{\theta_0}^t(\mathbf{X}_t = \mathbf{x}) \\ &= \int_{\{\omega \in \Omega \mid \mathbf{X}_t(\omega) = \mathbf{x}\}} \exp[(\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta_0)) \cdot \boldsymbol{\tau}_t(\mathbf{X}_t(\omega)) - (\zeta_t(\theta) - \zeta_t(\theta_0))] P_{\theta_0}^t(d\omega) \end{aligned}$$

To establish eq. (2.10), define  $q_t(\omega) := 1$  and  $\mathbf{B}_t(\omega) := \boldsymbol{\tau}_t(\mathbf{X}_t(\omega))$  for all  $\omega \in \Omega$ , which makes  $q_t$  and  $\mathbf{B}_t$  adapted to  $\{\mathcal{F}_t\}_t$ . Let  $a_t(\boldsymbol{\theta}) := e^{-(\zeta_t(\boldsymbol{\theta}) - \zeta_t(\boldsymbol{\theta}_0))}$ . By the  $P_{\theta_0}^t$ -almost sure uniqueness of Radon-Nikodým densities (Shiryaev, 2016, chap. 2, § 6.7, Radon-Nikodým theorem, p. 233), we have, for  $P_{\theta_0}^t$ -almost all  $\omega \in \Omega$ ,

$$\begin{aligned} \frac{dP_\theta^t}{dP_{\theta_0}^t}(\omega) &= \exp[(\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta_0)) \cdot \boldsymbol{\tau}_t(\mathbf{X}_t(\omega)) - (\zeta_t(\theta) - \zeta_t(\theta_0))] \\ &= a_t(\boldsymbol{\theta}) q_t(\omega) \exp((\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta_0)) \cdot \mathbf{B}_t(\omega)). \quad \square \end{aligned} \quad (2.12)$$

**2.3.2.3 Initial-Condition Exponential Families.** We now introduce notation terminology, most of it from Küchler and Sørensen (1997, chap. 6), that will allow us to condition joint

probabilities for an  $\mathcal{S}$ -valued stochastic process  $X = \{X_t\}_{t \in \mathcal{T}}$  on the *initial event* or *initial condition*  $\{X_0 = x\}$  for some  $x \in \mathcal{S}$ . We do not usually care what  $x$  is. Assume that the filtration  $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$  is generated by  $X$ . Let  $S \subseteq \mathcal{S}$ . Let  $Q_S := \{Q_{\theta,x}\}_{\theta \in \Theta, x \in S}$  be a set of probability measures on  $(\Omega, \mathcal{F})$  such that  $Q_{\theta,x}(X_0 = x) = 1$  for all  $\theta \in \Theta$  and  $x \in S$ , and such that  $x \mapsto Q_{\theta,x}(A)$  is measurable with respect to  $\mathcal{S}$  for every  $\theta \in \Theta$  and every  $A \in \mathcal{F}$ . We say that  $Q_S$  is an *S-initial-condition exponential family for X*<sup>14</sup> if there exist

- some  $\theta_0 \in \Theta$  such that for every  $\theta \in \Theta$ , every  $x \in S$ , and every  $t \in \mathcal{T}$ , we have that  $Q_{\theta,x}^t \sim Q_{\theta_0,x}^t$ ;
- non-random functions  $\eta: \Theta \rightarrow \mathbb{R}^\ell$  and  $\zeta_t: \Theta \rightarrow \mathbb{R}$  for each  $t \in \mathcal{T}$  such that  $\zeta_0(\theta) = 0$  for all  $\theta \in \Theta$ ;
- an  $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$ -adapted,  $\mathbb{R}^\ell$ -valued stochastic process  $B$  such that  $B_0 = \mathbf{0}$ ;

such that

$$\frac{dQ_{\theta,x}^t}{dQ_{\theta_0,x}^t} = \exp(\eta(\theta) \cdot B_t - \zeta_t(\theta)) \quad (2.13)$$

for all  $x \in S$  and  $t \in \mathcal{T}$ . If  $S = \mathcal{S}$  we say that  $Q := Q_{\mathcal{S}}$  is an *initial-condition exponential family for X*.

Exponential families of stochastic processes and initial-condition exponential families are related via the distribution on the initial condition of the stochastic process. Let  $\pi = \{\pi_\theta\}_{\theta \in \Theta}$  be a family of distributions on  $\mathcal{S}$ , and define (Küchler & Sørensen, 1997, § 6.1, p. 65)

$$Q_{\theta,\pi_\theta}(E) := \int_{\mathcal{S}} Q_{\theta,x}(E) \pi_\theta(dx) \quad (2.14)$$

for all  $E \in \mathcal{F}$  and  $\theta \in \Theta$ . This implies that  $Q_{\theta,\pi_\theta}(X_0 \in S) = \pi_\theta(S)$  for all measurable  $S \in \mathcal{S}$ . For this reason we call  $\pi_\theta$  the *initial distribution* of  $Q_{\theta,\pi_\theta}$ .

**Lemma 2.3.6** (Küchler and Sørensen, 1997, Prop. 6.1.2, p. 66).  *$\{Q_{\theta,\pi_\theta}\}_{\theta \in \Theta}$  is an exponential family of the stochastic process  $X$  if and only if  $\{\pi_\theta\}_{\theta \in \Theta}$  is an exponential family of initial distribu-*

<sup>14</sup>↑Küchler and Sørensen (1997, Def. 6.1.1, p. 66) uses the term *S-conditional exponential family*, but we renamed it here to avoid clashing with eq. (2.7).

tions on  $\mathcal{S}$  and there exists  $S \in \mathcal{S}$  such that  $\pi_{\theta_0}(S) = 1$  for some  $\theta_0 \in \Theta$  and  $\{Q_{\theta,x}\}_{\theta \in \Theta, x \in S}$  is an  $S$ -initial-condition exponential family for  $X$ .

The following corollary puts initial-condition exponential families in the more familiar terms of conditional probabilities.

**Corollary 2.3.7.** *Let  $P := \{P_{\theta}\}_{\theta \in \Theta}$  be an exponential family of the  $\mathcal{S}$ -valued stochastic process  $X = \{X_t\}_{t \in \mathcal{T}}$  on  $(\Omega, \mathcal{F})$ . Suppose that  $X_0$  takes on at most countably many values  $P_{\theta_0}$ -almost surely for some  $\theta_0 \in \Theta$ . Define  $Q_{\theta,x}(E) := P_{\theta}(E \mid X_0 = x)$  for each  $E \in \mathcal{F}$ ,  $\theta \in \Theta$ , and  $x \in \mathcal{S}$  with  $P_{\theta_0}(X_0 = x) > 0$ . (Assume the  $\sigma$ -algebra  $\mathcal{S}$  on  $\mathcal{S}$  contains all countable subsets of  $\mathcal{S}$ .<sup>15</sup>) Then  $\{Q_{\theta,x}\}_{\theta \in \Theta, x \in \mathcal{S}}$  is an initial-condition exponential family for  $X$ .*

*Proof.* For each  $\theta \in \Theta$ , define  $\pi_{\theta}$  to be the law of  $X_0$ , meaning  $\pi_{\theta}(A) := P_{\theta}(X_0 \in A)$  for all  $A \in \mathcal{S}$ . If we can show that  $\{Q_{\theta,\pi_{\theta}}\}_{\theta \in \Theta} = P$  then it's an exponential family and, by lemma 2.3.6,  $\{Q_{\theta,x}\}_{\theta \in \Theta, x \in \mathcal{S}}$  is an initial-condition exponential family for  $X$ .

To prove this equality, we need to show that eq. (2.14) equals  $P_{\theta}(E)$  for all  $\theta \in \Theta$  and  $E \in \mathcal{F}$ . Define  $S := \{x \in \mathcal{S} \mid P_{\theta_0}(X_0 = x) > 0\}$ , which is countable, and, by hypothesis,  $\pi_{\theta_0}(S) = P_{\theta_0}(S) = 1$ . Moreover,  $\pi_{\theta}(S) = P_{\theta}(S) = 1$  for all  $\theta \in \Theta$  because the probability measures in  $P$  are all equivalent (Küchler & Sørensen, 1997, chap. 3, § 1, p. 20). Thus, for all  $\theta \in \Theta$  and  $E \in \mathcal{F}$ , the inverse image of the mapping  $x \mapsto Q_{\theta,x}(E) = P_{\theta}(E \mid X_0 = x)$  is a subset of  $S$  and is thereby countable, so the mapping is measurable. We compute eq. (2.14) for any  $\theta \in \Theta$  and  $E \in \mathcal{F}$  to be

$$\int_{\mathcal{S}} Q_{\theta,x}(E) \pi_{\theta}(dx) = \sum_{x \in S} Q_{\theta,x}(E) \pi_{\theta}(\{x\}) = \sum_{x \in S} P_{\theta}(E \mid X_0 = x) P_{\theta}(X_0 = x) = P_{\theta}(E). \quad \square$$

We have the following partial converse.

**Lemma 2.3.8.** *Suppose  $Q$  is the  $S$ -initial-condition exponential family for  $X$  in the definition above. For all  $x \in S$ ,  $\{Q_{\theta,x}\}_{\theta \in \Theta}$  is an exponential family for the stochastic process  $Y = \{Y_t\}_{t \in \mathcal{T}}$  defined by  $Y_0 := x$  and  $Y_t := X_t$  if  $t > 0$ .*

<sup>15</sup>↑ This is a weak, purely technical requirement. The Borel  $\sigma$ -algebra for  $\mathbb{R}^n$ , discrete spaces, and other Polish spaces all satisfy it.

*Proof.* Fix any  $x \in S$ .  $Q_{\theta,x}^t \ll Q_{\theta_0,x}^t$  for all  $t \in \mathcal{T}$  and all  $\theta \in \Theta$ , and  $Q_{\theta_0,x}$  is  $\sigma$ -finite because it's finite. Thus we may choose  $\mu := Q_{\theta_0,x}$  as the dominating measure. Put  $a_t(\theta) := e^{-\zeta_t(\theta)}$  for all  $\theta \in \Theta$  and  $q_t(\omega) := 1$  for all  $\omega \in \Omega$ . Then, from eq. (2.13), since  $Q_{\theta,x}(X_0 \neq x) = 0$  and  $\mu^t = Q_{\theta_0,x}^t$ , we have

$$\begin{aligned} \frac{dQ_{\theta,x}^t}{d\mu^t} &= \frac{dQ_{\theta,x}^t}{dQ_{\theta_0,x}^t} = \exp(\eta(\theta) \cdot \mathbf{B}_t - \zeta_t(\theta)) = a_t(\theta) q_t \exp(\eta(\theta) \cdot \mathbf{B}_t) \\ &= a_t(\theta) q_t \exp(\eta(\theta) \cdot \mathbf{B}_t \mathbb{1}(X_0 = x)) \end{aligned}$$

$\mu^t$ -almost surely for all  $\theta \in \Theta$  and  $t \in \mathcal{T}$ . This satisfies eq. (2.10). Finally,  $\mathbf{B}_t \mathbb{1}(X_0 = x)$  is adapted to the filtration that  $Y$  generates because  $\mathbf{B}_t \mathbb{1}(X_0 = x) = 0$  is constant on  $\{X_0 \neq x\}$ .  $\square$

**2.3.2.4 The Main Theorem.** All of that work was just so we could present the following lemma by Küchler and Sørensen, which we will use to prove the main theorem of this subsection.

**Lemma 2.3.9** (Küchler and Sørensen, 1997, Cor. 6.3.4; Küchler and Sørensen, 1998, Cor. 4.10). *Suppose  $\mathcal{S}$  is a Polish space (Küchler & Sørensen, 1997, Condition 6.3.1, p. 71; Küchler & Sørensen, 1998, Assumption 4.1, p. 11). Let  $\mathcal{T} = \mathbb{N}$  and let  $Q = \{Q_{\theta,x}\}_{\theta \in \Theta, x \in \mathcal{S}}$  be an initial-condition exponential family for  $X$ , where  $X = \{X_t\}_{t \in \mathbb{N}}$  is a Markov chain with transition functions  $T_\theta: \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$  for each  $\theta \in \Theta$ , so that  $T_\theta(a, B) = Q_{\theta,a}(X_1 \in B)$ . Then the transition functions  $T_\theta(a, B)$  form an exponential family of probabilities on  $(\mathcal{S}, \mathcal{S})$ . In particular, if  $dQ_{\theta,x}^t / dQ_{\theta_0,x}^t$  is given by eq. (2.13), we have*

$$\frac{T_\theta(a, db)}{T_{\theta_0}(a, db)} = \exp[\eta(\theta) \cdot \tau(a, b) - \zeta_1(\theta)], \quad (2.15)$$

where  $\tau$  is a measurable function on  $\mathcal{S} \times \mathcal{S}$  such that  $\mathbf{B}_1(\omega) = \tau(X_0(\omega), X_1(\omega))$ .

Conversely, if eq. (2.15) holds, then one can construct an initial-condition exponential family for the Markov chain  $X$  with a representation eq. (2.13), where the  $\eta$  is the same as in eq. (2.15),  $\mathbf{B}_t := \sum_{i=1}^t \tau(X_{i-1}, X_i)$ , and  $\zeta_t(\theta) := t \zeta_1(\theta)$ .

The essence of the proof, which we omit here, is that when time is discrete and  $\mathcal{S}$  is Polish, the fact that  $\mathbf{B}_1$  is measurable with respect to  $\mathcal{F}_1 = \sigma(X_0, X_1)$  means that there is

some measurable function  $\tau: \mathcal{S}^2 \rightarrow \mathbb{R}^\ell$  such that  $B_1 = \tau(X_0, X_1)$  by lemma 2.3.1. Further, in the case when  $\mathcal{S}$  is discrete,  $T_\theta(a, b) = Q_{\theta,a}(X_1 = b)$  is the law of  $X_1$  under  $Q_{\theta,a}$ ; lemma 2.3.5 says the law must be an exponential family if and only if  $\{Q_{\theta,a}\}_{\theta \in \Theta}$  is an exponential family for a stochastic process, which it must be by lemma 2.3.8.

**Lemma 2.3.10.** *Suppose that  $\mathcal{S}$  is at most countable and that  $T := \{T_\theta\}_{\theta \in \Theta}$  is a family of transition matrices and  $\theta_0$  is a fixed element of  $\Theta$ . Then eq. (2.15) holds for all  $b \in \mathcal{S}$  for which  $T_{\theta_0}(a, b) \neq 0$  if and only if  $T$  is the MEF such that*

$$T_\theta(a, b) = \kappa(a, b) \exp[\eta(\theta) \cdot \tau(a, b) - \zeta(\theta)] \quad (2.16)$$

for all  $\theta \in \Theta$  and all  $a, b \in \mathcal{S}$ , where  $\kappa(a, b) = T_{\theta_0}(a, b)$  and  $\zeta = \zeta_1$ .

*Proof.* Most of this proof is just reading eq. (2.15) correctly. In lemma 2.3.9,  $T_\theta$  is a transition function, so  $T_\theta(a, B) = Q_{\theta,a}^1(X_1 \in B)$  for all  $\theta \in \Theta$ ,  $a \in \mathcal{S}$ , and  $B \in \mathcal{S}$ . That means that  $T_\theta(a, \cdot)$  is the law of  $X_1$  under  $Q_{\theta,a}^1$ . Since  $Q_{\theta,a}^1 \ll Q_{\theta_0,a}^1$  by the definition of an initial-condition exponential family, we also have  $T_\theta(a, \cdot) \ll T_{\theta_0}(a, \cdot)$ . The left-hand side of eq. (2.15) is the Radon-Nikodým density of  $T_\theta(a, \cdot)$  with respect to  $T_{\theta_0}(a, \cdot)$ . By lemma 2.3.2 with  $\mu = T_\theta(a, \cdot)$  and  $\nu = T_{\theta_0}(a, \cdot)$ , we therefore have that eq. (2.15) equals  $T_\theta(a, b)/T_{\theta_0}(a, b)$ . Read in this light, eq. (2.15) is exactly the same as eq. (2.16) after replacing  $\kappa(a, b)$  with  $T_{\theta_0}(a, b)$ , and  $\zeta$  with  $\zeta_1$ : when  $T_{\theta_0}(a, b) \neq 0$ , just multiply or divide both sides of eq. (2.16) or eq. (2.15) to get the other; when  $T_{\theta_0}(a, b) = 0$ , absolute continuity guarantees us that  $T_\theta(a, b) = 0$ .

Equation (2.16) looks like the definition of MEF in eq. (2.8) on page 14 when we replace  $P_\theta(a, b)$  with  $T_\theta(a, b)$ . Notice that  $\theta_0$  is fixed and does not depend on  $\theta$ , so  $\kappa$  does not depend on the parameter. □

**Theorem 2.3.11.** *Let  $\mathcal{T} = \mathbb{N}$ ,  $\mathcal{S}$  be at most countable, and  $X = \{X_t\}_{t \in \mathbb{N}}$  be an  $\mathcal{S}$ -valued stochastic process. Suppose that  $X$  is a Markov chain under all the probability measures in the family  $Q := \{Q_{\theta,x}\}_{\theta \in \Theta, x \in \mathcal{S}}$ , each with a corresponding transition matrix in  $T := \{T(\theta)\}_{\theta \in \Theta}$  and satisfying  $Q_{\theta,x}(X_0 = x) = 1$ . That is,  $T_{a,b}(\theta) = Q_{\theta,a}(X_1 = b)$  for all  $\theta \in \Theta$  and all  $a, b \in \mathcal{S}$ . Define  $L_{\theta,x_0}^t$  to be the law of  $X_1, \dots, X_t$  under  $Q_{\theta,x_0}^t$  for all  $\theta \in \Theta$ ,  $t \in \mathbb{N}$ , and  $x_0 \in \mathcal{S}$ .*

Then  $T$  is the MEF in eq. (2.16) on the previous page with  $\kappa(a, b) = T_{\theta_0}(a, b)$  for some  $\theta_0 \in \Theta$  and all  $a, b \in \mathcal{S}$  if and only if  $\{L_{\theta, x_0}^t\}_{\theta \in \Theta}$  is the exponential family on  $\mathcal{S}^t$  given by

$$\begin{aligned} L_{\theta, x_0}^t(x_1, \dots, x_t) &:= L_{\theta, x_0}^t(\{(x_1, \dots, x_t)\}) \\ &= \exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, x_{i+1}) - t\zeta(\theta)\right) \prod_{i=0}^{t-1} \kappa(x_i, x_{i+1}) \end{aligned} \quad (2.17)$$

for all  $x_0, x_1, \dots, x_t \in \mathcal{S}$  and all  $t \in \mathbb{N}$ , and  $\kappa(a, b) = L_{\theta_0, a}^1(b)$  for some  $\theta_0 \in \Theta$  and all  $a, b \in \mathcal{S}$ .

*Proof.* The right-hand side of eq. (2.17) is an exponential family representation with sufficient statistic  $\sum_{i=0}^{t-1} \tau(X_i, X_{i+1})$ , carrier measure  $\prod_{i=0}^{t-1} \kappa(x_i, x_{i+1})$ , parameter function  $\eta$ , parameter space  $\Theta$ , and log-partition function  $t\zeta$ .

( $\implies$ ). Suppose  $T$  is the MEF in eq. (2.16) with  $\kappa(a, b) = T_{\theta_0}(a, b)$  for some  $\theta_0 \in \Theta$  and all  $a, b \in \mathcal{S}$ . Fix  $t \in \mathbb{N}$ ,  $x_0, x_1, \dots, x_t \in \mathcal{S}$ , and  $\theta \in \Theta$ . Since  $X$  is a Markov chain under  $Q_{\theta, x_0}$ , we have

$$L_{\theta, x_0}^t(x_1, \dots, x_t) = \prod_{i=0}^{t-1} T_{\theta}(x_i, x_{i+1}) = \prod_{i=0}^{t-1} \kappa(a, b) \exp[\eta(\theta) \cdot \tau - \zeta(\theta)],$$

which equals the right-hand side of eq. (2.17).

By hypothesis,  $\kappa(a, b) = T_{\theta_0}(a, b) = L_{\theta_0, a}^1(b)$  for any  $a, b \in \mathcal{S}$ .

( $\impliedby$ ). Suppose  $\{L_{\theta, x_0}^t\}_{\theta \in \Theta}$  is the exponential family on  $\mathcal{S}^t$  given by eq. (2.17) and  $\kappa(a, b) = T_{\theta_0}(a, b)$  for some  $\theta_0 \in \Theta$ .

We want to apply lemma 2.3.5, so we first prove the consistency condition

$$L_{\theta, x_0}^{t+1}(\{(x_1, \dots, x_t)\} \times \mathcal{S}) = L_{\theta, x_0}^t(x_1, \dots, x_t).$$

Fix  $t \in \mathbb{N}$ ,  $x_0, x_1, \dots, x_t \in \mathcal{S}$ , and  $\theta \in \Theta$ . In the second step below we factor, and all the work afterward is simplifying:

$$\begin{aligned} &L_{\theta, x_0}^{t+1}(\{(x_1, \dots, x_t)\} \times \mathcal{S}) \\ &= \sum_{y \in \mathcal{S}} \exp\left[\eta(\theta) \cdot \left(\sum_{i=0}^{t-1} \tau(x_i, x_{i+1}) + \tau(x_t, y)\right) - (t+1)\zeta(\theta)\right] \kappa(x_t, y) \prod_{i=0}^{t-1} \kappa(x_i, x_{i+1}) \\ &= \left(\exp\left[\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, x_{i+1}) - t\zeta(\theta)\right] \prod_{i=0}^{t-1} \kappa(x_i, x_{i+1})\right) \sum_{y \in \mathcal{S}} \kappa(x_t, y) e^{\eta(\theta) \cdot \tau(x_t, y) - \zeta(\theta)} \\ &= L_{\theta, x_0}^t(x_1, \dots, x_t) \sum_{y \in \mathcal{S}} L_{\theta, x_t}^1(y) = L_{\theta, x_0}^t(x_1, \dots, x_t). \end{aligned}$$



By lemma 2.3.5, if  $x_0 \in \mathcal{S}$ , then  $\{Q_{\theta, x_0}\}_{\theta \in \Theta}$  is an exponential family for the stochastic process  $X$ . Moreover, since  $L_{\theta, x_0}^t(x_1, \dots, x_t) = Q_{\theta, x_0}^t(X_1 = x_1, \dots, X_t = x_t) = 0$  if and only if  $\prod_{i=0}^{t-1} \kappa(x_i, x_{i+1}) = 0$ , we can conclude that  $Q_{\theta, x_0}^t \sim Q_{\theta_0, x_0}^t$  for all  $t \in \mathbb{N}$ ,  $x_0 \in \mathcal{S}$ , and  $\theta, \theta_0 \in \Theta$ . Plugging  $L$  into eq. (2.11) and replacing  $P_\theta$  in eq. (2.11) with  $Q_{\theta, x_0}$ , setting  $\tau_t(x) := \sum_{i=0}^{t-1} \tau(x_i, x_{i+1})$ , and setting  $\zeta_t(\theta) := t\zeta(\theta)$ , we get from eq. (2.12) that

$$\frac{dQ_{\theta, x}^t}{dQ_{\theta_0, x}^t}(\omega) = \exp \left[ (\eta(\theta) - \eta(\theta_0)) \cdot \sum_{i=0}^{t-1} \tau(X_i(\omega), X_{i+1}(\omega)) - t(\zeta(\theta) - \zeta(\theta_0)) \right] \quad (2.18)$$

$Q_{\theta_0}^t$ -almost all  $\omega \in \Omega$ , all  $\theta \in \Theta$ , all  $t \in \mathbb{N}$ , and all  $x_0 \in \mathcal{S}$ . This satisfies eq. (2.13).  $\sum_{i=0}^{t-1} \tau(X_i, X_{i+1})$  is adapted to  $\sigma(X_0, \dots, X_t)$ , and  $Q_{\theta, x_0}(X_0 = x_0) = 1$  for all  $\theta \in \Theta$  and  $x_0 \in \mathcal{S}$ .  $x_0 \mapsto Q_{\theta, x_0}(E)$  is  $\mathcal{S}$ -measurable for all  $E \in \mathcal{F}$  and all  $\theta \in \Theta$  because  $\mathcal{S} = 2^{\mathcal{S}}$ , so every function is measurable. Therefore  $\{Q_{\theta, x_0}\}_{\theta \in \Theta, x_0 \in \mathcal{S}}$  is an initial-condition exponential family for  $X$ .

Moreover, by hypothesis, we have  $\kappa(a, b) = L_{\theta_0, a}^1(b) = \kappa(a, b)e^{\eta(\theta_0) \cdot \tau(a, b) - \zeta(\theta_0)}$  for all  $a, b \in \mathcal{S}$ . Since  $T_\theta(a, b) = L_{\theta, a}^1(b)$  for any  $a, b \in \mathcal{S}$  and every  $\theta \in \Theta$ , we have that either  $\eta(\theta) \cdot \tau(a, b) = \zeta(\theta)$  or  $T_\theta(a, b) = 0$ . But  $\prod_{i=0}^{t-1} \kappa(X_i, X_{i+1}) = 0$  implies  $L_{\theta, x_0}^t(X_1, \dots, X_t) = 0$  by eq. (2.17), which implies that  $Q_{\theta, x_0}^t \left( \prod_{i=0}^{t-1} \kappa(X_i, X_{i+1}) = 0 \right) = 0$ , which in turn implies that eq. (2.18) equals zero on  $\left\{ \prod_{i=0}^{t-1} \kappa(X_i, X_{i+1}) = 0 \right\}$  for all  $\theta \in \Theta$ . Thus  $\sum_{i=0}^{t-1} \eta(\theta) \cdot \tau(X_i, X_{i+1}) = t\zeta(\theta)$   $Q_{\theta_0, x_0}$ -almost surely for all  $x_0 \in \mathcal{S}$ . Hence, eq. (2.18) equals eq. (2.13)  $Q_{\theta_0, x_0}$ -almost surely.

By lemma 2.3.9, eq. (2.15) obtains. By lemma 2.3.10, eq. (2.16) holds.  $\square$

**2.3.3 CEF Likelihood Functions.** Algebraically, the difference between MEFS and CEFs is the asymmetry of the log-partition functions. To show this, we first need a lemma that makes it easier to apply the distributive law.

**Lemma 2.3.12.** *Let  $\mathcal{T}$  be a non-empty, finite set. Let  $\{B_i\}_{i \in \mathcal{T}}$  be a family of sets;  $B := \bigcup_{i \in \mathcal{T}} B_i$ ;  $R$  be a commutative ring;  $f_i: B_i \rightarrow R$  for each  $i \in \mathcal{T}$ ; and  $\mathcal{F}_{\mathcal{T}} = \{y \in B^{\mathcal{T}} \mid y_i \in B_i \text{ for all } i \in \mathcal{T}\}$ . Suppose that, for all  $i \in \mathcal{T}$ , either  $B_i$  is finite; or  $B_i$  is countable and  $R$  is endowed with a topology in which  $\sum_{b \in B_i} f_i(b)$  converges. Then*

$$\prod_{i \in \mathcal{T}} \sum_{b \in B_i} f_i(b) = \sum_{y \in \mathcal{F}_{\mathcal{T}}} \prod_{i \in \mathcal{T}} f_i(y_i). \quad (2.19)$$

*Proof.* We proceed by induction on the cardinality of  $\mathcal{T}$ . Assume that  $B_i \neq \emptyset$  for all  $i \in \mathcal{T}$  because otherwise both sides of eq. (2.19) would contain sums over empty sets and thus be zero, rendering the lemma true anyway. For both the base and inductive cases, we'll need to fix some specific element  $j \in \mathcal{T}$ .

*Base case.* If  $|\mathcal{T}| = 1$ , then  $\mathcal{T} = \{j\}$ ,  $\mathcal{F}_{\mathcal{T}} = \{j\} \times B_j$ , and

$$\prod_{i \in \mathcal{T}} \sum_{b \in B_i} f_i(b) = \sum_{b \in B_j} f_j(b) = \sum_{y \in B_j^{(j)}} f_j(y_j) = \sum_{y \in \mathcal{F}_{\mathcal{T}}} \prod_{i \in \mathcal{T}} f_i(y_i).$$

*Inductive case.* Suppose that  $|\mathcal{T}| > 1$  and eq. (2.19) holds for index sets one smaller than  $\mathcal{T}$ , such as the set  $\mathcal{U} = \mathcal{T} \setminus \{j\}$ . Then the inductive hypothesis tells us that

$$\prod_{i \in \mathcal{U}} \sum_{b \in B_i} f_i(b) = \sum_{y \in \mathcal{F}_{\mathcal{U}}} \prod_{i \in \mathcal{U}} f_i(y_i). \quad (2.20)$$

Since  $\mathcal{U}$  is finite and  $\sum_{b \in B_i} f_i(b)$  is a finite sum or converges in  $R$  for all  $i \in \mathcal{U}$ , eq. (2.20)'s left-hand side is in  $R$ . Therefore its right-hand side is either a finite sum or converges in  $R$ . Thus,

$$\prod_{i \in \mathcal{T}} \sum_{b \in B_i} f_i(b) = \left( \sum_{b \in B_j} f_j(b) \right) \left( \prod_{i \in \mathcal{U}} \sum_{b \in B_i} f_i(b) \right) = \left( \sum_{b \in B_j} f_j(b) \right) \left( \sum_{y \in \mathcal{F}_{\mathcal{U}}} \prod_{i \in \mathcal{U}} f_i(y_i) \right)$$

Either because  $B_j$  is finite or because  $\sum_{b \in B_j} f_j(b)$  converges and  $\sum_{y \in \mathcal{F}_{\mathcal{U}}} \prod_{i \in \mathcal{U}} f_i(y_i)$  is a constant with respect to  $b$ , we can distribute:

$$= \sum_{b \in B_j} \left( f_j(b) \sum_{y \in \mathcal{F}_{\mathcal{U}}} \prod_{i \in \mathcal{U}} f_i(y_i) \right)$$

Either because  $\mathcal{F}_{\mathcal{U}}$  is finite or because  $\sum_{y \in \mathcal{F}_{\mathcal{U}}} \prod_{i \in \mathcal{U}} f_i(y_i)$  converges and  $f_j(b)$  is a constant with respect to  $y$ , we can distribute:

$$\begin{aligned} &= \sum_{b \in B_j} \left( \sum_{y \in \mathcal{F}_{\mathcal{U}}} f_j(b) \prod_{i \in \mathcal{U}} f_i(y_i) \right) \\ &= \sum_{y_j \in B_j} \sum_{y \in \mathcal{F}_{\mathcal{U}}} \prod_{i \in \mathcal{T}} f_i(y_i) = \sum_{y \in \mathcal{F}_{\mathcal{T}}} \prod_{i \in \mathcal{T}} f_i(y_i). \end{aligned} \quad \square$$

In the lemma,  $\mathcal{T}$ 's being non-empty is important because otherwise the left side of eq. (2.19) would be one and the right side zero.

**Theorem 2.3.13.** Let  $\mathcal{S}$  be at most countable and  $\theta \in \Theta$ . Suppose  $X$  is a Markov chain whose transition matrix  $P_\theta$  is given in eq. (2.7). Write the vector  $(\zeta(a, \theta))_{a \in \mathcal{S}}$  as  $\zeta(\theta) \in \mathbb{R}^{\mathcal{S}}$ . For time  $t > 0$  and  $x \in \mathcal{S}^{t+1}$  the joint probability mass function for  $X_1, \dots, X_t$  conditional on  $X_0 = x_0$  is

$$L_{\theta, x_0}^t(x_1, \dots, x_t) = \exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, x_{i+1}) - \zeta(\theta)^\top n_t \mathbf{1}\right) \prod_{i=0}^{t-1} \kappa(x_i, x_{i+1}) \quad (2.21)$$

$$= \frac{\exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, x_{i+1})\right) \prod_{i=0}^{t-1} \kappa(x_i, x_{i+1})}{\sum_{\substack{y \in \mathcal{S}^{t+1} \\ y_0 = x_0}} \exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, y_{i+1})\right) \prod_{i=0}^{t-1} \kappa(x_i, y_{i+1})}. \quad (2.22)$$

where  $n_t$  is the transition-count matrix for  $x_1, \dots, x_t$ .

*Proof.* Fix time  $t \geq 1$ . Equation (2.21) follows from the usual calculations with Markov chains (Hoel et al., 1972, chap. 1).

$$L_{\theta, x_0}^t(x_1, \dots, x_t) = \prod_{i=0}^{t-1} P_\theta(x_i, x_{i+1}) = \frac{\exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, x_{i+1})\right) \prod_{i=0}^{t-1} \kappa(x_i, x_{i+1})}{\prod_{i=0}^{t-1} \exp(\zeta(x_i, \theta))}.$$

In the denominator, we group factors by values of  $x_i$ :

$$\begin{aligned} \prod_{i=0}^{t-1} e^{\zeta(x_i, \theta)} &= \prod_{a \in \mathcal{S}} \prod_{\substack{i=0 \\ x_i=a}}^{t-1} e^{\zeta(a, \theta)} = \prod_{a \in \mathcal{S}} \left( e^{\zeta(a, \theta)} \right)^{n_t(a, +)} = \exp\left( \sum_{a \in \mathcal{S}} \zeta(a, \theta) (n_t \mathbf{1})_a \right) \\ &= \exp(\zeta(\theta)^\top n_t \mathbf{1}) \end{aligned}$$

since the number of times  $i < t$  that  $x_i = a$  is  $n_t(a, +)$  by eq. (2.1).

Now, to prove eq. (2.22), we use eq. (2.4) to write

$$\prod_{i=0}^{t-1} e^{\zeta(x_i, \theta)} = \prod_{i=0}^{t-1} \sum_{b \in \mathcal{S}} \kappa(x_i, b) \exp(\eta(\theta) \cdot \tau(x_i, b))$$

We apply lemma 2.3.12 as follows. Let  $\mathcal{T} := \{0, \dots, t-1\}$ , and, for each  $i \in \mathcal{T}$ , let  $B_i := \mathcal{S}$  (which is finite) and  $f_i(b) := \kappa(x_i, b) \exp(\eta(\theta) \cdot \tau(x_i, b))$  for each  $b \in \mathcal{S}$ . The range of all the  $f_i$ s is  $\mathbb{R}$ , a topological field, and, we have the following sum or (absolute) convergence from eq. (2.4) for each finite or countably infinite  $B_i, i \in \mathcal{T}$ :

$$\sum_{b \in B_i} f_i(b) = \sum_{b \in \mathcal{S}} \kappa(x_i, b) \exp(\eta(\theta) \cdot \tau(x_i, b)) = e^{\zeta(x_i, \theta)},$$

which is finite because we assumed at the outset that  $\eta(\theta)$  lies inside the natural parameter space. The set  $\mathcal{F}_{\mathcal{T}}$  from lemma 2.3.12 is just  $\mathcal{S}^t$ . Plugging these definitions into lemma 2.3.12 allows us to write

$$\begin{aligned} \prod_{i=0}^{t-1} \sum_{b \in \mathcal{S}} \kappa(x_i, b) \exp(\eta(\theta) \cdot \tau(x_i, b)) &= \sum_{y \in \mathcal{S}^t} \prod_{i=0}^{t-1} \kappa(x_i, y_i) \exp(\eta(\theta) \cdot \tau(x_i, y_i)) \\ &= \sum_{y \in \mathcal{S}^t} \exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, y_i)\right) \prod_{i=0}^{t-1} \kappa(x_i, y_i) \end{aligned}$$

Reindex  $(y_0, \dots, y_{t-1})$  as  $(x_0, y_1, \dots, y_t)$  to match indexing of  $x = (x_0, x_1, \dots, x_t)$ :

$$= \sum_{\substack{y \in \mathcal{S}^{t+1} \\ y_0 = x_0}} \exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(x_i, y_{i+1})\right) \prod_{i=0}^{t-1} \kappa(x_i, y_{i+1}). \quad \square$$

We get eq. (2.17) and a similar result for CAEFS from eq. (2.21) as follows. First, in the CAEF case, suppose  $\zeta(a, \theta) = \phi(\theta)\psi_a$ , and denote the vector  $\psi = (\psi_a \mid a \in \mathcal{S})$ . Then

$$\prod_{a \in \mathcal{S}} \left( e^{\zeta(a, \theta)} \right)^{n_t(a, +)} = \left( e^{\phi(\theta)} \right)^{\sum_{a \in \mathcal{S}} \psi_a n_t(a, +)} = \left( e^{\phi(\theta)} \right)^{\psi \cdot \mathbf{n}_t(\cdot, +)}.$$

Then, in the MEF case,  $\psi_a = 1$  for all  $a \in \mathcal{S}$ , and  $\sum_{a \in \mathcal{S}} n_t(a, +) = t$ .

Theorem 2.3.13 facilitates our intuition about theorem 2.3.11 should be true. In the MEF case, applying eq. (2.4) to eq. (2.17) shows that

$$e^{t\zeta(\theta)} = \sum_{\substack{y \in \mathcal{S}^{t+1} \\ y_0 = x_0}} \exp\left(\eta(\theta) \cdot \sum_{i=0}^{t-1} \tau(y_i, y_{i+1})\right) \prod_{i=0}^{t-1} \kappa(y_i, y_{i+1}). \quad (2.23)$$

The difference between the  $x_i$  and the  $y_i$  in the first slot is the difference between MEFS' having exponential family likelihood functions with sufficient statistics

$$\sum_{i=0}^{t-1} \tau(X_i, X_{i+1}), \quad (2.24)$$

the sum of the sufficient statistics seen so far, and CEFS' having exponential family likelihood functions with sufficient statistics  $N_t$ .

The difference in dimension of the sufficient statistics between the CEF case in eq. (2.21) and the MEF case in eq. (2.17) can be large when  $\mathcal{S}$  is large. For example, if  $\mathcal{S}$  is

the set of graphs on the vertex set  $[n]$  for some  $n \in \mathbb{N}$ , the minimal representation of  $N_t$  has  $|\mathcal{S}|^2 - |\mathcal{S}|$  degrees of freedom (Stefanov, 1991, § 2) where

$$|\mathcal{S}|^2 - |\mathcal{S}| = \left(2^{\binom{n}{2}}\right)^2 - 2^{\binom{n}{2}} = \mathcal{O}\left(2^{n^2}\right).$$

In contrast, eq. (2.24) has dimension  $\ell$ .

The following example illustrates how eq. (2.22) is not equal to eq. (2.17).

**Example 2.3.14.** Let the state space  $\mathcal{S} = \{1, 2\}$ , the dimension of the parameter and sufficient statistic be  $\ell = 1$ , the parameter  $\theta = 1$ ,  $\kappa_{ab} = 1$  for all  $a, b \in \mathcal{S}$ , and

$$\tau = \begin{bmatrix} \log 1 & \log 2 \\ \log 3 & \log 4 \end{bmatrix}, \text{ so that } P_{ab} = \frac{e^{\tau(a,b)}}{\sum_{s \in \mathcal{S}} e^{\tau(a,s)}} \implies P = \begin{bmatrix} 1/3 & 2/3 \\ 3/7 & 4/7 \end{bmatrix}.$$

This is a CEF not an MEF because the denominator three in the first row does not equal the denominator seven in the second row. For  $t = 2$ , consider calculating the joint probability that  $(X_0, X_1, X_2) = (x_0, x_1, x_2) = (1, 2, 1)$  conditional on  $X_0 = 1$ . This probability is  $\frac{2}{3} \times \frac{3}{7} = \frac{2}{7}$ .

The numerators of eqs. (2.17) and (2.22) are both

$$\exp(\tau_{12} + \tau_{21}) = 2 \times 3 = 6.$$

The denominator of eq. (2.22) is

$$\begin{aligned} & \sum_{\substack{y \in \mathcal{S}^3 \\ y_0 = x_0}} \exp(\tau(x_0, y_1) + \tau(x_1, y_2)) \\ &= \exp(\tau_{11} + \tau_{21}) + \exp(\tau_{11} + \tau_{22}) + \exp(\tau_{12} + \tau_{21}) + \exp(\tau_{12} + \tau_{22}) \\ &= 1 \times 3 + 1 \times 4 + 2 \times 3 + 2 \times 4 = 21. \end{aligned}$$

The denominator of eq. (2.17) is

$$\begin{aligned} & \sum_{\substack{y \in \mathcal{S}^3 \\ y_0 = x_0}} \exp(\tau(y_0, y_1) + \tau(y_1, y_2)) \\ &= \exp(\tau_{11} + \tau_{11}) + \exp(\tau_{11} + \tau_{12}) + \exp(\tau_{12} + \tau_{21}) + \exp(\tau_{12} + \tau_{22}) \\ &= 1 \times 1 + 1 \times 2 + 2 \times 3 + 2 \times 4 = 17. \end{aligned}$$

Equation (2.22) gives the correct joint probability of  $\frac{2}{7}$  whereas eq. (2.17) gives the incorrect value of  $\frac{6}{17}$ . □

Despite this, some CEFS are MEFS. A first example is when  $\tau(a, b)$  is constant in  $a$ , meaning that transition probabilities from the current state to the next state do not depend on the current state. Such a  $\tau$  would ignore the  $x_t$  term in the denominator of eq. (2.22) and make the Markov chain an IID sequence. The next subsection discusses the structure of MEF transition matrices to constrain the search for such models.

**2.3.4 The Structure of MEF Transition Matrices.** Let  $P = \{P_\theta \mid \theta \in \Theta\}$  be the MEF defined in eq. (2.8). Since  $P_\theta$  is a stochastic matrix,  $1 = \sum_{b \in \mathcal{S}} P_\theta(a, b)$ , so

$$e^{\zeta(\theta)} = \sum_{b \in \mathcal{S}} \kappa(a, b) \exp(\eta(\theta) \cdot \tau(a, b)) = \sum_{b \in \mathcal{S}} \kappa(c, b) \exp(\eta(\theta) \cdot \tau(c, b)). \quad (2.25)$$

for all  $a, c \in \mathcal{S}$  and all  $\theta \in \Theta$ .

Since exponential functions  $x \mapsto e^{rx}$  are linearly independent for different scalars  $r$  (Axler, 2015; Tsumura, 2016, August 17/2017), we can set some of the coefficients equal (The first to point this out in the scalar case were Gani, 1955, pp. 356–357; Bhat & Gani, 1960, p. 454; a scalar example can be found in Bhat, 1988, Example 2). To handle vector parameters we need some notation. Define

$$R(\theta) := \{\eta(\theta) \cdot \tau(a, b) \in \mathbb{R} \mid a, b \in \mathcal{S} \text{ and } \kappa(a, b) \neq 0\},$$

the set of possible values of the exponents, for each  $\theta \in \Theta$ . The first implication of eq. (2.25) is that, for all  $a \in \mathcal{S}$  and all  $\theta \in \Theta$ , we have

$$R(\theta) = \{\eta(\theta) \cdot \tau(a, b) \in \mathbb{R} \mid b \in \mathcal{S} \text{ and } \kappa(a, b) \neq 0\}. \quad (2.26)$$

Equation (2.26) means that every row of  $P_\theta$  has the same values in the exponent. (See example 2.3.17 on the following page.) Moreover,  $|R(\theta)| \leq |\mathcal{S}|$  for every  $\theta \in \Theta$ . This proves the following theorem.

**Theorem 2.3.15.** *For all  $a, b, c \in \mathcal{S}$  such that  $\kappa(a, b) \neq 0$  and all  $\theta \in \Theta$ , there exists  $x \in \mathcal{S}$  such that  $\eta(\theta) \cdot \tau(a, b) = \eta(\theta) \cdot \tau(c, x)$ , i.e.,  $\eta(\theta) \cdot (\tau(a, b) - \tau(c, x)) = 0$ .*

Even if eq. (2.8) is a minimal representation, we cannot conclude from theorem 2.3.15 that  $\tau(a, b) = \tau(c, x)$ . The value of  $x$  may depend on the values of  $\theta$  or of  $a, b$ , or  $c$ . However, we can say something weaker in the scalar case.

**Proposition 2.3.16** (Gani, 1955, pp. 357–358). For the MEF defined in eq. (2.8) with  $\ell = 1$  (i.e.,  $\eta(\theta)$  is a scalar), and supposing that  $\eta(\Theta)$  contains a number other than zero, we have, for all  $a, c \in \mathcal{S}$ ,

$$\{\tau(a, b) \in \mathbb{R} \mid b \in \mathcal{S}\} = \{\tau(c, b) \in \mathbb{R} \mid b \in \mathcal{S}\}.$$

*Proof.* Pick  $a, b, c \in \mathcal{S}$  arbitrarily and  $\theta \in \Theta$  with  $\eta(\theta) \neq 0$ . From theorem 2.3.15 obtain  $x \in \mathcal{S}$  with  $\eta(\theta)\tau(a, b) = \eta(\theta)\tau(c, x)$ , so  $\tau(a, b) = \tau(c, x)$ . This proves both set inclusions.  $\square$

We now examine the coefficients in the sum in eq. (2.25). For each  $c, d \in \mathcal{S}$ ,  $\theta \in \Theta$ , and  $r \in R(\theta)$ , define  $S_c(\theta, r)$  to be the set of all columns in row  $c$  and  $S^d(\theta, r)$  to be the set of all rows in column  $d$  in the transition matrix  $P_\theta$  whose exponent on  $e$  is equal to  $r$ :

$$S_c(\theta, r) = \{d \in \mathcal{S} \mid \eta(\theta) \cdot \tau(c, d) = r\}, \quad (2.27)$$

$$S^d(\theta, r) = \{c \in \mathcal{S} \mid \eta(\theta) \cdot \tau(c, d) = r\}.$$

Then, from eq. (2.25), we have our second implication:

$$\sum_{c \in S_a(\theta, r)} \kappa(a, c) = \sum_{c \in S_b(\theta, r)} \kappa(b, c) \quad \text{for all } r \in R(\theta) \quad (2.28)$$

for all  $a, b \in \mathcal{S}$ , all  $\theta \in \Theta$ . This sum can never be zero lest it violate eq. (2.26).

**Example 2.3.17** (Gani, 1955, Example iv.3.3, pp. 358–359). The following is an example of an MEF with a scalar parameter. Set  $\mathcal{S} := \{1, 2, 3\}$ ,  $\Theta := (0, \infty)$ ,  $\eta(\theta) := \log(\theta)$ ,

$$\tau := \begin{bmatrix} 1 & 1 & 3 \\ 3 & 3 & 1 \\ 1 & 3 & 3 \end{bmatrix}, \quad \text{and} \quad \kappa := \begin{bmatrix} 2 & 1 & 1 \\ \frac{1}{3} & \frac{2}{3} & 3 \\ \frac{11}{4} & 1 & \frac{1}{4} \end{bmatrix}, \quad \text{so} \quad P_\theta = \frac{1}{3\theta + \theta^3} \begin{bmatrix} 2\theta & \theta & \theta^3 \\ \frac{1}{3}\theta^3 & \frac{2}{3}\theta^3 & 3\theta \\ \frac{11}{4}\theta & \theta^3 & \frac{1}{4}\theta^3 \end{bmatrix}.$$

Further,  $R(\theta) = \{\log \theta, 3 \log \theta\}$ , so

$$\begin{array}{lll} S^1(\theta, \log \theta) = \{1, 3\} & S^2(\theta, \log \theta) = \{1\} & S^3(\theta, \log \theta) = \{2\} \\ S^1(\theta, 3 \log \theta) = \{2\} & S^2(\theta, 3 \log \theta) = \{2, 3\} & S^3(\theta, 3 \log \theta) = \{1, 3\} \\ S_1(\theta, \log \theta) = \{1, 2\} & S_2(\theta, \log \theta) = \{3\} & S_3(\theta, \log \theta) = \{1\} \\ S_1(\theta, 3 \log \theta) = \{3\} & S_2(\theta, 3 \log \theta) = \{1, 2\} & S_3(\theta, 3 \log \theta) = \{2, 3\}. \quad \square \end{array}$$

We now apply similar logic for eigenvectors of  $P_\theta$ . The word *every* in the first sentence is key to applying linear independence in the proof. The importance of the theorem is if it can be applied to reversible MEFs to obtain bounds on the second largest eigenvalue as such bounds would help determine mixing times (see Levin & Peres, 2017, § 12.2).

**Theorem 2.3.18.** *For the MEF defined in eq. (2.8) and every  $\theta \in \Theta$ , suppose  $\mathbf{u} \in \mathbb{C}^{\mathcal{S}}$  and  $\lambda \in \mathbb{C}$  are such that  $\mathbf{u}^\top P_\theta = \lambda \mathbf{u}^\top$ . Then for all  $b \in \mathcal{S}$  and  $\theta \in \Theta$ ,*

$$\lambda u_b = \frac{\sum_{a \in S^b(\theta, r)} u_a \kappa(a, b)}{\sum_{a \in S_b(\theta, r)} \kappa(b, a)} \quad \text{for any } r \in R(\theta).$$

*In particular,  $P_\theta$  is symmetric if and only if*

$$\sum_{a \in S_b(\theta, r)} \kappa(b, a) = \sum_{a \in S^b(\theta, r)} \kappa(a, b) \quad \text{for any } r \in R(\theta)$$

*for all  $b \in \mathcal{S}$  if and only if*

$$\log \frac{\kappa(a, b)}{\kappa(b, a)} = \eta(\theta) \cdot (\tau(a, b) - \tau(b, a))$$

*for all  $a, b \in \mathcal{S}$ .*

*Proof.* The  $b$ th coordinate of the equation  $\mathbf{u}^\top P_\theta = \lambda \mathbf{u}^\top$  is  $\lambda u_b = \sum_{a \in \mathcal{S}} u_a P_\theta(a, b)$ , so, multiplying both sides by  $e^{\zeta(\theta)}$  and plugging in eq. (2.4), we have

$$\lambda u_b \sum_{c \in \mathcal{S}} \kappa(d, c) \exp(\eta(\theta) \cdot \tau(d, c)) = \sum_{a \in \mathcal{S}} u_a \kappa(a, b) \exp(\eta(\theta) \cdot \tau(a, b)),$$

where  $d$  could be any element of  $\mathcal{S}$ , including  $b$ . By the linear independence of the exponential functions, we have

$$\lambda u_b \sum_{c \in S_b(\theta, r)} \kappa(b, c) = \sum_{a \in S^b(\theta, r)} u_a \kappa(a, b)$$

for each  $r \in R(\theta)$ . The sum on the left is never zero by eq. (2.28).

Symmetry in a stochastic matrix happens if and only if the matrix is doubly stochastic if and only if  $\mathbf{1}$  is a left eigenvector corresponding to eigenvalue 1. Plugging these in for  $\mathbf{u}$  and  $\lambda$  respectively gives the second to last part of the theorem. The last part follows from expanding and simplifying  $P_\theta(a, b) = P_\theta(b, a)$ .  $\square$



We end with a theorem of a more statistical flavor. It allows us to dispose of temporal dependence when we just need the *mean parameter*  $\mathbb{E}[\tau(X_t, X_{t+1})]$ , which is crucial for maximum likelihood estimation (Fienberg & Rinaldo, 2012a; E. L. Lehmann & Casella, 1998, Thm. 6.2, pp. 125–126). We will prove a counterpart in theorem 2.4.9.

**Theorem 2.3.19.** *Let  $X = \{X_t\}_{t \in \mathbb{N}}$  be a MEF with transition matrices given by eq. (2.8). Let  $\theta \in \Theta$  such that  $\eta(\theta)$  is in the interior of the natural parameter space and  $\eta$  is differentiable at  $\theta$  with Jacobian matrix  $J_\theta$ . Then, for any  $t \in \mathbb{N}$ ,  $\mathbb{E}_\theta[\tau(X_t, X_{t+1})] = J_\theta^{-1} \nabla \zeta(\theta)$ .*

*Proof.* By Feigin (1981, Thm. 1(i)), if  $\eta$  is the identity, and  $X$ 's transition matrix is more generally the CEF given in eq. (2.7), then  $\mathbb{E}_\theta[\tau(X_t, X_{t+1}) \mid X_0, \dots, X_t] = \nabla \zeta(X_t, \theta)$  for any  $t \in \mathbb{N}$ . If  $\eta$  is not the identity function we apply the chain rule, replacing  $\nabla \zeta(X_t, \theta)$  with  $J_\theta^{-1} \nabla \zeta(X_t, \theta)$ , as in E. L. Lehmann and Casella (1998, Problem 5.6(b), p. 66). However, because  $X$ 's transition matrix is the MEF eq. (2.8),  $\zeta$  is constant with respect to  $X_t$ , so we must replace  $J_\theta^{-1} \nabla \zeta(X_t, \theta)$  with  $J_\theta^{-1} \nabla \zeta(\theta)$ . Taking expectations on both sides we get, for any  $t \in \mathbb{N}$ ,

$$J_\theta^{-1} \nabla \zeta(\theta) = \mathbb{E}_\theta (\mathbb{E}_\theta[\tau(X_t, X_{t+1}) \mid X_0, \dots, X_t]) = \mathbb{E}_\theta[\tau(X_t, X_{t+1})]. \quad \square$$

**2.3.4.1 CEF Representations of MEFs.** By theorem 2.3.11, if a stochastic process  $X = \{X_t\}_{t \in \mathbb{N}}$  is a Markov chain under any of the transition matrices in  $T := \{T_\theta\}_{\theta \in \Theta}$ , then  $T$  is an MEF if and only if  $X$ 's distribution  $L$  has an exponential family representation. The theorem makes explicit how the representations of  $T$  and  $L$  relate. This allows us to recognize an exponential family of distributions from a transition matrix, or an MEF from a distribution. However, this just pushes recognizing an MEF from other CEFs from  $L$  back down to  $T$ . In this subsection, we give some necessary conditions to recognize the MEF  $T$  from theorem 2.3.11 when  $T$  is presented as a CEF. In the scalar-parameter case, proposition 2.3.16 is a powerful tool for this purpose already—see subsection 2.5.4 for some examples of this. The results below are more useful for vector parameters.

Suppose  $T$  is the MEF in eq. (2.16) on page 29:

$$T_\theta(a, b) = \kappa(a, b) \exp(\eta(\theta) \cdot \tau(a, b) - \zeta(\theta)) \quad (2.16 \text{ revisited})$$

with  $T_{\theta_0}(a, b) = \kappa(a, b)$  for some  $\theta_0 \in \Theta$  and all  $a, b \in \mathcal{S}$ . Each row of  $T_{\theta_0}$  is a probability

mass function, so we have that  $\sum_{b \in \mathcal{S}} \kappa(a, b) = 1$  for all  $a \in \mathcal{S}$ . Further, since  $\kappa(a, b) = \kappa(a, b) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}_0) \cdot \boldsymbol{\tau}(a, b) - \zeta(\boldsymbol{\theta}_0))$ , we have that  $\boldsymbol{\eta}(\boldsymbol{\theta}_0) \cdot \boldsymbol{\tau}(a, b) = \zeta(\boldsymbol{\theta}_0)$  for all  $a, b \in \mathcal{S}$ . When  $T$  is a minimal representation,  $\boldsymbol{\tau}$  has affinely independent coordinates, so  $\boldsymbol{\eta}(\boldsymbol{\theta}_0) = \mathbf{0}$  and  $\zeta(\boldsymbol{\theta}_0) = 0$  in such a case.

Further, suppose that for some  $k: \mathcal{S}^2 \rightarrow [0, \infty)$ ,  $\mathbf{h}: \Theta \rightarrow \mathbb{R}^n$ ,  $\mathbf{m}: \mathcal{S}^2 \rightarrow \mathbb{R}^n$ , and  $B: \mathcal{S} \times \Theta \rightarrow \mathbb{R} \cup \{\infty\}$ , we also have the following CEF representation of  $T$ :

$$T_{\boldsymbol{\theta}}(a, b) = k(a, b) \exp(\mathbf{h}(\boldsymbol{\theta}) \cdot \mathbf{m}(a, b) - B(a, \boldsymbol{\theta})) \quad (2.29)$$

for all  $\boldsymbol{\theta} \in \Theta$  and  $a, b \in \mathcal{S}$ . The following results provide necessary conditions to recognize the CEF representation as the MEF it really is. However, the most important follows directly from eqs. (2.16) and (2.29):  $\kappa(a, b) = 0 \iff k(a, b) = 0$  for all  $a, b \in \mathcal{S}$ .

**Lemma 2.3.20.** *If  $n = \ell$ ,  $\mathbf{h} = \boldsymbol{\eta}$  and  $\mathbf{m} = \boldsymbol{\tau}$  then*

$$\frac{e^{B(c, \boldsymbol{\theta})}}{e^{B(a, \boldsymbol{\theta})}} = \frac{\sum_{b \in \mathcal{S}} k(c, b)}{\sum_{b \in \mathcal{S}} k(a, b)}.$$

*Proof.* From the equality of eqs. (2.16) and (2.29), for any  $\boldsymbol{\theta} \in \Theta$  and  $a, b \in \mathcal{S}$  for which  $k(a, b) \neq 0$ , we have  $e^{\zeta(\boldsymbol{\theta})} / e^{B(a, \boldsymbol{\theta})} = \kappa(a, b) / k(a, b)$ . The right side cannot depend on  $\boldsymbol{\theta}$  and the left side cannot depend on  $b$ , so both sides equal a constant  $p(a)$  depending on  $a$ . This way we can write  $p(a)k(a, b) = \kappa(a, b) = T_{\boldsymbol{\theta}_0}(a, b)$ . Hence  $p(a) = 1 / \sum_{b \in \mathcal{S}} k(a, b)$ . For any  $a, c \in \mathcal{S}$ , we may compute the ratio

$$\frac{e^{\zeta(\boldsymbol{\theta})} / e^{B(a, \boldsymbol{\theta})}}{e^{\zeta(\boldsymbol{\theta})} / e^{B(c, \boldsymbol{\theta})}} = \frac{p(a)}{p(c)} = \frac{\sum_{b \in \mathcal{S}} k(c, b)}{\sum_{b \in \mathcal{S}} k(a, b)}. \quad \square$$

**Lemma 2.3.21.** *If  $k = \kappa$ ,  $n = \ell$ ,  $\mathbf{h} = \boldsymbol{\eta}$  is continuous, and  $\boldsymbol{\eta}(\Theta)$  contains an  $\ell$ -dimensional open set, then eq. (2.29) is an MEF representation if and only if  $\boldsymbol{\tau}(a, b) - \mathbf{m}(a, b) = \boldsymbol{\tau}(c, b) - \mathbf{m}(c, b)$  for all  $a, b, c \in \mathcal{S}$ .*

*Proof.* From the equality of eqs. (2.16) and (2.29), for any  $\boldsymbol{\theta} \in \Theta$  and  $a, b \in \mathcal{S}$  for which  $\kappa(a, b) \neq 0$ , we have  $\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot (\boldsymbol{\tau}(a, b) - \mathbf{m}(a, b)) = B(a, \boldsymbol{\theta}) - \zeta(\boldsymbol{\theta})$ . Subtracting this equation with  $a$  replaced with  $c$  from the equation with  $a$  yields that

$$\boldsymbol{\eta}(\boldsymbol{\theta}_i) \cdot (\boldsymbol{\tau}(a, b) - \mathbf{m}(a, b) - \boldsymbol{\tau}(c, b) + \mathbf{m}(c, b)) = B(a, \boldsymbol{\theta}_i) - B(c, \boldsymbol{\theta}_i).$$

for each  $i \in \{1, \dots, \ell\}$ . Let  $A$  be the matrix whose columns are  $\ell$  linearly independent vectors  $\eta(\theta_1), \dots, \eta(\theta_\ell)$ . Let  $\mathbf{d}$  be the vector whose  $i$ th row is  $B(a, \theta_i) - B(c, \theta_i)$ . Then

$$A(\tau(a, b) - \mathbf{m}(a, b) - \tau(c, b) + \mathbf{m}(c, b)) = \mathbf{d}.$$

The columns of  $A$  are linearly independent, so  $\tau(a, b) - \mathbf{m}(a, b) = \tau(c, b) - \mathbf{m}(c, b)$  if and only if  $B(a, \theta_i) = B(c, \theta_i)$  for each  $i \in \{1, \dots, \ell\}$ . However,  $a, b, c \in \mathcal{S}$  were arbitrary, the  $\theta_i$  were arbitrary, and  $\eta$  was continuous. We can perturb the  $\theta_i$ s in an open set in  $\Theta$  to show the foregoing result for all  $\theta$  on the open set. Since  $\eta$  is continuous, so is  $B(a, \cdot)$  for each  $a$ . Thus  $B(a, \cdot) = B(c, \cdot)$  for all  $a$ , and  $c$ , i.e., eq. (2.29) is an MEF representation of  $T$ .  $\square$

Recall from eq. (2.27) that  $S_a(\theta, r) = \{b \in \mathcal{S} \mid \eta(\theta) \cdot \tau(a, b) = r\}$ .

**Lemma 2.3.22.** *If  $k = \kappa$ , then for all  $a \in \mathcal{S}$ ,  $\theta \in \Theta$ , and  $r \in \mathbb{R}$ , then there exists  $q \in \mathbb{R}$  such that  $\{b \in \mathcal{S} \mid \eta(\theta) \cdot \tau(a, b) = r\} = \{b \in \mathcal{S} \mid \mathbf{h}(\theta) \cdot \mathbf{m}(a, b) = q\}$ .*

*Proof.* From the equality of eqs. (2.16) and (2.29), for any  $\theta \in \Theta$  and  $a, b \in \mathcal{S}$  for which  $\kappa(a, b) \neq 0$ ,

$$\eta(\theta) \cdot \tau(a, b) - \mathbf{h}(\theta) \cdot \mathbf{m}(a, b) = \zeta(\theta) - B(a, \theta).$$

Fix  $a \in \mathcal{S}$ ,  $\theta \in \Theta$ , and  $r \in \mathbb{R}$  such that  $\eta(\theta) \cdot \tau(a, b) = r$  for some  $b \in \mathcal{S}$ . Suppose  $c$  is another element of  $\mathcal{S}$  such that  $\eta(\theta) \cdot \tau(a, c) = r$ . Then

$$\eta(\theta) \cdot \tau(a, b) - \mathbf{h}(\theta) \cdot \mathbf{m}(a, b) = \zeta(\theta) - B(a, \theta) = \eta(\theta) \cdot \tau(a, c) - \mathbf{h}(\theta) \cdot \mathbf{m}(a, c).$$

Canceling  $\eta(\theta) \cdot \tau(a, c)$  on both sides yields that  $\mathbf{h}(\theta) \cdot \mathbf{m}(a, b) = \mathbf{h}(\theta) \cdot \mathbf{m}(a, c) =: q$ . Equality of the sets is obtained by swapping the roles of  $\mathbf{m}$  and  $\tau$ .  $\square$

**Lemma 2.3.23.** *If  $k = \kappa$  and  $\mathbf{m} = \tau$  and eq. (2.16) is a minimal representation, then eq. (2.29) is an MEF representation such that  $\mathbf{h} = \eta$  and  $B(a, \theta) = \zeta(\theta)$  for all  $a \in \mathcal{S}$  and  $\theta \in \Theta$ .*

*Proof.* From the equality of eqs. (2.16) and (2.29), for any  $\theta \in \Theta$  and  $a, b \in \mathcal{S}$  for which  $\kappa(a, b) \neq 0$ ,

$$(\eta(\theta) - \mathbf{h}(\theta)) \cdot \tau(a, b) = \zeta(\theta) - B(a, \theta).$$

The right side is constant with respect to  $b$ , so the minimality of eq. (2.16) implies that  $\zeta(\theta) = B(a, \theta)$  and  $\eta(\theta) = \mathbf{h}(\theta)$  for all  $a \in \mathcal{S}$  and  $\theta \in \Theta$ .  $\square$

## 2.4 Permutation-Uniform Markov Chains

In this section, we show that if a transition matrix is *permutation uniform*, meaning that every row is a permutation of every other row, then we may identify the corresponding Markov chain with an IID sequence on the same state space. This identification, which is the main theorem of this section and is presented in theorem 2.4.5, preserves the functional form of the transition matrix when it is transformed to the distribution for the IID sequence. In this way, we can translate an MEF's likelihood function to an exponential family distribution for an IID sequence, perform statistical analysis on that IID sequence, and draw conclusions about the Markov chain. Autoregressive processes on discrete state spaces and several of the examples in section 2.5 provide applications for this technique. Shalizi and Rinaldo (2013, § 6) calls for ways of analyzing independent random variables in place of dependent random variables in Markov chains of networks. The novelty of the techniques in this paper are disposing of the temporal dependence in a Markov chain, and maintaining interpretability of parameters and sufficient statistics while doing so.

The main concept in this section is *permutation uniformity*, the property that “every row is a permutation of every other row.” We make this rough definition precise as follows. Let  $D$  be any set. For each  $a \in \mathcal{S}$ , define the  $a$ th *row* of a function or matrix  $f: \mathcal{S} \times \mathcal{S} \rightarrow D$  to be the function or row vector  $b \mapsto f(a, b)$ . We write permutations on  $\mathcal{S}$  (i.e., bijections  $\mathcal{S} \rightarrow \mathcal{S}$ ) juxtaposed with other permutations on  $\mathcal{S}$  to denote composition and juxtaposed with elements of  $\mathcal{S}$  to denote application. Let  $\pi := \{\pi_a\}_{a \in \mathcal{S}}$  be a set of permutations on  $\mathcal{S}$ . We say that  $f$  is *permutation uniform*, or *p-uniform*, under  $\pi$  if  $f(a, \pi_a^{-1}c) = f(b, \pi_b^{-1}c)$  for all  $a, b, c \in \mathcal{S}$ . We say that a Markov chain is a *permutation-uniform Markov chain*, or a *p-uniform chain*, if its transition matrix is p-uniform.

**Example 2.4.1.** The following are examples of p-uniform matrices or functions.

$$\begin{bmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{bmatrix}, \quad \begin{bmatrix} 2/7 & 4/7 & 1/7 \\ 4/7 & 1/7 & 2/7 \\ 1/7 & 2/7 & 4/7 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} & \begin{bmatrix} 1 \\ 6 \end{bmatrix} & \begin{bmatrix} 9 \\ 2 \end{bmatrix} \\ \begin{bmatrix} 9 \\ 2 \end{bmatrix} & \begin{bmatrix} 3 \\ 4 \end{bmatrix} & \begin{bmatrix} 1 \\ 6 \end{bmatrix} \\ \begin{bmatrix} 1 \\ 6 \end{bmatrix} & \begin{bmatrix} 3 \\ 4 \end{bmatrix} & \begin{bmatrix} 9 \\ 2 \end{bmatrix} \end{bmatrix}. \quad \square$$

Generalizing the definition of *p-uniform* from Yano and Yasutomi (2011b, Def. 1.4),

we can characterize  $p$ -uniform transition matrices as follows.

**Lemma 2.4.2.** *A function  $f: \mathcal{S} \times \mathcal{S} \rightarrow D$  for some set  $D$  is  $p$ -uniform if and only if there exists a function  $g: \mathcal{S} \rightarrow D$  and a set  $\{\pi_a\}_{a \in \mathcal{S}}$  of permutations such that*

$$f(a, b) = g(\pi_a b) \quad \text{for all } a, b \in \mathcal{S}. \quad (2.30)$$

*In this case,  $g$  is unique up to permutation.*

*If  $\mathcal{S}$  is at most countable and  $f$  is a  $p$ -uniform stochastic matrix on  $\mathcal{S}$ , then  $g$  is a stochastic vector, i.e., a probability mass function.*

*Proof.* Equation (2.30) follows directly from the definition of permutation uniformity.

Suppose  $g, h: \mathcal{S} \rightarrow D$  and  $\{\pi_a\}_{a \in \mathcal{S}}$  and  $\{s_a\}_{a \in \mathcal{S}}$  are sets of permutations such that  $f(a, b) = g(\pi_a b) = h(s_a b)$  for all  $a, b$ . Then for any  $b \in \mathcal{S}$ ,  $g(b) = g(\pi_a \pi_a^{-1} b) = f(a, \pi_a^{-1} b) = h(s_a \pi_a^{-1} b)$ . That is,  $g$  is the composition of  $h$  and the permutation  $s_a \pi_a^{-1}$ .

The last part of the proof follows from the definition of a stochastic matrix.  $\square$

The selection of the permutations  $\pi_a$  is unique up to permutation only trivially in the sense that every permutation is the permutation of another permutation. When  $\mathcal{S}$  is finite and every row of a transition matrix  $P$  is the uniform distribution  $\mathcal{S}$ , then any choice of permutations will satisfy  $P_{ab} = 1/|\mathcal{S}|$ .

**Example 2.4.3** (Continuation of example 2.4.1). Consider the arrays in example 2.4.1. For each one, the function  $g$  of lemma 2.4.2 could be the first row of the array and  $\pi_1$  could be the identity permutation. Then the permutation vectors for the second or third rows of the respective arrays in example 2.4.1 are

$$(2, 1), \quad \begin{pmatrix} (2, 3, 1) \\ (3, 1, 2) \end{pmatrix}, \quad \begin{pmatrix} (1, 2, 3) \\ (1, 2, 3) \end{pmatrix}, \quad \begin{pmatrix} (3, 1, 2) \\ (2, 1, 3) \end{pmatrix}. \quad \square$$

The following lemma will be useful in subsection 2.5.2 when we discuss finite exchangability.

**Lemma 2.4.4.** *Suppose  $\mathcal{S}$  bears an equivalence relation  $\sim$ , and let  $f$  and  $g$  be as they are in lemma 2.4.2. Let  $a, b$ , and  $c$  generically represent elements of  $\mathcal{S}$ . Consider the following conditions:*

$$(a) \ b \sim c \implies g(b) = g(c). \qquad (c) \ b \sim c \implies \pi_a b \sim \pi_a c \text{ for all } a.$$

$$(b) \ b \sim c \implies f(a, b) = f(a, c) \text{ for all } a. \qquad (d) \ b \sim c \implies \pi_a^{-1} b \sim \pi_a^{-1} c \text{ for all } a.$$

Then (c)'s converse  $\iff$  (d); (d)'s converse  $\iff$  (c); (b) and (d)  $\implies$  (a); (a) and (c)  $\implies$  (b); (b)'s converse and (a)  $\implies$  (d); and (a)'s converse and (b)  $\implies$  (c). If  $\pi$  is closed under inversion or if  $\mathcal{S}$  is finite, then (c)  $\iff$  (d). Finally, if there is an  $a$  for which  $b \sim c \implies f(a, b) = f(a, c)$  and (c) and (d) both hold, then (b) holds.

*Proof.* Assume (c)'s converse and  $b \sim c$ . Then  $\pi_a \pi_a^{-1} b \sim \pi_a \pi_a^{-1} c$  for all  $a$ . Condition (c)'s converse implies  $\pi_a^{-1} b \sim \pi_a^{-1} c$  for all  $a$ . Hence (d).

Assume (d) and  $\pi_a b \sim \pi_a c$  for all  $a$ . Condition (d) implies  $\pi_d^{-1} \pi_a b \sim \pi_d^{-1} \pi_a c$  for all  $a$  for all  $d \in \mathcal{S}$ , including  $d = a$ , so  $\pi_a^{-1} \pi_a b \sim \pi_a^{-1} \pi_a c$ , and so  $b \sim c$ . Hence (c)'s converse.

Assume (d)'s converse and  $b \sim c$ . Then  $\pi_a^{-1} \pi_a b \sim \pi_a^{-1} \pi_a c$  for all  $a$ . Condition (d)'s converse implies  $\pi_a b \sim \pi_a c$  for all  $a$ . Hence (c).

Assume (c) and  $\pi_a^{-1} b \sim \pi_a^{-1} c$  for all  $a$ . Condition (c) implies  $\pi_d \pi_a^{-1} b \sim \pi_d \pi_a^{-1} c$  for all  $a$  for all  $d \in \mathcal{S}$ , including  $d = a$ , so  $\pi_a \pi_a^{-1} b \sim \pi_a \pi_a^{-1} c$ , and so  $b \sim c$ . Hence (d)'s converse.

Assume (b) and (d) and  $b \sim c$ . Condition (d) implies  $\pi_a^{-1} b \sim \pi_a^{-1} c$  for all  $a$ . Equation (2.30) and condition (b) then imply  $g(b) = f(a, \pi_a^{-1} b) = f(a, \pi_a^{-1} c) = g(c)$  for all  $a$ . Hence (a).

Assume (a) and (c) and  $b \sim c$ . Condition (c) implies  $\pi_a b \sim \pi_a c$  for all  $a$ . Equation (2.30) and condition (a) then imply  $f(a, b) = g(\pi_a b) = g(\pi_a c) = f(a, c)$  for all  $a$ . Hence (b).

Assume (b)'s converse and (a) and  $b \sim c$ . Equation (2.30) and condition (a) imply  $f(a, \pi_a^{-1} b) = g(b) = g(c) = f(a, \pi_a^{-1} c)$  for all  $a$ . The converse of (b) then implies  $\pi_a^{-1} b \sim \pi_a^{-1} c$  for all  $a$ . Hence (d).

Assume (a)'s converse and (b) and  $b \sim c$ . Equation (2.30) and condition (b) imply  $g(\pi_a b) = f(a, b) = f(a, c) = g(\pi_a c)$  for all  $a$ . The converse of (a) then implies  $\pi_a b \sim \pi_a c$  for all  $a$ . Hence (c).

Suppose  $\pi$  is closed under inverses, and assume (c) and  $b \sim c$ . Since  $\pi$  contains its inverses, to each  $a$  belongs some  $d(a) \in \mathcal{S}$  such that  $\pi_{d(a)} = \pi_a^{-1} \iff \pi_{d(a)}^{-1} = \pi_a$ . Condition (c) implies  $\pi_a b \sim \pi_a c$  for all  $a$ . Therefore  $\pi_{d(a)}^{-1} b \sim \pi_{d(a)}^{-1} c$  for all  $a$ . Suppose by way

of contradiction that there is some  $e \in \mathcal{S}$  such that  $\pi_e^{-1} \neq \pi_{d(a)}^{-1}$  for all  $a$ . But  $\pi_{d(e)} = \pi_e^{-1}$  and  $\pi_{d(d(e))} = \pi_{d(e)}^{-1} = \pi_e$ . Thus  $\pi_e^{-1} = \pi_{d(a)}^{-1}$  for  $a = d(e)$ , contradicting our choice of  $e$ . Therefore,  $a \mapsto \pi_{d(a)}^{-1}$  is surjective onto  $\{\pi_a^{-1}\}_{a \in \mathcal{S}}$ , and  $\pi_a^{-1}b \sim \pi_a^{-1}c$  for all  $a$ . Hence (d).

Suppose  $\pi$  is closed under inverses, and assume (d) and  $b \sim c$ . Condition (d) implies  $\pi_a^{-1}b \sim \pi_a^{-1}c$  for all  $a$ . Using  $d$  as defined in the previous paragraph, we have  $\pi_{d(a)}^{-1}b \sim \pi_{d(a)}^{-1}c$  for all  $a$ , and therefore  $\pi_a b \sim \pi_a c$  for all  $a$ . Hence (c).

Suppose  $\mathcal{S}$  is finite, and assume (c) and  $b \sim c$ . Condition (c) implies  $\pi_a^n b \sim \pi_a^n c$  for all  $a$  and all  $n \in \mathbb{N}$  (by induction on  $n$ ). Since  $\mathcal{S}$  is finite, to each  $a$  belongs some  $n(a) \in \mathbb{N}$  such that  $\pi_a^{n(a)} = \pi_a^{-1}$  (Clark, 1971/1984, paras. 35, 41). Therefore  $\pi_a^{-1}b \sim \pi_a^{-1}c$  for all  $a$ . Hence (d).

Suppose  $\mathcal{S}$  is finite, and assume (d) and  $b \sim c$ . Condition (d) implies  $\pi_a^{-n}b \sim \pi_a^{-n}c$  for all  $a$  and all  $n \in \mathbb{N}$ . Using  $n$  as defined in the previous paragraph, we have, for each  $a$ ,  $\pi_a^{-n(a)} = \left(\pi_a^{n(a)}\right)^{-1} = \left(\pi_a^{-1}\right)^{-1} = \pi_a$ . Therefore  $\pi_a b \sim \pi_a c$  for all  $a$ . Hence (c).

Finally, assume (c) and (d) and there is an  $a$  for which  $b \sim c \implies f(a, b) = f(a, c)$ . If  $b \sim c$ , then  $\pi_a^{-1}b \sim \pi_a^{-1}c$  by (d), and eq. (2.30) implies  $g(b) = f(a, \pi_a^{-1}b) = f(a, \pi_a^{-1}c) = g(c)$ ; hence (a). Then (a) and (c)  $\implies$  (b).  $\square$

**2.4.1 Literature Review.** Rosenblatt (1959, § 3) first named and studied  $p$ -uniform chains, calling them simply *uniform chains*.<sup>16</sup> Rosenblatt showed that when a  $p$ -uniform chain  $X$  on a countable state space with transition matrix  $P$  has distinct numbers in every entry of a row of  $P$ ,  $X_i$  and  $X_{i+1}$  almost surely uniquely determine a random variable  $Z_{i+1}$ , and the  $Z_i$ s are IID and independent of  $X_{i-1}, X_{i-2}, \dots$  (Rosenblatt, 1959, Lem. 3).

Rosenblatt (1960) constructs an IID sequence of random variables  $Z_t$ , uniformly distributed on the interval  $[0, 1]$ , such that  $X_t$  is a function of all the preceding  $Z_i$ s. Under the conditions of the theorem, this is applicable to a wider variety of Markov chains than theorem 2.4.5, but offers less control over the  $Z$  process. In Rosenblatt's construction,  $X$  can be any discrete-valued, irreducible, aperiodic Markov chain indexed by time set  $\mathbb{Z}$  rather than  $\mathbb{N}$ , meaning there's no initial value. The infinitude of the past is crucial to the proof, as it relies on the Borel-Cantelli lemma. However, since the  $Z_t$ s are  $[0, 1]$ -valued, they are

---

<sup>16</sup>We add the  $p$ - to  $p$ -uniform both to follow Yano and Yasutomi (2011b, Def. 1.4) and to avoid confusion with uniform distributions. The  $p$  stands for *permutation*.

defined on a different sample space from the  $X_t$ s, and the function that links the  $Z_t$ s to the  $X_t$ s is not constructed, as its existence is inferred from the some intermediate limit results. Wu and Mielniczuk (2010, § 4) gives a more readable description of the theory. In the extension of this result to continuous state spaces given by Hanson (1963), indexing time by  $\mathbb{Z}$  rather than  $\mathbb{N}$  turns out to be crucial. Blum and Hanson (1963) and Rosenblatt (1963) give further results in this vein, and Laurent (2010, Problem 1) summarizes some of the results.

Diaconis and Freedman (1999) gives a model of Markov chains induced by iterating functions from the state space to itself chosen randomly independently. We will explore this connection in some detail in subsection 2.4.2.

Rubshtein (2004) classifies Markov shifts of  $p$ -uniform Markov chains, which the author calls  *$p$ -uniform stochastic graphs*, in the language of dynamic systems.

**2.4.2 Independence.** In this subsection, we derive IID  $Z_t$ s uniquely determined by a  $p$ -uniform Markov chain  $X$  with transition matrix  $P_{\theta_0}$  for some fixed  $\theta_0 \in \Theta$ . In contrast with our inspiration for this technique, Rosenblatt (1959, Lem. 3), we can drop “distinct” and the “uniquely” from the requirements of the transition probabilities, and our input Markov chain  $X$  does not go infinitely into the past.

Our  $Z_t$ s’ distributions can stand in for the  $X_t$ s’, thus making statistical inference on the time-dependent  $X_t$ s easier by using the independent  $Z_t$ s. This makes the Markov chain  $X$  resemble a random walk. We explore this connection more after the main theorems.

$X$ ’s being  $p$ -uniform imposes enough structure for us to observe the desired IID sequence  $Z$  from  $X$ , which we mean in the following sense. We say that the  $\mathcal{S}$ -valued sequence of random variables  $Z_1, \dots, Z_t, t \in \mathbb{N}_{>0}$ , has a joint distribution *similar* under  $\mathbb{P}$  to that of  $\mathcal{S}$ -valued sequence of random variables  $X_0, X_1, \dots, X_t$ ’s for some probability measure  $\mathbb{P}$  if and for all  $x_0, x_1, \dots, x_t \in \mathcal{S}$  there exists  $z_1, \dots, z_t \in \mathcal{S}$  such that

$$\mathbb{P}(X_t = x_t, \dots, X_1 = x_1 \mid X_0 = x_0) = \mathbb{P}(Z_t = z_t, \dots, Z_1 = z_1).$$

For  $\{Z_t\}_{t=1}^{\infty}$  to be IID with a *common law* (whose PMF under  $\mathbb{P}$  is)  $\mu$  means that  $\mathbb{P}(Z_t = z) = \mu(z)$  for all  $t \in \mathbb{N}_{>0}$  and all  $z \in \mathcal{S}$ . As has been our wont, we do not worry about the



(common) sample probability space for  $X$  and  $Z$  or the nature of the measure  $\mathbb{P}$ , but rather just assume that some such  $\mathbb{P}$  is fixed for the current discussion. In the parameterized case such as when  $X$  is a Markov chain whose transition matrix is drawn from a CEF, precisely which  $\mathbb{P}$  we are discussing will depend on the parameter. The next theorem does not assume a CEF.

**Theorem 2.4.5.** *Suppose  $\mathcal{S}$  is at most countable. Let  $\pi := \{\pi_a\}_{a \in \mathcal{S}}$  be a set of permutations on  $\mathcal{S}$ , and let  $X := \{X_t\}_{t \in \mathbb{N}}$  be an  $\mathcal{S}$ -valued stochastic process.  $X$  is a Markov chain on  $\mathcal{S}$  under probability measure  $\mathbb{P}$  with transition matrix  $P$   $p$ -uniform under  $\pi$  if and only if there exists a probability measure  $\mu$  on  $\mathcal{S}$ , a sequence of  $\mathcal{S}$ -valued random variables  $Z := \{Z_t\}_{t=1}^\infty$ , and an  $\mathcal{S}$ -valued random variable  $X_0$  such that*

- (a)  $Z$  is IID with common law  $\mu$ ;
- (b)  $X_{t+1} = \pi_{X_t}^{-1} Z_{t+1}$   $\mathbb{P}$ -almost surely for all  $t \in \mathbb{N}$ ;
- (c)  $P(a, b) = \mu(\pi_a b)$  for all  $a, b \in \mathcal{S}$ ; and
- (d)  $X_0, Z_1, \dots, Z_t$  are mutually independent for all  $t \in \mathbb{N}_{>0}$ .

When such a  $Z$  exists,  $X$  and  $Z$  have similar joint distributions. Further, for all  $t \in \mathbb{N}$ ,  $Z_{t+1}$  and any random vector of the form  $(X_{i_1}, \dots, X_{i_n}, X_t)$  for  $i_1, \dots, i_n \in \{0, 1, \dots, t-1\}$  and  $n \in \{0, 1, \dots, t-1\}$  are pairwise independent.

If we are given a  $p$ -uniform Markov chain  $X$  with transition matrix  $P$ , lemma 2.4.2 already tells us what  $\mu$  has to be: any permutation of any row of  $P$ . We are free to choose any permutation of a row of  $P$  that is convenient because we can choose any permutations  $\{\pi_a\}_{a \in \mathcal{S}}$  that make  $P(a, b) = \mu(\pi_a b)$  true. In this sense, once we're given  $P$ , we have no choice over  $\mu$ ; we can only choose the permutations  $\pi_a$ . The choice of permutations  $\pi_a$  determines the sequence  $Z$  almost surely from  $X$ . Applying theorem 2.4.5 thus comes down to choosing  $\pi_a$ s. Conversely if we are given  $Z$  with the common law  $\mu$ , different choices of  $\pi_a$ s give rise to different Markov chains  $X$ . We will see examples of this phenomenon in section 2.5.

In the proof below, we can define random variables as we see fit because we have implicitly assumed the use of the discrete  $\sigma$ -algebras on  $\mathcal{S}$  and  $\mathcal{S} \times \mathcal{S}$ . The function  $(a, b) \mapsto \pi_a b$  is measurable with respect to the discrete  $\sigma$ -algebra on  $\mathcal{S} \times \mathcal{S}$ .

*Proof of theorem 2.4.5. Forward Implication.* Suppose  $X$  is a Markov chain on  $\mathcal{S}$  under probability measure  $\mathbb{P}$  with transition matrix  $P$  p-uniform under  $\pi$ .

Lemma 2.4.2 provides for the existence of a probability measure  $\mu$  on  $\mathcal{S}$  such that  $P(a, b) = \mu(\pi_a b)$ , condition (c).  $X_0$  is an  $\mathcal{S}$ -valued random variable because  $X$  is a Markov chain.

We can define the  $\mathcal{S}$ -valued random variables  $Z = \{Z_t\}_{t \in \mathbb{N}_{>0}}$  by

$$Z_{t+1} := \pi_{X_t} X_{t+1} \quad \text{for all } t \in \mathbb{N}.$$

This equality holds everywhere in the sample space and  $\pi_a$  is invertible for all  $a \in \mathcal{S}$ , so  $X_{t+1} = \pi_{X_t}^{-1} Z_{t+1}$  holds  $\mathbb{P}$ -almost surely for all  $t \in \mathbb{N}$ , establishing condition (b). In the remainder of this direction of the proof, we will make use of the equivalence between  $Z_t = z$  and  $X_t = \pi_{X_{t-1}}^{-1} z$  when  $t \in \mathbb{N}_{>0}$ .

Now we must show that  $Z$  is IID with common law  $\mu$ , which is condition (a). The work to prove it will also prove condition (d). To do so, we fix an arbitrary sequence  $\{z_t\}_{t \in \mathbb{N}_{>0}} \subseteq \mathcal{S}$  and time  $T \in \mathbb{N}_{>0}$ , and we show that

$$\mathbb{P}(Z_1 = z_1, \dots, Z_T = z_T) = \mu(z_1) \cdots \mu(z_T). \quad (2.31)$$

For each  $t \in [T]$ ,  $Z_t = z_t$  determines a transition. Stringing together all  $T$  transitions determines the state of the Markov chain through time  $T$  if we know the starting position  $X_0$ . This suggests starting the proof with the left-hand-side expression in eq. (2.31) and using the law of total probability:

$$\mathbb{P}(Z_1 = z_1, \dots, Z_T = z_T) = \sum_{a \in \mathcal{S}} \mathbb{P}(Z_1 = z_1, \dots, Z_T = z_T \mid X_0 = a) \mathbb{P}(X_0 = a). \quad (2.32)$$

We can formalize what we meant by “stringing together” the transitions as follows. For each  $t \in \mathbb{N}$ , let  $x_t : \mathcal{S} \rightarrow \mathcal{S}$  be defined recursively by

$$x_t(a) := \begin{cases} a & t = 0 \\ \pi_{x_{t-1}(a)}^{-1} z_t & t > 0. \end{cases}$$

For a given  $a \in \mathcal{S}$ ,  $\{x_t(a)\}_{t \in \mathbb{N}}$  is a deterministic sequence of elements of  $\mathcal{S}$ . Let  $a \in \mathcal{S}$  such that  $\mathbb{P}(X_0 = a) > 0$ .

As we now prove by induction, the event  $A$  in which  $Z_1 = z_1, \dots, Z_T = z_T$ , and  $X_0 = a$  is almost surely the same event as the event  $B$  in which  $X_0 = x_0(a), \dots, X_T = x_T(a)$ . Let  $t \in [T]$ . First suppose we are in event  $A$ , so  $Z_t = z_t$  and  $X_0 = a = x_0(a)$ . Under the induction hypothesis that  $X_{t-1} = x_{t-1}(a)$ , we have  $X_t = \pi_{X_{t-1}}^{-1} Z_t = \pi_{x_{t-1}(a)}^{-1} z_t = x_t(a)$ . Thus we are in event  $B$  as well. Second suppose we are in event  $B$ , so  $X_0 = x_0(a) = a$ ,  $X_{t-1} = x_{t-1}(a)$ , and  $X_t = x_t(a)$ . Then  $Z_t = \pi_{X_{t-1}} X_t = \pi_{x_{t-1}(a)} x_t(a) = \pi_{x_{t-1}(a)} \pi_{x_{t-1}(a)}^{-1} z_t = z_t$ . Thus we are in event  $A$  as well. Consequently, we may expand the left term in the summand of eq. (2.32) using these equivalent events for  $a \in \mathcal{S}$ :

$$\mathbb{P}(Z_1 = z_1, \dots, Z_T = z_T \mid X_0 = a) = \mathbb{P}(X_1 = x_1(a), \dots, X_T = x_T(a) \mid X_0 = a) \quad (2.33)$$

$X$  is a Markov chain so we can expand the right side of eq. (2.33) as

$$\mathbb{P}(X_1 = x_1(a), \dots, X_T = x_T(a) \mid X_0 = a) = \prod_{t=1}^T P(x_{t-1}(a), x_t(a)). \quad (2.34)$$

Applying condition (c), we can then write for any  $t \in [T]$

$$P(x_{t-1}(a), x_t(a)) = \mu(\pi_{x_{t-1}(a)} x_t(a)) = \mu(\pi_{x_{t-1}(a)} \pi_{x_{t-1}(a)}^{-1} z_t) = \mu(z_t). \quad (2.35)$$

Putting together eqs. (2.33) to (2.35), we get

$$\mathbb{P}(Z_1 = z_1, \dots, Z_T = z_T \mid X_0 = a) = \prod_{t=1}^T \mu(z_t). \quad (2.36)$$

This is *almost* what we need. Combining eq. (2.36) with eq. (2.32) yields our goal from eq. (2.31):

$$\begin{aligned} \mathbb{P}(Z_1 = z_1, \dots, Z_T = z_T) &= \sum_{a \in \mathcal{S}} \mathbb{P}(Z_1 = z_1, \dots, Z_T = z_T \mid X_0 = a) \mathbb{P}(X_0 = a) \\ &= \sum_{a \in \mathcal{S}} \left[ \prod_{t=1}^T \mu(z_t) \right] \mathbb{P}(X_0 = a) = \prod_{t=1}^T \mu(z_t) \sum_{a \in \mathcal{S}} \mathbb{P}(X_0 = a) = \prod_{t=1}^T \mu(z_t). \end{aligned}$$

This establishes condition (a). Between eqs. (2.31) and (2.36), we see that  $X_0, Z_1, \dots, Z_T$  are independent, establishing condition (d).

**Backward Implication.** We must show that  $X = \{X_t\}_{t \in \mathbb{N}}$ , defined in condition (b), is a Markov chain with a transition matrix  $P$ , defined in condition (c), and that  $P$  is p-uniform

under  $\pi$ . To do so, we fix an arbitrary time  $t \in \mathbb{N}$  and vector  $\boldsymbol{x} = (x_0, \dots, x_{t+1}) \in \mathcal{S}^{t+2}$  such that  $\mathbb{P}(X_t = x_t) > 0$ , and we show that

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, \dots, X_t = x_t) = \mu(\pi_{x_t} x_{t+1}) \quad (2.37)$$

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t) = \mu(\pi_{x_t} x_{t+1}) \quad (2.38)$$

Proving eqs. (2.37) and (2.38) together will show that  $X$  has the Markov property, and proving eq. (2.38) will show that  $X$  is p-uniform by lemma 2.4.2 and eq. (2.30). Since the time  $t$  is arbitrary, proving eq. (2.38) will also establish that  $X$  is homogeneous, so that  $P(a, b) = \mu(\pi_a b)$  is the transition matrix for  $X$ . By lemma 2.4.2, this will establish that  $P$  is p-uniform under  $\pi$ .

As we now prove by induction, the event  $C$  in which  $X_0 = x_0, \dots, X_T = x_T$  (on which the left side of eq. (2.37) is conditioned) is, for all  $T \in \mathbb{N}$ , almost surely the same event as event  $D$  in which  $X_0 = x_0$  and  $Z_1 = \pi_{x_0} x_1, \dots, Z_T = \pi_{x_{T-1}} x_T$ . We prove this for the case when  $T > 0$ . Let  $s \in [T]$ . First suppose we are in event  $C$ , so  $X_{s-1} = x_{s-1}$ ,  $X_s = x_s$ , and  $X_0 = x_0$ . We have  $x_s = X_s = \pi_{X_{s-1}}^{-1} Z_s = \pi_{x_{s-1}}^{-1} Z_s$  a.s., so  $Z_s = \pi_{x_{s-1}} x_s$  a.s. Thus we are in event  $D$  as well. Second suppose we are in event  $D$ , so  $Z_s = \pi_{x_{s-1}} x_s$ ,  $Z_1 = \pi_{x_0} x_1$ , and  $X_0 = x_0$ . Under the induction hypothesis that  $X_{s-1} = x_{s-1}$  a.s., we have  $X_s = \pi_{X_{s-1}}^{-1} Z_s = \pi_{x_{s-1}}^{-1} \pi_{x_{s-1}} x_s = x_s$  a.s. Thus we are in event  $C$  as well.

We prove eq. (2.37) as follows. The first equality uses  $C = D$  a.s. for  $T = t$ , and the second uses  $Z$ 's PMF  $\mu$  from condition (a) and the independence of  $X_0, Z_1, \dots, Z_{t+1}$  from condition (d).

$$\begin{aligned} \mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, X_1 = x_1, \dots, X_t = x_t) \\ = \mathbb{P}(Z_{t+1} = \pi_{x_t} x_{t+1} \mid X_0 = x_0, Z_1 = \pi_{x_0} x_1, \dots, Z_t = \pi_{x_{t-1}} x_t) = \mu(\pi_{x_t} x_{t+1}). \end{aligned}$$

We prove eq. (2.38) as follows. The first and third equalities below apply  $C = D$

a.s. for  $T = t - 1$ , and the second uses conditions (a) and (d).

$$\begin{aligned}
& \mathbb{P}(X_{t+1} = x_{t+1}, X_t = x_t \mid X_0 = x_0, X_1 = x_1, \dots, X_{t-1} = x_{t-1}) \\
&= \mathbb{P}(Z_{t+1} = \pi_{x_t} x_{t+1}, Z_t = \pi_{x_{t-1}} x_t \mid X_0 = x_0, Z_1 = \pi_{x_0} x_1, \dots, Z_{t-1} = \pi_{x_{t-2}} x_{t-1}) \\
&= \mu(\pi_{x_t} x_{t+1}) \mathbb{P}(Z_t = \pi_{x_{t-1}} x_t \mid X_0 = x_0, Z_1 = \pi_{x_0} x_1, \dots, Z_{t-1} = \pi_{x_{t-2}} x_{t-1}) \\
&= \mu(\pi_{x_t} x_{t+1}) \mathbb{P}(X_t = x_t \mid X_0 = x_0, X_1 = x_1, \dots, X_{t-1} = x_{t-1}).
\end{aligned}$$

Apply the law of total probability with fixed  $x_t$  and  $x_{t+1}$  and sum over  $\mathbf{x} = (x_0, \dots, x_{t-1})$ :

$$\begin{aligned}
& \mathbb{P}(X_{t+1} = x_{t+1}, X_t = x_t) \\
&= \sum_{\mathbf{x} \in \mathcal{S}^t} \mathbb{P}(X_{t+1} = x_{t+1}, X_t = x_t \mid X_0 = x_0, \dots, X_{t-1} = x_{t-1}) \mathbb{P}(X_0 = x_0, \dots, X_{t-1} = x_{t-1}) \\
&= \mu(\pi_{x_t} x_{t+1}) \sum_{\mathbf{x} \in \mathcal{S}^t} \mathbb{P}(X_t = x_t \mid X_0 = x_0, \dots, X_{t-1} = x_{t-1}) \mathbb{P}(X_0 = x_0, \dots, X_{t-1} = x_{t-1}) \\
&= \mu(\pi_{x_t} x_{t+1}) \mathbb{P}(X_t = x_t).
\end{aligned}$$

Dividing both sides by  $\mathbb{P}(X_t = x_t)$  yields eq. (2.38), as desired.

**Finishing the Proof.** The fact that  $C = D$  a.s. proves  $X$  and  $Z$  have similar distributions. Finally, to prove that  $Z_{t+1}$  and  $(X_{i_1}, \dots, X_{i_n}, X_t)$  are independent, let  $z \in \mathcal{S}$  and observe that

$$\begin{aligned}
& \mathbb{P}(Z_{t+1} = z \mid X_t = x_t, X_{i_n} = x_{i_n}, \dots, X_{i_1} = x_{i_1}) \\
&= \mathbb{P}(X_{t+1} = \pi_{x_t}^{-1} z \mid X_t = x_t) = \mu(\pi_{x_t} \pi_{x_t}^{-1} z) = \mathbb{P}(Z_{t+1} = z). \quad \square
\end{aligned}$$

Diaconis and Freedman (1999) proposed a model of Markov chains induced by random functions, of which theorem 2.4.5 turns out to be a sub-model. Let  $\Omega$  be an arbitrary set. Fix some family  $F = \{f_\omega\}_{\omega \in \Omega}$  of functions from  $\mathcal{S}$  to itself indexed by  $\Omega$ . Let  $\mu$  be a probability measure defined on  $\Omega$ . A Markov chain *induced by*  $F$  and  $\mu$  is a process  $X = \{X_t\}_{t \in \mathbb{N}}$  starting at  $X_0 = x_0 \in \mathcal{S}$  and such that, for  $t \in \mathbb{N}$ ,  $X_{t+1} = f_{W_{t+1}}(X_t)$ , where  $W = \{W_t\}_{t \in \mathbb{N}}$  is an IID sequence of  $\Omega$ -valued random variables with common law  $\mu$ . When  $\Omega$  is discrete, we have that

$$\mathbb{P}(X_{t+1} = b \mid X_t = a) = \mathbb{P}(f_{W_{t+1}}(a) = b) = \sum_{\substack{\omega \in \Omega \\ f_\omega(a) = b}} \mu(\{\omega\}).$$

Adopting the notation of theorem 2.4.5, let  $X$  be a  $p$ -uniform Markov chain with permutations  $\{\pi_a\}_{a \in \mathcal{S}}$  and IID sequence  $Z$  with common law  $\mu$ . Setting  $\Omega = \mathcal{S}$  and  $f_a(b) = \pi_b^{-1}a$ , we have that  $X$  is induced by  $F$  and  $\mu$ :

$$X_{t+1} = \pi_{X_t}^{-1}Z_{t+1} = f_{Z_{t+1}}(X_t).$$

In this case,  $\{\omega \in \Omega \mid f_\omega(a) = b\} = \{\pi_a b\}$ , so that  $\mathbb{P}(X_{t+1} = b \mid X_t = a) = \mu(\pi_a b)$  as desired under  $p$ -uniformity.

$P$ -uniformity forces us to restrict the functions  $f_a$ . Since the  $\pi_a$ s are bijective,  $a \mapsto f_a$  is injective in the set of functions  $\mathcal{S} \rightarrow \mathcal{S}$ , and  $a \mapsto f_a(b)$  is bijective in  $\mathcal{S}$ . The latter assertion follows straight from the definition of  $f_a$ ; to see the former, note that

$$f_a = f_b \iff f_a(c) = f_b(c) \forall c \iff \pi_c^{-1}a = \pi_c^{-1}b \forall c \iff a = b.$$

**Example 2.4.6** (Modular Autoregressive Model (Diaconis & Freedman, 1999, Example 6.2, p. 66)). Let  $\mathcal{S} = \mathbb{Z}/n\mathbb{Z}$  be the set of  $n - 1$  integers modulo  $n$ . Fix some initial state  $x_0 \in \mathcal{S}$  and set  $X_0 = x_0$ . Then define  $X_{t+1} = X_t + Z_{t+1} \pmod{n}$ , where the  $Z_t$ s are uniform, IID random variables taking values zero or one each with probability a half:  $\mu(0) = \mu(1) = 1/2$ . Then  $Z_{t+1} = X_{t+1} - X_t \pmod{n}$ , and, since modular subtraction is bijective,  $X$  is a  $p$ -uniform chain with  $\pi_{ij} = j - i \pmod{n}$  for each  $i, j \in \mathcal{S}$ . In terms of iterating random functions,  $f_j(i) = \pi_i^{-1}j = j + i \pmod{n}$ , and we can write  $X_{t+1} = f_{Z_{t+1}}(X_t)$ . Notice that we apply only either  $f_0(i) = i$  or  $f_1(i) = i + 1 \pmod{n}$ .

The transition matrix  $P$  is defined by  $P(i, j) = \mu(\pi_{ij}) = \mu(j - i \pmod{n})$ , which is a half if  $i - j \pmod{n}$  is either zero or one, and is zero otherwise.  $X$  is irreducible and aperiodic, its stationary distribution is uniform, and it converges to the stationary distribution at an exponential rate.  $f_0$  and  $f_1$  are Lipschitz continuous with Lipschitz constant one under the metric  $\rho$  on  $\mathcal{S}$  defined by  $\rho(i, j) = \min\{j - i \pmod{n}, i - j \pmod{n}\}$ , the shortest modular addition distance from  $i$  to  $j$  in either direction.  $\square$

**Example 2.4.7** (Vector Autoregressive Model (Hoff, 2015, p. 1171)). Hoff (2015, p. 1171) introduces a multilinear tensor autoregressive model for network data similar to VAR models (cf. Wooldridge, 1999, August/2012, p. 657). Let  $\mathbf{X}_i$  be the vectorization of a network's

weighted adjacency matrix at time  $i$ . The bilinear version of Hoff's model is

$$\mathbf{X}_i = \theta \mathbf{X}_{i-1} + \mathbf{Z}_i, \quad \mathbb{E}(\mathbf{Z}_i) = \mathbf{0}, \quad \mathbb{E}(\mathbf{Z}_i \mathbf{Z}_j^\top) = \begin{cases} \Sigma & i = j \\ \mathbf{0} & i \neq j, \end{cases}$$

where  $\theta$  and  $\Sigma$  are matrices of parameters to be estimated. Hoff used variations of this model on a time series of verbal and material diplomatic actions among 25 countries between 2004 and 2014. Geographically nearby countries' actions were the best predictors of each country's actions, with the exceptions of the United Kingdom and Australia and of the United States and certain other countries. The analysis also found that the relations between two countries depends on other countries' relations.

With a couple simple restrictions, this model becomes a  $p$ -uniform Markov chain. First, we must restrict all values to rationals so that the state space for the  $\mathbf{X}_i$ s is countable. Second, we must assume that the  $\mathbf{Z}_i$ s are IID, which is compatible with the assumptions that  $\mathbb{E}(\mathbf{Z}_i) = \mathbf{0}$  and  $\mathbb{E}(\mathbf{Z}_i \mathbf{Z}_j^\top) = \mathbf{0}$  if  $i \neq j$ . Then the set  $\pi$  of permutations under which the  $\mathbf{X}_i$ s are  $p$ -uniform are those for which  $\pi_{\mathbf{X}_{i-1}} \mathbf{X}_i = \mathbf{X}_i - \theta \mathbf{X}_{i-1} = \mathbf{Z}_i$ .  $\square$

More resources on random walks on finite sets include Diaconis (1988), Hildebrand (2005), Hirayama and Yano (2013), and Yano and Yasutomi (2011a, 2011b).

**2.4.3 Consequences of Independence.** When  $X$  is a  $p$ -uniform Markov chain, its corresponding IID sequence  $Z$ —which may not be unique (even almost surely) as a function of the underlying state space—conveys all the probabilistic information about  $X$ , modulo the initial value  $X_0$ . One common way to measure information is entropy. The *entropy* of a probability vector  $\mathbf{p}$  on  $\mathcal{S}$  is  $H(\mathbf{p}) = -\sum_{a \in \mathcal{S}} p_a \log p_a$ , where we define  $0 \log 0 := 0$  (Shiryaev, 2016, chap. 1, § 5.4). If  $P$  is a transition matrix on  $\mathcal{S}$  with a stationary distribution  $\mathbf{v}$ , the *entropy rate* of the corresponding Markov chain  $X$  is  $H(X) = -\sum_{a \in \mathcal{S}} v_a \sum_{b \in \mathcal{S}} P(a, b) \log P(a, b)$  (Yano & Yasutomi, 2011b, § 1.3).

**Theorem 2.4.8.** *Let  $X$  be a  $p$ -uniform Markov chain on  $\mathcal{S}$  with transition matrix  $P$  and let  $Z$  be the same as it was in the statement of theorem 2.4.5 with common law  $\mu$ . If  $P$  has a stationary distribution  $\mathbf{v}$ , the entropy rate of  $X$  equals the entropy of  $Z$ 's common law  $\mu$ .*

*Proof.* Since  $P(a, b) = \mu(\pi_a b)$  for all  $a, b \in \mathcal{S}$ , the entropy rate of  $X$  reduces to

$$H(X) = - \sum_{a \in \mathcal{S}} v_a \sum_{b \in \mathcal{S}} \mu(\pi_a b) \log \mu(\pi_a b).$$

However, for each  $a \in \mathcal{S}$ ,  $\pi_a$  in the summand above is a permutation, so

$$- \sum_{b \in \mathcal{S}} \mu(\pi_a b) \log \mu(\pi_a b) = - \sum_{b \in \mathcal{S}} \mu(b) \log \mu(b) = H(\mu).$$

Therefore

$$H(X) = - \sum_{a \in \mathcal{S}} v_a \sum_{b \in \mathcal{S}} \mu(\pi_a b) \log \mu(\pi_a b) = \sum_{a \in \mathcal{S}} v_a H(\mu) = H(\mu) \sum_{a \in \mathcal{S}} v_a = H(\mu). \quad \square$$

A statistical consequence of a Markov chain's  $p$ -uniformity is that we can apply IID convergence theorems to the Markov chain. In the case of  $p$ -uniform Markov chains that are MEFs, theorem 2.4.9 strengthens results in the literature for CEFs where the function  $\tau$  in the theorem is the sufficient statistic. Stefanov (1995, Prop. 1.1) proved convergence in probability of the time-average of an MEF's sufficient statistic. Theorem 2.4.9 strengthens this convergence to almost-sure and  $L^1$  convergence. Coupling theorem 2.4.9 with theorem 2.3.19 gives a limit to the mean parameter of the MEF.

Feigin (1981, Thm. 1, p. 598) has such a result for CEFs, but only in expectation conditional on the past state of the chain, with which condition theorem 2.3.19 dispenses.

**Theorem 2.4.9.** *Suppose  $X$  is a Markov chain on  $\mathcal{S}$  under  $\mathbb{P}$  with transition matrix  $P$   $p$ -uniform under  $\pi$ . Let  $\tau: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^\ell$  be  $p$ -uniform under  $\pi$  such that  $\sum_{b \in \mathcal{S}} \tau(a, b)P(a, b)$  converges absolutely for all  $a \in \mathcal{S}$ . Then, for any  $a \in \mathcal{S}$  and any  $s \in \mathbb{N}$ ,*

$$\frac{1}{t} \sum_{i=0}^{t-1} \tau(X_i, X_{i+1}) \rightarrow \mathbb{E}_{\mathbb{P}}[\tau(X_s, X_{s+1})] = \sum_{b \in \mathcal{S}} \tau(a, b)P(a, b) \quad \text{as } t \rightarrow \infty$$

where convergence is both almost sure and in  $L^1$  under  $\mathbb{P}$ . In particular, the limit does not depend on  $s$ , the expectation does not depend on the initial distribution of  $X_0$ , and the equation does not depend on  $a$ .

*Proof.* By theorem 2.4.5,  $Z_i := \pi_{X_{i-1}} X_i$ ,  $i \in \mathbb{N}_{>0}$ , is a sequence of IID,  $\mathcal{S}$ -valued random variables with some common law  $\mu$ . Since  $\tau$  is  $p$ -uniform under  $\pi$ , lemma 2.4.2 allows us to define  $m: \mathcal{S} \rightarrow \mathbb{R}^\ell$  by  $m(b) := \tau(a, \pi_a^{-1} b)$  for all  $a, b \in \mathcal{S}$ .



We claim that  $\mathbf{m}(Z_1)$  has finite expectation. Since  $\sum_{b \in \mathcal{S}} \tau(a, b)P(a, b)$  converges absolutely for all  $a \in \mathcal{S}$ , the Riemann rearrangement theorem says that every rearrangement converges absolutely to the same value (Rudin, 1976, Thm. 3.55). For each  $a \in \mathcal{S}$ ,  $\pi_a^{-1}$  is bijective, so

$$\sum_{b \in \mathcal{S}} \tau(a, b)P(a, b) = \sum_{b \in \mathcal{S}} \tau(a, \pi_a^{-1}b)P(a, \pi_a^{-1}b) \quad (2.39)$$

converges absolutely. Further, theorem 2.4.5 says that  $\mu$  satisfies  $P(a, b) = \mu(\pi_a b)$  for all  $b \in \mathcal{S}$ , so

$$\sum_{b \in \mathcal{S}} \tau(a, \pi_a^{-1}b)P(a, \pi_a^{-1}b) = \sum_{b \in \mathcal{S}} \tau(a, b)\mu(b) = \mathbb{E}_\mu[\mathbf{m}(Z_1)]. \quad (2.40)$$

Combining eqs. (2.39) and (2.40) yields that

$$\sum_{b \in \mathcal{S}} \tau(a, b)P(a, b) = \mathbb{E}_\mu[\mathbf{m}(Z_1)] \quad (2.41)$$

converges absolutely. Thus  $\mathbf{m}(Z_1)$  has finite expectation.

This allows us to apply Kolmogorov's strong law of large numbers (Jacod & Protter, 2004, Thm. 20.2):

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=0}^{t-1} \mathbf{m}(Z_{i+1}) = \mathbb{E}_\mu[\mathbf{m}(Z_1)] \quad (2.42)$$

holds both  $\mu$ -almost surely and in  $L^1$ . Since the  $Z_i$ s have the law  $\mu$  under  $\mathbb{P}$ , eq. (2.42) holds  $\mathbb{P}$ -almost surely. For any  $t \in \mathbb{N}_{>0}$ , we have

$$\frac{1}{t} \sum_{i=0}^{t-1} \mathbf{m}(Z_{i+1}) = \frac{1}{t} \sum_{i=0}^{t-1} \tau(X_i, \pi_{X_i}^{-1}Z_{i+1}) = \frac{1}{t} \sum_{i=0}^{t-1} \tau(X_i, X_{i+1}). \quad (2.43)$$

For any  $s \in \mathbb{N}$ , the definition of expectation and  $X$ 's Markov property yield

$$\mathbb{E}_\mathbb{P}[\tau(X_s, X_{s+1})] = \sum_{a \in \mathcal{S}} \sum_{b \in \mathcal{S}} \tau(a, b) \mathbb{P}(X_s = a)P(a, b) = \sum_{a \in \mathcal{S}} \mathbb{P}(X_s = a) \sum_{b \in \mathcal{S}} \tau(a, b)P(a, b),$$

and, by eq. (2.41),

$$\begin{aligned} &= \sum_{a \in \mathcal{S}} \mathbb{P}(X_s = a) \mathbb{E}_\mu[\mathbf{m}(Z_1)] \\ &= \mathbb{E}_\mu[\mathbf{m}(Z_1)]. \end{aligned} \quad (2.44)$$

Combining eqs. (2.41) to (2.44) yields the result.  $\square$

**2.4.4 Infinite State Spaces.** Infinite state spaces cannot always usefully bear p-uniform functions.

**Theorem 2.4.10.** *Suppose  $\mathcal{S}$  is countably infinite and  $f: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  is p-uniform. Then  $\sum_{a,b \in \mathcal{S}} f(a,b)$  converges absolutely if and only if  $f$  is identically zero.*

*Proof.* Suppose  $f: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  is p-uniform. Using lemma 2.4.2, pick  $g: \mathcal{S} \rightarrow \mathbb{R}$  and a set  $\{\pi_a\}_{a \in \mathcal{S}}$  of permutations such that  $f(a,b) = g(\pi_a b)$  for all  $a,b \in \mathcal{S}$ . Order  $\mathcal{S}$  as, say,  $\{s_1, s_2, \dots\} = \mathcal{S}$ .

For each  $a \in \mathcal{S}$ , the  $\ell^1(\mathcal{S})$  norm of row  $a$  of  $f$  is

$$\|f(a, \cdot)\|_1 = \sum_{b \in \mathcal{S}} |f(a,b)| = \sum_{b \in \mathcal{S}} |g(\pi_a b)| = \sum_{i=1}^{\infty} |g(\pi_a s_i)|.$$

Because of the absolute value bars, the series either diverges or converges absolutely. By the Riemann rearrangement theorem, for all  $a$ ,  $\|f(a, \cdot)\|_1 = \sum_{b \in \mathcal{S}} |g(b)| =: h$  either converges absolutely to the common value  $h$  or diverges to  $h = \infty$  (because  $\pi_a$  is a bijection) (Rudin, 1976, Thm. 3.55). In particular, our choice of order for  $\mathcal{S}$  did not matter.

By Fubini's theorem (J. K. Hunter & Nachtergaele, 2005, Thm. 12.41),  $\sum_{a,b \in \mathcal{S}} f(a,b)$  converges absolutely if and only if

$$\sum_{a,b \in \mathcal{S}} |f(a,b)| = \sum_{a \in \mathcal{S}} \left[ \sum_{b \in \mathcal{S}} |f(a,b)| \right].$$

The right-hand side equals  $\sum_{a \in \mathcal{S}} \|f(a, \cdot)\|_1 = \sum_{a \in \mathcal{S}} h$ . However,  $\sum_{a \in \mathcal{S}} h$  converges if and only if  $h = 0$  if and only if  $f$  is identically zero.  $\square$

The Riemann rearrangement theorem says that a conditionally convergent series that does not converge absolutely can be rearranged to converge to any value at all or to diverge (Rudin, 1976, Thm. 3.54). If we can pick a good ordering of  $\mathcal{S} = \{s_1, s_2, \dots\}$  and the permutations  $\{\pi_a\}_{a \in \mathcal{S}}$  are well chosen, a p-uniform  $f$  could yield a sequence of  $\|f(s_i, \cdot)\|_1$  whose series converges. But this requires getting lucky.

**Corollary 2.4.11.** *Suppose  $\mathcal{S}$  is countably infinite, and let  $f: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  be p-uniform. Order  $\mathcal{S} = \{s_1, s_2, \dots\}$ . Then  $\sum_{i,j=1}^{\infty} f(s_i, s_j)$  converges only if  $f$  is identically zero or  $\sum_{j=1}^{\infty} f(a, s_j)$  converges conditionally but not absolutely for all  $a \in \mathcal{S}$ .*

*Proof.* The proof of theorem 2.4.10 explains why  $\sum_{j=1}^{\infty} f(a, s_j)$  must converge but not absolutely for all  $a \in \mathcal{S}$ . If  $f$  is identically zero, then the series converges absolutely and thus also conditionally.  $\square$

**2.4.5 Finite State Spaces.** In this subsection we discuss the linear algebraic structure of  $p$ -uniform stochastic matrices over non-empty, finite state spaces with  $n \in \mathbb{N}$  elements. For  $i \in [n]$ , let  $e_i$  be the  $i$ th standard basis vector in  $\mathbb{R}^n$  and let  $\mathbf{1} = (1, \dots, 1)$ . If  $\boldsymbol{\mu}$  is an  $n$ -vector, then  $\mathbf{1}^\top \boldsymbol{\mu} = \mathbf{1} \cdot \boldsymbol{\mu} = \sum_{i=1}^n \mu_i$ .

Does being  $p$ -uniform constrain a matrix's rank? According to the next theorem, not only is the answer no as long as the rows of the matrix aren't the uniform distribution, but the complexity is low for the the permutations needed to achieve a full rank,  $p$ -uniform matrix. In particular, the set  $R$  in the statement contains only  $2\binom{n}{2}$  permutations.

**Theorem 2.4.12.** *Let  $\boldsymbol{\mu}$  be a stochastic vector in  $\mathbb{R}^n$ . Pick  $i < j \in [n]$  such that  $\mu_i \neq \mu_j$ , or, if those don't exist, pick any  $i, j \in [n]$ . Denote the set of permutations on  $[n]$  by  $S_n$ . Let  $R \subseteq S_n$  be the set of permutations on  $[n]$  including permutations of the forms  $(k\ i)(\ell\ j)$  and  $(i\ j)(k\ j)(\ell\ i)$  for  $k < \ell \in [n]$ , where we interpret permutations of the form  $(i\ i)$  to be the identity. Let  $A$  and  $B$  be the sets of  $n$ -vectors obtained by applying some permutation in  $R$  or  $S_n$ , respectively, to  $\boldsymbol{\mu}$ . Denote  $V := \text{span } A$ .*

*Then  $V = \text{span } B$ . If  $\boldsymbol{\mu}$  is the uniform distribution, then  $V = \text{span}\{\boldsymbol{\mu}\} = \text{span}\{\mathbf{1}\}$ . If  $\boldsymbol{\mu}$  is not the uniform distribution, then  $V = \mathbb{R}^n$ .*

*Proof.* Since  $A \subseteq B \subseteq \mathbb{R}^n$ , we have  $A \subset \text{span } A = V \subseteq \text{span } B \subseteq \mathbb{R}^n$ , and hence  $\dim V \leq \dim \text{span } B$ .  $\boldsymbol{\mu}$ 's entries are nonnegative and sum to one, so  $\boldsymbol{\mu}$  is not the zero vector, and hence  $0 < \dim V$ . If  $\boldsymbol{\mu}$  is the uniform distribution, then all entries of  $\boldsymbol{\mu}$  are equal, so  $A = B = \{\boldsymbol{\mu}\}$ , and  $V = \text{span } B = \text{span}\{\mathbf{1}\}$  since  $\mathbf{1} = n\boldsymbol{\mu}$ .  $\text{span}\{\mathbf{1}\}$  is the set of all vectors whose coordinates are constant.

Thus we may assume  $\boldsymbol{\mu}$  is not uniform, so it has two entries not equal to each other,  $\mu_i \neq \mu_j$ , where  $i < j \in [n]$ . If we can prove that  $V = \mathbb{R}^n$ , then  $V \subseteq \text{span } B = \mathbb{R}^n$  as well.

First we show that  $V^\perp \subseteq \text{span}\{\mathbf{1}\}$ . Our proof of this claim follows Lahtonen (2011). Suppose to the contrary that  $V^\perp \not\subseteq \text{span}\{\mathbf{1}\}$ , so that for some  $\mathbf{u} \in V^\perp$ ,  $u_k \neq u_\ell$  for some

$k < \ell \in [n]$ . Let  $\alpha \in A$  be the vector whose entries are those of  $\mu$  under the permutation  $(k\ i)(\ell\ j)$ , and  $\beta \in A$  be the same but under  $(i\ j)(k\ j)(\ell\ i) = (i\ k\ j\ \ell)$ . In other words,

$$\begin{aligned}\beta_k &= \mu_j = \alpha_\ell & \beta_i &= \mu_k = \alpha_i \\ \beta_\ell &= \mu_i = \alpha_k & \beta_j &= \mu_\ell = \alpha_j\end{aligned}$$

and  $\beta_h = \mu_h = \alpha_h$ , for all  $h \in [n] \setminus \{i, j, k, \ell\}$ . Then  $\alpha - \beta = (\mu_i - \mu_j)e_k + (\mu_j - \mu_i)e_\ell$ . Since  $\mathbf{u} \in V^\perp$  and  $\alpha, \beta \in A \subset V$ , we also have  $\mathbf{u}^\top \alpha = 0 = \mathbf{u}^\top \beta$ . Hence

$$\begin{aligned}0 &= 0 - 0 = \mathbf{u}^\top \alpha - \mathbf{u}^\top \beta = \mathbf{u}^\top (\alpha - \beta) \\ &= (\mu_i - \mu_j)u_k + (\mu_j - \mu_i)u_\ell \\ &= (\mu_i - \mu_j)(u_k - u_\ell).\end{aligned}$$

But  $\mu_i \neq \mu_j$  and  $u_k \neq u_\ell$ , so this is a contradiction. Therefore  $V^\perp$  contains no vector with unequal coordinates, i.e.,  $V^\perp \subseteq \text{span}\{\mathbf{1}\}$ .

That  $V^\perp = \{\mathbf{0}\}$ , and thus  $V = \mathbb{R}^n$ , follows from the fact that  $\mathbf{1} \notin V^\perp$ . This is because  $\mu \in V$  and  $\mathbf{1}^\top \mu = 1 \neq 0$ .  $V^\perp$  is a subspace contained in  $\text{span}\{\mathbf{1}\}$  but not containing  $\mathbf{1}$ . The only such subspace is  $\{\mathbf{0}\}$ .  $\square$

The following characterization of  $p$ -uniform matrices will facilitate the next several results.

**Lemma 2.4.13.** *Let  $P$  be a  $p$ -uniform  $n \times n$  matrix, and let  $\mu$  be a permutation of a row of  $P$ . Then, for some permutation matrices  $\Pi_r$ ,  $r \in [n]$ , we have*

$$P = \sum_{r=1}^n \mathbf{e}_r \mu^\top \Pi_r. \quad (2.45)$$

Further, for all stochastic vectors  $\mathbf{v} \in \mathbb{R}^n$  there exists a doubly stochastic matrix  $D \in \mathbb{R}^{n \times n}$  such that  $\mathbf{v}^\top P = \mu^\top D$ .

*Proof.* For each row  $r$ , let  $\Pi_r$  be the  $n \times n$  permutation matrix such that the  $r$ th row of  $P$  is  $\mu^\top \Pi_r$  (so  $\Pi_r$  is the permutation matrix corresponding to  $\pi_r^{-1}$  in the notation of lemma 2.4.2).

This proves eq. (2.45).

Let  $\mathbf{v}$  be a stochastic  $n$ -vector, so  $\sum_{r=1}^n v_r = 1$  and  $\mathbf{v} > 0$ . Then we have  $\mathbf{v}^\top P = \sum_{r=1}^n v_r \boldsymbol{\mu}^\top \Pi_r = \boldsymbol{\mu}^\top (\sum_{r=1}^n v_r \Pi_r)$ . By the Birkhoff-von Neumann theorem (see, e.g., Bertsimas & Tsitsiklis, 1997, Exercise 7.26; but the theorem may have first appeared in König, 1936, chap. xiv, § 3), the matrix  $D = \sum_{r=1}^n v_r \Pi_r$  is doubly stochastic.  $\square$

For much more on the relationship between doubly stochastic matrices and the set of permutations of a vector, see Marshall et al. (2011). This includes a characterization of the convex hull of the set of permutations of a vector  $\boldsymbol{\mu}$  as the image of  $\boldsymbol{\mu}$  under all the doubly stochastic matrices. We include just one application here to the stationary distributions of  $p$ -uniform stochastic matrices. If  $\mathbf{x} \in \mathbb{R}^n$ , let  $x_{(i)}$  denote the  $i$ th *order statistic* of  $\mathbf{x}$ , meaning the  $i$ th smallest entry of  $\mathbf{x}$ , so  $x_{(1)} \leq \dots \leq x_{(n)}$ . If  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , we write  $\mathbf{x} < \mathbf{y}$  and say that  $\mathbf{y}$  *majorizes* (Marshall et al., 2011, Def. 1.1.A.1, p. 8)  $\mathbf{x}$  if

$$\sum_{i=1}^k x_{(n-i+1)} \leq \sum_{i=1}^k y_{(n-i+1)} \quad \text{for all } k \in [n-1] \quad \text{and} \quad \sum_{i=1}^n x_i = \sum_{i=1}^n y_i.$$

**Theorem 2.4.14.** *Let  $P$  be an  $n \times n$ ,  $p$ -uniform, stochastic matrix with a stationary distribution  $\mathbf{v}$ . Then any permutation of a row of  $P$  majorizes  $\mathbf{v}$ .*

*Proof.* Let  $\boldsymbol{\mu}$  be any permutation of a row of  $P$ , as in lemma 2.4.2. By lemma 2.4.13, there exists some doubly stochastic,  $n \times n$  matrix  $D$  such that  $\boldsymbol{\mu}^\top D = \mathbf{v}^\top P = \mathbf{v}^\top$ , so  $\mathbf{v} = D^\top \boldsymbol{\mu}$ . Since  $D$  is doubly stochastic, so is  $D^\top$ . By the Hardy-Littlewood-Pólya theorem (Marshall et al., 2011, Thm. 1.2.B.2, p. 33),  $\mathbf{v} < \boldsymbol{\mu}$ .  $\square$

A corollary (Marshall et al., 2011, Cor. 1.2.B.3, p. 34) to the Hardy-Littlewood-Pólya theorem implies that

$$\mathbf{v} \in \text{conv hull}\{\Pi \boldsymbol{\mu} \mid \Pi \text{ is a permutation matrix}\},$$

but this is weaker than the simpler statement that  $\mathbf{v}$  is in the convex hull of the rows of  $P$ .

The next lemma helps bound the norm of a  $p$ -uniform matrix.

**Lemma 2.4.15.** *Let  $P$  be a  $p$ -uniform  $n \times n$  matrix, and let  $\boldsymbol{\mu}$  be a permutation of a row of  $P$ . The diagonal entries of  $PP^\top$  are all equal to  $\|\boldsymbol{\mu}\|_2^2$ .*

*Proof.* The  $r, c$  entry of  $PP^T$  is the dot product of the  $r$ th and  $c$ th rows of  $P$ . Using eq. (2.45), the  $r$ th row of  $P$  is  $\boldsymbol{\mu}^T \Pi_r$ , for some permutation, hence orthogonal, matrix  $\Pi_r$ . Thus the  $r, r$  entry of  $PP^T$  is  $(\boldsymbol{\mu}^T \Pi_r)(\boldsymbol{\mu}^T \Pi_r)^T = \boldsymbol{\mu}^T \Pi_r \Pi_r^T \boldsymbol{\mu} = \boldsymbol{\mu}^T I \boldsymbol{\mu} = \boldsymbol{\mu}^T \boldsymbol{\mu} = \|\boldsymbol{\mu}\|_2^2$ .  $\square$

Recall that if  $P$  is an  $n \times n$  real matrix, its *Hilbert-Schmidt* or *Frobenius norm*  $\|P\|_F = \sqrt{\text{tr } PP^T}$ , the the square root of the sum of squares of the singular values of  $P$ , which is also equal to the square root of the sum of squares of the entries of  $P$  (Trefethen & Bau, 1997, pp. 22–23). The *1-norm*  $\|P\|_1$ , *2-norm*  $\|P\|_2$ , and  *$\infty$ -norm*  $\|P\|_\infty$  are the maximum absolute column sum, largest singular value, and maximum absolute row sum respectively (Trefethen & Bau, 1997, pp. 20–21 and Thm. 5.3).  $\rho(P)$  is the *spectral radius*, i.e., the maximum absolute value of an eigenvalue of  $P$  (Trefethen & Bau, 1997, Exercise 3.2). Golub and Van Loan (1983/2013, Eq. 2.3.12 on p. 72) says that  $\frac{1}{\sqrt{n}}\|P\|_1 \leq \|P\|_2$ . We can say much more about p-uniform matrices and especially p-uniform, stochastic matrices.

**Lemma 2.4.16** (Norms). *Let  $P$  be a p-uniform  $n \times n$  matrix whose rows are permutations of the  $n$ -vector  $\boldsymbol{\mu}$ . Then  $\|P\|_F = \sqrt{n}\|\boldsymbol{\mu}\|_2$ ,  $\sqrt{\frac{n}{r}}\|\boldsymbol{\mu}\|_2 \leq \|P\|_2 \leq \sqrt{n}\|\boldsymbol{\mu}\|_2$  where  $r = \text{rank } P$ , and  $\|P\|_1 \leq n\|\boldsymbol{\mu}\|_\infty$ . These inequalities are sharp.*

*Additionally, if  $\boldsymbol{\mu}$  is a stochastic vector so that  $P$  is a stochastic matrix, then  $|\det P| \leq 1$  and*

$$\frac{1}{n} \leq \|\boldsymbol{\mu}\|_\infty \leq \|\boldsymbol{\mu}\|_2 \leq \|\boldsymbol{\mu}\|_1 = \|P\|_\infty = 1 = \rho(P) \leq \|P\|_2 \leq \begin{cases} \|P\|_F = \sqrt{n}\|\boldsymbol{\mu}\|_2 \\ \sqrt{\|P\|_1} \leq \sqrt{n}\|\boldsymbol{\mu}\|_\infty \end{cases} \leq \sqrt{n}.$$

*These inequalities are sharp. Finally,  $\|P\|_1 = 1$  if and only if  $P$  is doubly stochastic if and only if  $\|P\|_2 = 1$ .*

*Proof.* First suppose  $P$  is p-uniform and that its rows are permutations of the vector  $\boldsymbol{\mu}$ . Each diagonal entry of  $PP^T$  equals  $\|\boldsymbol{\mu}\|_2^2$  by lemma 2.4.15, so  $\|P\|_F^2 = \text{tr } PP^T = \sum_{i=1}^n \|\boldsymbol{\mu}\|_2^2 = n\|\boldsymbol{\mu}\|_2^2$ . Denote  $P$ 's singular values as  $s_1 \geq s_2 \geq \dots \geq s_r > 0 = s_{r+1} = \dots = s_n$  (Trefethen & Bau, 1997, Thm. 5.1). By Trefethen and Bau (1997, Thm. 5.3),  $\|P\|_2^2 = s_1^2 \leq s_1^2 + \dots + s_r^2 = \|P\|_F^2 = n\|\boldsymbol{\mu}\|_2^2$ . Thus  $n\|\boldsymbol{\mu}\|_2^2 = \sum_{i=1}^r s_i^2 \leq \sum_{i=1}^r s_1^2 = rs_1^2$ . These inequalities are sharp, for example in the case

$\boldsymbol{\mu} = \frac{1}{n}\mathbf{1}$  so that  $r = 1$ . By Trefethen and Bau (1997, Example 3.3),

$$\|P\|_1 = \max_{c \in [n]} \sum_{i=1}^n |P_{i,c}| \leq \sum_{i=1}^n \max_{c \in [n]} |P_{i,c}| \leq \sum_{i=1}^n \|\boldsymbol{\mu}\|_\infty = n\|\boldsymbol{\mu}\|_\infty.$$

This inequality is sharp, for example in the case that  $P = \mathbf{1}e_1^\top$ , the matrix whose first column is all ones and has zeros everywhere else.

Now suppose further that  $\boldsymbol{\mu}$  is a stochastic vector so that  $P$  is a stochastic matrix. This means that  $1 = \sum_{i=1}^n \boldsymbol{\mu}_i \leq \sum_{i=1}^n \|\boldsymbol{\mu}\|_\infty = n\|\boldsymbol{\mu}\|_\infty$ , so  $1/n \leq \|\boldsymbol{\mu}\|_\infty$ . This is sharp by  $\boldsymbol{\mu} = \frac{1}{n}\mathbf{1}$ . That  $\|\boldsymbol{\mu}\|_\infty \leq \|\boldsymbol{\mu}\|_2 \leq \|\boldsymbol{\mu}\|_1$  comes from Golub and Van Loan (1983/2013, Eqs. 2.2.5–7 on p. 69), and is sharp by  $\boldsymbol{\mu} = e_1$ .  $\|\boldsymbol{\mu}\|_1$ , the sum of the entries of  $\boldsymbol{\mu}$ , is the sum of the entries of every row of  $P$  and thus equals  $\|\boldsymbol{\mu}\|_1 = \|P\|_\infty$ , the largest such sum.  $P$  is stochastic, so  $\|P\|_\infty = 1$ , and  $P\mathbf{1} = \mathbf{1}$ .

The latter equation shows that  $1 \leq \rho(P)$ . By Gerschgorin's circle theorem (Trefethen & Bau, 1997, Exercise 24.2), every eigenvalue  $\lambda$  of  $P$  lies in one of the closed circular disks in the complex plane centered at a diagonal entry  $P_{i,i}$  of  $P$  with radius  $\sum_{c \neq i} P_{i,c}$ .  $P$ 's being stochastic means that  $\sum_{c \neq i} P_{i,c} = 1 - P_{i,i}$ . Therefore  $|P_{i,i} - \lambda| \leq 1 - P_{i,i}$ , so  $|\lambda| \leq 1$ , and thus  $\rho(P) \leq 1$ . Consequently,  $1 = \rho(P)$  and  $|\det P| \leq (\rho(P))^n = 1$ .

It is always the case that  $\rho(P) \leq \|P\|_2$  (Trefethen & Bau, 1997, Exercise 3.2). We already showed above that  $\|P\|_2^2 \leq \|P\|_F^2 = n\|\boldsymbol{\mu}\|_2^2 \leq n$ .

Golub and Van Loan (1983/2013, Cor. 2.3.2 on p. 73) says that  $\|P\|_2 \leq \sqrt{\|P\|_1 \|P\|_\infty}$  always, so in our case in which  $\|P\|_\infty = 1$ , we also have  $\|P\|_2 \leq \sqrt{\|P\|_1}$ . This is sharp when both values equal one, for example when  $\boldsymbol{\mu} = \frac{1}{n}\mathbf{1}$ . We already showed above that  $\|P\|_1 \leq n\|\boldsymbol{\mu}\|_\infty \leq n$ .

Finally we prove that  $\|P\|_1 = 1$  if and only if  $P$  is doubly stochastic if and only if  $\|P\|_2 = 1$ . First suppose that  $P$  is doubly stochastic. Then every column of  $P$  sums to one, so  $\|P\|_1 = 1$ . Second suppose that  $\|P\|_1 = 1$ . Since  $1 \leq \|P\|_2 \leq \sqrt{\|P\|_1} = 1$ , we have  $\|P\|_2 = 1$ . Third suppose that  $\|P\|_2 = 1$  but that  $P$  is not column stochastic (i.e.,  $P$  is a stochastic matrix but not a doubly stochastic matrix) (The remainder of this proof follows user1551, 2015). Then  $\mathbf{1}^\top P$  is not a scalar multiple of  $\mathbf{1}$  and thus the Cauchy-Schwarz theorem (Axler, 1997, Thm. 6.6) says that  $|\mathbf{1}^\top P\mathbf{1}| < \|\mathbf{1}^\top P\|_2 \|\mathbf{1}\|_2$ .  $P$  is stochastic so  $|\mathbf{1}^\top P\mathbf{1}| = n$ , and we know that

$\|\mathbf{1}\|_2 = \sqrt{n}$ . Thus  $\sqrt{n} < \|\mathbf{1}^\top P\|_2$ . By the definition of the 2-norm (Trefethen & Bau, 1997, Eq. 3.5),  $\|\mathbf{1}^\top P\|_2 \leq \|\mathbf{1}\|_2 \|P\|_2$ , but  $\|P\|_2 = 1$ , so  $\sqrt{n} < \|\mathbf{1}^\top P\|_2 \leq \|\mathbf{1}\|_2 = \sqrt{n}$ , a contradiction. Therefore  $P$  is doubly stochastic.  $\square$

The bound  $\sqrt{\frac{n}{r}} \|\boldsymbol{\mu}\|_2 \leq \|P\|_2$  is not useful—we already know  $1 \leq \|P\|_2$ —unless  $\sqrt{\frac{r}{n}} \leq \|\boldsymbol{\mu}\|_2$ . The rank  $r$  of  $P$  depends not just on  $\boldsymbol{\mu}$  but also on the permutations  $\pi_i$  that each row  $P_i^\top$  is of  $\boldsymbol{\mu}$ . For example, if  $\boldsymbol{\mu} = \frac{1}{n}\mathbf{1}$ ,  $r = 1$  since every row is the same regardless of the permutations. If  $\boldsymbol{\mu} = e_1$ ,  $P$  could be the matrix  $\mathbf{1}e_1^\top$  or the identity  $I$ , or something in between, depending on the permutations. The first has rank  $r = 1$  and the second rank  $r = n$ . The first has  $\|\mathbf{1}e_1^\top\|_2 = \sqrt{n}$  and the second has  $\|I\|_2 = 1$ .

**2.4.5.1 Symmetry.** A sufficient condition for a stochastic matrix to be doubly stochastic is that it be symmetric, because then  $\mathbf{1}^\top = (P\mathbf{1})^\top = \mathbf{1}^\top P^\top = \mathbf{1}^\top P$ . In this sub-subsection we look at some necessary and sufficient conditions for this.

Let  $\pi := \{\pi_i\}_{i=1}^n$  be a set of permutations of  $[n]$ . We say that  $\pi$  is *symmetric* if  $\pi_i j = \pi_j i$  for all  $i, j \in [n]$ . A *Latin square* is an  $n \times n$  matrix that contains each number one through  $n$  exactly once in each row and exactly once in each column (Marshall et al., 2011, Example 1.2.H.1, p. 61). We say that  $\pi$  *determines a Latin square* if the matrix  $(\pi_i j)_{i,j=1}^n$  is a Latin square. A *quasi-group*  $(Q, *)$  is a set  $Q$  together with a binary operation  $*$ :  $Q \times Q \rightarrow Q$  such that, for each  $x, y \in Q$ , there is a unique  $u \in Q$  with  $u * x = y$  and a unique  $v \in Q$  with  $x * v = y$  (Robinson, 2003, Exercise 10.4.4, p. 207). We say that  $\pi$  *determines a quasi-group* if  $*$ :  $[n] \times [n] \rightarrow [n]$  defined by  $i * j = \pi_i j$  makes  $([n], *)$  a quasi-group. The *multiplication table* of a quasi-group is the matrix  $(i * j)_{i,j \in Q}$ . Every Latin square is the multiplication table of some finite quasi-group and every finite quasi-group's multiplication table is a Latin square (Robinson, 2003, Exercises 10.4.4 and 10.4.5, p. 207).

**Lemma 2.4.17.** *Let  $P$  be a  $p$ -uniform  $n \times n$  matrix with permutations  $\pi := \{\pi_i\}_{i=1}^n$ , and let  $\boldsymbol{\mu}$  be an  $n$ -vector such that  $P_{ij} = \mu_{\pi_i j}$ . If  $\pi$  is symmetric, then it determines a symmetric Latin square and a commutative quasi-group, and  $P$  is symmetric. Conversely, if  $P$  is symmetric and the entries of  $\boldsymbol{\mu}$  are all distinct, then  $\pi$  is symmetric.*

*Proof.* ( $\implies$ ). Suppose  $\pi$  is symmetric, so  $\pi_i j = \pi_j i$  for all indices  $i$  and  $j$ . Then  $P_{ij} = \mu_{\pi_i j} =$



$\mu_{\pi_j i} = P_{ji}$  for all indices  $i$  and  $j$ . Define the  $n \times n$  matrix  $A := (\pi_{ij})_{i,j=1}^n$ . Let  $i$  and  $k$  be numbers from one to  $n$ .  $\pi_i$  is a permutation, so there is exactly one  $j$  such that  $A_{ij} = \pi_{ij} = k$ . Thus  $k$  shows up exactly once on row  $i$  of  $A$ . Now we know that  $k$  also appears at least once in column  $j$  at row  $i$ . Suppose it exists also in another row, say  $r$ , in column  $j$ . Then  $\pi_j i = \pi_j j = \pi_r j = \pi_j r$ .  $\pi_j$  is a permutation, so we can cancel it from both sides of the equation, leaving  $i = r$ . This violates our choice of  $r$ , so no such  $r$  exists. Thus  $A$  is Latin square, and moreover a symmetric one since  $A_{ij} = \pi_{ij} = \pi_{ji} = A_{ji}$  for all indices  $i$  and  $j$ . That is,  $\pi$  determines a symmetric Latin square.

Finally,  $\pi$  determines a quasi-group where  $([n], *)$ ,  $*$ :  $[n] \times [n] \rightarrow [n]$  defined by  $i * j = \pi_{ij}$ . This is because if  $i, k \in [n]$ , then  $k = u * i = \pi_u i = \pi_i u$  and  $k = i * v = \pi_i v$  imply that  $u = \pi_i^{-1} k = v$  uniquely. The quasi-group is commutative because  $i * j = \pi_{ij} = \pi_{ji} = j * i$ .  $P$  is symmetric because  $P_{ij} = \mu_{\pi_{ij}} = \mu_{\pi_{ji}} = P_{ji}$ .

( $\Leftarrow$ ) Suppose  $P$  is symmetric and the entries of  $\mu$  are unique.  $P$ 's symmetry implies that for some indices  $i$  and  $j$ ,  $\mu_{\pi_{ij}} = P_{ij} = P_{ji} = \mu_{\pi_{ji}}$ . The uniqueness of the entries of  $\mu$  implies that  $\pi_{ij} = \pi_{ji}$ .  $\square$

A *loop* is a quasi-group that also has an identity element (Smith & Romanowska, 1999, chap. 1, § 4, p. 87).

**Corollary 2.4.18.** *If  $\pi$  is symmetric and there is some  $z \in [n]$  such that  $\pi_z$  is the identity permutation, then  $z$  is the only  $y \in [n]$  such that  $\pi_y$  is the identity permutation, and  $\pi$  determines a commutative loop with identity element  $z$  and where the inverse of  $i \in [n]$  is  $\pi_i^{-1} z$ .*

*Proof.* By lemma 2.4.17,  $\pi$  determines a commutative quasi-group where  $([n], *)$ ,  $*$ :  $[n] \times [n] \rightarrow [n]$  defined by  $i * j = \pi_{ij}$ . To see that  $z$  is an identity of the quasi-group, take  $i \in [n]$ . Since  $\pi_z$  is the identity,  $i = \pi_z i$ . By symmetry,  $\pi_z i = \pi_i z$ . Thus  $i = z * i = i * z$ .

If  $y \in [n]$  were such that  $\pi_y$  were the identity permutation, then  $\pi_i y = i * y = i = i * z = \pi_i z$ , so  $y = z$  since  $\pi_i$  is a permutation. Consequently  $z$  is the unique such element.

Finally  $(\pi_i^{-1} z) * i = i * (\pi_i^{-1} z) = \pi_i \pi_i^{-1} z = z$ , so  $\pi_i^{-1} z$  is the inverse of  $i$  in the quasi-group.  $\square$

Loops are important because they arise from a commonly used set operator.

**Example 2.4.19.** If  $A$  and  $B$  are sets, then the *symmetric difference*  $A\Delta B$  is defined as  $A\Delta B := (A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A)$ , all the elements in exactly one of  $A$  or  $B$ . If the state space  $\mathcal{S}$  is a field of sets (closed under intersection and union), then we can define permutations  $\pi_A: \mathcal{S} \rightarrow \mathcal{S}$  for each  $A \in \mathcal{S}$  by  $\pi_A B := A\Delta B$ . The symmetric difference operator is commutative because unions are commutative. The empty set  $\emptyset$  is the identity for  $\Delta$  because  $\emptyset\Delta B = B$ . Every set is its own inverse under  $\Delta$  because  $A\Delta A = \emptyset$ . Thus, by corollary 2.4.18,  $\{\pi_A\}_{A \in \mathcal{S}}$  determines a loop  $(\mathcal{S}, \Delta)$  with identity  $\emptyset$  and where the inverse of each element is the element itself. (In fact, because  $\Delta$  is associative,  $(\mathcal{S}, \Delta)$  is an abelian group.)  $\square$

The next example shows that symmetric permutations always exist for finite state spaces.

**Example 2.4.20.** For each  $i \in [n]$ , let  $i \% n$  denote the integer value (rather than the modular congruence class) remaining after  $n$  divides  $i$ , but identify 0 with  $n$ . More precisely,

$$i \% n = \begin{cases} j & \text{if } i \equiv j \pmod{n} \text{ for some } j \in \{1, \dots, n-1\} \\ n & \text{if } i \equiv 0 \pmod{n}. \end{cases}$$

In particular,  $(i \% n)$  is the unique integer in  $[n] = \{1, \dots, n\}$  such that  $(i \% n) \equiv i \pmod{n}$ .

Define the permutation  $\pi_i$  for  $i \in [n]$  by  $\pi_i j = (i + j - 1) \% n$ .  $\pi$  inherits its symmetry from addition's commutativity. For each  $j \in [n]$ , the identity permutation is  $\pi_1$  because  $\pi_1 j = (1 + j - 1) \% n = j \% n = j$ . Thus, corollary 2.4.18 says that  $\pi$  determines a loop with identity 1. The inverse of  $j$  is the unique solution  $x$  to the equation  $1 = \pi_j x$ . We can verify that  $x = \pi_j^{-1} 1 = 2 - j$  because  $\pi_j x = (j + 2 - j - 1) \% n = 1 \% n = 1$ .

To give a sense of the structure of  $\pi$ , below is the matrix  $(\pi_i j)_{i,j=1}^n$  for  $n = 7$ , and, for contrast, the *circulant matrix*  $C$  for  $n$ -vector  $\mu$  defined by  $C = (\mu_{[i-j \pmod{n}] + 1})_{i,j=1}^n$  (Trefethen

& Bau, 1997, p. 318).

$$\underbrace{\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 3 & 4 & 5 & 6 & 7 & 1 \\ 3 & 4 & 5 & 6 & 7 & 1 & 2 \\ 4 & 5 & 6 & 7 & 1 & 2 & 3 \\ 5 & 6 & 7 & 1 & 2 & 3 & 4 \\ 6 & 7 & 1 & 2 & 3 & 4 & 5 \\ 7 & 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix}}_{\text{matrix for } \pi} \quad \underbrace{\begin{bmatrix} 1 & 7 & 6 & 5 & 4 & 3 & 2 \\ 2 & 1 & 7 & 6 & 5 & 4 & 3 \\ 3 & 2 & 1 & 7 & 6 & 5 & 4 \\ 4 & 3 & 2 & 1 & 7 & 6 & 5 \\ 5 & 4 & 3 & 2 & 1 & 7 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 & 7 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix}}_{\text{circulant matrix for } (1, \dots, 7)}$$

□

The effect of symmetry on transition matrices is on the eigenspace, specifically the stationary distribution.

**Corollary 2.4.21.** *Let  $P$  be a  $p$ -uniform  $n \times n$  stochastic matrix with a symmetric set of permutations  $\{\pi_i\}_{i=1}^n$  such that  $P_{i, \pi_i^{-1}j} = P_{k, \pi_k^{-1}j}$  for all  $i, j, k \in [n]$ . Then  $P$  is symmetric and doubly stochastic, so the uniform distribution is a stationary distribution. If  $P$  is also irreducible, then  $P$ 's unique stationary distribution is the uniform distribution.*

*Proof.*  $P$ 's symmetry and double stochasticity follow from lemma 2.4.17 and the paragraph before it. The second statement follows from the Perron-Frobenius theorem (Godsil & Royle, 2001, Thm. 8.8.1). □

In the symmetric  $\pi$  case, the doubly stochastic matrix  $D$  that lemma 2.4.13 says must exist in the equation  $\frac{1}{n} \mathbf{1}^\top P = \boldsymbol{\mu}^\top D$  is  $D = \frac{1}{n} \sum_{r=1}^n \Pi_r = \frac{1}{n} \mathbf{1} \mathbf{1}^\top$ , where  $\Pi_r$  is the permutation matrix corresponding to  $\pi_r^{-1}$  according to the proof of lemma 2.4.13. That  $\frac{1}{n} \sum_{r=1}^n \Pi_r^\top = \frac{1}{n} \mathbf{1} \mathbf{1}^\top$  follows from the symmetry of  $\pi$  because if  $\Pi_q^\top$  and  $\Pi_r^\top$  both had a one in the  $k$ th row, then  $\pi_q k = \pi_r k$  and thus  $\pi_k q = \pi_k r$  by symmetry. Since  $\pi_k$  is a permutation,  $q = r$ .

The following corollary applies theorem 2.4.14 to derive a general theorem about majorization.

**Corollary 2.4.22.** *Every stochastic  $n$ -vector  $\boldsymbol{\mu}$  majorizes  $\frac{1}{n} \mathbf{1}$ , meaning that*

$$\frac{k}{n} \geq \sum_{i=1}^k \mu_{(i)} \quad \text{for all } k \in [n]. \quad (2.46)$$

*Proof.* Use the permutations  $\pi$  defined in example 2.4.20 to come up with the  $n \times n$  matrix  $P$  by  $P_{ij} = \mu_{\pi_{ij}}$ , which is doubly stochastic with stationary distribution  $\frac{1}{n}\mathbf{1}$  by corollary 2.4.21. From theorem 2.4.14,  $\frac{1}{n}\mathbf{1} < \mu$ , and by the definition of majorization, the inequalities in eq. (2.46) follow.  $\square$

**2.4.6 Permutation Uniformity and CEFS.** In this subsection, we show that p-uniformity preserves MEF structure; we have already discussed convergence of mean parameters of p-uniform MEFS in and around theorem 2.4.9.

Feigin (1981) presented a theorem about CAEFS in a spirit similar to our present investigation of CEFS' relationship with IID sequences. It supposed  $X$  was real valued and  $\mathcal{P}$  was a CAEF as in eq. (2.7) whose natural parameter space is open. Feigin (1981, Thm. 3) derived an additive process (partial sums of IID real-valued random variables) with the same law as a certain function of  $X$  when  $\tau(a, \cdot)$  is invertible in the second slot. On a finite state space and with  $\kappa(a, \cdot)$  constant, invertibility of  $\tau(a, \cdot)$  in the second slot is equivalent to the transition matrix's having distinct numbers in every entry of a row. This is similar to the distinctness requirement of Rosenblatt (1959, Lem. 3). Feigin (1981, Thm. 4) derived strong consistency of MLE and a central limit theorem for the Fisher information of a CAEF under the invertibility assumption.

In terms of theorem 2.4.5, the next proposition and its corollary show that if  $X$  is p-uniform, then its distribution comes from an MEF if and only if  $Z$ 's distribution comes from an exponential family. We say that  $\mathcal{P}$  is *p-uniform* under  $\pi$  if  $P_{\theta}(a, \pi_a^{-1}c) = P_{\theta}(b, \pi_b^{-1}c)$  for all  $a, b, c \in \mathcal{S}$  and all  $\theta \in \Theta$ . In this case, for each  $\theta \in \Theta$ , lemma 2.4.2 says that there exists a PMF  $\mu_{\theta}$  such that  $P_{\theta}(a, b) = \mu_{\theta}(\pi_a b)$  for all  $a, b \in \mathcal{S}$ . Put  $\mathcal{M} := \{\mu_{\theta}\}_{\theta \in \Theta}$ .

**Proposition 2.4.23.** *Suppose  $\mathcal{P}$  is p-uniform as above. If  $\mathcal{P}$  is the CEF given in eq. (2.7), then  $\mathcal{M}$  is an exponential family:  $\mu_{\theta}(b) = \kappa(a, \pi_a^{-1}b) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}(a, \pi_a^{-1}b) - \zeta(a, \boldsymbol{\theta}))$  for any  $a, b \in \mathcal{S}$  and  $\boldsymbol{\theta} \in \Theta$ . Conversely, if  $\mathcal{M}$  is the exponential family given in eq. (2.5), then  $\mathcal{P}$  is an MEF:  $P_{\theta}(a, b) = \kappa(\pi_a b) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}(\pi_a b) - \zeta(\boldsymbol{\theta}))$  for any  $a, b \in \mathcal{S}$  and  $\boldsymbol{\theta} \in \Theta$ .*

*Proof.* Prove the forward implication by plugging  $\mu_{\theta}(b) = \mu_{\theta}(\pi_a \pi_a^{-1}b) = P_{\theta}(a, \pi_a^{-1}b)$  into eq. (2.7). Prove the converse by plugging  $P_{\theta}(a, b) = \mu_{\theta}(\pi_a b)$  into eq. (2.5).  $\square$

**Corollary 2.4.24.** *If  $\mathcal{P}$  is  $p$ -uniform and a CEF, then it is an MEF.*

*Proof.* Proposition 2.4.23 says that, if  $\mathcal{P}$  is  $p$ -uniform and a CEF, then  $\mathcal{M}$  is an exponential family, which in turn implies that  $\mathcal{P}$  is an MEF.  $\square$

Next, observation 2.4.25 and theorem 2.4.26 give necessary and sufficient conditions for an MEF to be  $p$ -uniform in terms of the  $p$ -uniformity of the carrier measure  $\kappa$  and sufficient statistic  $\tau$ . First we give the sufficient condition.

*Observation 2.4.25.* If  $\mathcal{P}$  is the CEF given in eq. (2.7) and  $\tau$  and  $\kappa$  are  $p$ -uniform both under  $\pi$ , then  $\mathcal{P}$  is  $p$ -uniform under  $\pi$ .

*Proof.* For all  $a, b, c \in \mathcal{S}$  and all  $\theta \in \Theta$ ,

$$P_{\theta}(a, \pi_a^{-1}b) = \kappa(a, \pi_a^{-1}b) \exp\left(\eta(\theta) \cdot \tau(a, \pi_a^{-1}b) - \zeta(a, \theta)\right) \quad (2.47)$$

$$= \kappa(c, \pi_c^{-1}b) \exp\left(\eta(\theta) \cdot \tau(c, \pi_c^{-1}b) - \zeta(a, \theta)\right) \quad (2.48)$$

We set 1 equal to the sum of eq. (2.47) over all  $b \in \mathcal{S}$  and rearrange to obtain

$$\exp(\zeta(a, \theta)) = \sum_{b \in \mathcal{S}} \kappa(a, \pi_a^{-1}b) \exp\left(\eta(\theta) \cdot \tau(a, \pi_a^{-1}b)\right). \quad (2.49)$$

Doing the same thing to eq. (2.48) and replacing  $a$  with  $c$  in eq. (2.49) reveals that  $\zeta(a, \theta) = \zeta(c, \theta)$ . Thus eq. (2.48) equals

$$\kappa(c, \pi_c^{-1}b) \exp\left(\eta(\theta) \cdot \tau(c, \pi_c^{-1}b) - \zeta(c, \theta)\right) = P_{\theta}(c, \pi_c^{-1}b). \quad \square$$

The exact converse of the above Observation is not quite true. If  $\kappa$  is zero in a couple positions, then the corresponding values of  $\tau$  are unconstrained. We also need  $\Theta$  to be big enough to determine which hyperplanes the different values of  $\tau$  lie in. We manage the latter concern in theorem 2.4.26 by assuming that  $\eta$  has affinely independent entries.

**Theorem 2.4.26.** *Suppose that  $\mathcal{P}$  is the MEF from eq. (2.8) and is  $p$ -uniform under  $\pi$ . Then  $\kappa$  is  $p$ -uniform under  $\pi$ . Further, if  $\kappa$  is never zero and  $\eta$  has affinely independent entries, then  $\tau$  is  $p$ -uniform under  $\pi$ .*

*Proof.* Fix  $a, b, c \in \mathcal{S}$ . If  $\kappa(a, \pi_a^{-1}b) = 0$ , then  $0 = P_\theta(a, \pi_a^{-1}b) = P_\theta(c, \pi_c^{-1}b)$  for all  $\theta \in \Theta$ . Since  $\exp > 0$  on  $\mathbb{R}$ , we must have  $\kappa(c, \pi_c^{-1}b) = 0$  as well.

By the same token, suppose  $\kappa(a, \pi_a^{-1}b) > 0$ , and notice that  $\kappa(c, \pi_c^{-1}b) > 0$  as well. Then rearranging the equation  $P_\theta(a, \pi_a^{-1}b) = P_\theta(c, \pi_c^{-1}b)$  yields

$$\eta(\theta) \cdot [\tau(c, \pi_c^{-1}b) - \tau(a, \pi_a^{-1}b)] = \log \frac{\kappa(a, \pi_a^{-1}b)}{\kappa(c, \pi_c^{-1}b)} \quad \text{for all } \theta \in \Theta.$$

Since the right side is constant with respect to  $\theta$ , affine independence of  $\eta$ 's entries implies that  $\tau(c, \pi_c^{-1}b) - \tau(a, \pi_a^{-1}b) = \mathbf{0}$  and  $\log(\kappa(a, \pi_a^{-1}b)/\kappa(c, \pi_c^{-1}b)) = 0$ .  $\square$

A variety of hypotheses familiar in exponential family theory imply that  $\eta$  has affinely independent entries, justifying the application of theorem 2.4.26. These hypotheses assume alternately that 1. eq. (2.8) is a *minimal representation* of the  $a$ th row of  $P \in \mathcal{P}$ , meaning that  $\tau(a, \cdot)$  and  $\eta$  both have affinely independent entries (Barndorff-Nielsen, 1978, Cor. 8.1, p. 113; Küchler & Sørensen, 1997, p. 38; Wainwright & Jordan, 2008, p. 40); 2.  $\eta(\Theta)$  contains an open,  $\ell$ -dimensional set (Casella & Berger, 2002, Thms. 5.2.11 or 6.2.25, pp. 217, 288); or 3.  $\eta(\Theta)$  contains  $\ell + 1$  affinely independent vectors (E. L. Lehmann & Casella, 1998, Cor. 6.16, p. 39). Item 2 implies item 3, which in turn implies that  $\eta$  has affinely independent entries; the straightforward proof is as follows.

To prove the first implication, for  $i \in [\ell]$ , let  $e_i$  be the  $i$ th standard basis vector in  $\mathbb{R}^\ell$ . The open set contains an open ball centered at some vector  $c$  with radius some  $r$ . The list of  $\ell + 1$  vectors  $c, \frac{r}{2}e_1 + c, \dots, \frac{r}{2}e_\ell + c$  is affinely independent.

For the second implication, consider any  $\ell + 1$  affinely independent vectors in  $\eta(\Theta)$ , say,  $\eta(\theta_0), \dots, \eta(\theta_\ell)$ . Then  $\eta(\theta_1) - \eta(\theta_0), \dots, \eta(\theta_\ell) - \eta(\theta_0)$  are linearly independent (Bertsimas & Tsitsiklis, 1997, chap. 3, § 6, Def. 3.6, p. 120). Form an  $\ell \times \ell$  matrix  $A$  with these vectors as columns. Suppose  $\delta \in \mathbb{R}^\ell$  and  $h \in \mathbb{R}$  are such that  $\delta \cdot \eta(\theta) = h$  for all  $\theta \in \Theta$ . Then  $\delta \cdot (\eta(\theta_i) - \eta(\theta_0)) = 0$  for all  $i \in [\ell]$ , and hence  $A^\top \delta = \mathbf{0}$ . By the linear independence of  $A$ 's columns,  $\delta = \mathbf{0}$ , and thus  $h = 0$ . Hence  $\eta$  has affinely independent entries.

Recall from eq. (2.27) that for an MEF  $P$  defined in eq. (2.8), we defined  $R(\theta) = \{\eta(\theta) \cdot \tau_{cd} \mid \kappa_{cd} \neq 0\}$  and  $S_c(\theta, r) = \{d \mid \eta(\theta) \cdot \tau_{cd} = r\}$  for  $r \in R(\theta)$ . When  $P$  is p-uniform,

we can strengthen eq. (2.28) with the following result.

**Lemma 2.4.27.** *If  $P$  is a  $p$ -uniform MEF defined in eq. (2.8) such that  $\tau$  is  $p$ -uniform with the same permutations  $\{\pi_a\}_{a \in \mathcal{S}}$  as  $P$ , then, for all  $\theta \in \Theta$ ,  $r \in \mathbb{R}(\theta)$ , and  $a, b \in \mathcal{S}$ ,  $S_a(\theta, r)$  is isomorphic to  $S_b(\theta, r)$  under the bijection  $\pi_a^{-1}\pi_b$ .*

*Proof.* Suppose  $c \in S_a(\theta, r)$ .  $r = \eta(\theta) \cdot \tau(a, c) = \eta(\theta) \cdot \tau(b, \pi_b^{-1}\pi_a c)$ . Thus  $\pi_b^{-1}\pi_a c \in S_b(\theta, r)$ . The backward implications are also true.  $\square$

## 2.5 Markov Chains of Graphs

In this section, we apply the theories of CEFS and of  $p$ -uniform Markov chains to some of the network models that Hanneke et al. (2010) proposed. We find that some of them are  $p$ -uniform and thus MEFS. For two of the models, we can avoid MCMC for MLE, which is what Hanneke et al. used; instead we give a closed form for the MLE. We also explore the relationships among  $p$ -uniformity, MEFS, and statistical independence of the random edges in the graphs. The main result in theorems 2.5.8 and 2.5.9 is that we may replace  $t$  observations of a  $p$ -uniform Markov chain of graphs with a single observation of a corresponding multigraph. We introduce *exponential random  $t$ -multigraph models* for this purpose.

Let the state space  $\mathcal{S}$  be the set  $\mathcal{G}_{n,t}$  of loopless multigraphs on the vertex set  $[n] := \{1, \dots, n\}$  ( $n \geq 2$ ) where each edge has multiplicity at most  $t \in \mathbb{N}_{>0}$ . Each potential edge or non-edge comes from the set  $\mathcal{D}_n := \binom{[n]}{2}$  of *dyads* whose elements  $\{i, j\} \subseteq [n]$  we may write as  $ij$  or  $ji$  when the meaning is clear. Let  $N := |\mathcal{D}_n| = \binom{n}{2}$ . We identify a multigraph  $g$  with its *edge-multiplicity vector*  $\mathbf{g} \in \{0, 1, \dots, t\}^{\mathcal{D}_n} \equiv \mathcal{G}_{n,t}$ . This is the vectorization of the adjacency matrix of  $g$ . If  $\mathbf{g} \in \mathcal{G}_{n,t}$  and  $f \in \mathcal{D}_n$ , then  $g(f)$  is the *multiplicity* of dyad  $f$ . The *edge set*  $E(\mathbf{g})$  is  $\{f \in \mathcal{D}_n \mid g(f) > 0\}$ . The *complement* of  $\mathbf{g}$  is  $\bar{\mathbf{g}} := t\mathbf{1} - \mathbf{g}$ .

An *exponential random graph model* (ERGM) is an exponential family defined on simple graphs,  $\mathcal{G}_{n,1}$ . We introduce the analogous family of models for multigraphs.

**Definition 2.5.1.** We call an exponential family defined on  $\mathcal{G}_{n,t}$  an *exponential random  $t$ -multigraph model* ( $t$ -ERMGM).

The choice of which ERGM to use in practice depends on identifying the sufficient statistics appropriate for specific data. Those statistics may incorporate node covariates, leading to models that Fienberg et al. (1985) introduced and whose MLE Yan et al. (2018) investigated. Fienberg et al. contrasted *microanalytic* studies, such as Snijders (2001), employing node covariates with *macroanalytic* studies solely of network topology—our focus in the sequel. Many popular macroanalytic models, which Goldenberg et al. (2010) surveyed, rely on statistics built on subgraph counts. The simplest choice of subgraph is the single edge.

**Example 2.5.2** (Erdős-Rényi Graph Model). This ERGM arises by choosing edges of random graph  $G$  independently each with probability  $p$ . The probability of  $G = g$  is  $p^{E(g)}(1 - p)^{N - E(g)} = \exp[\log(p/(1 - p))E(g) + N \log(1 - p)]$ . The parameter function is  $\eta(p) = \log \frac{p}{1 - p}$ , sufficient statistic is  $E(g)$ , and the log-partition function is  $-N \log(1 - p)$  (Chatterjee & Diaconis, 2013, § 2.2).  $\square$

Chatterjee and Diaconis (2013) treated subgraph counts in a general context. Specific choices of subgraphs were made in the studies in Bannister et al. (2014) of counts of triangles, in Park and Newman (2004) of counts of two-stars, or in Snijders et al. (2006) of a complicated combination of degree counts, triangles, and 2-stars. Chatterjee et al. (2011), Holland and Leinhardt (1981), and Rinaldo et al. (2013) used degree sequences—equivalent to counting labeled  $k$ -stars.

**Example 2.5.3** ( $\beta$  Model). This is the ERGM whose sufficient statistic is the degree sequence  $\beta(G)$ . The probability of  $G = g$  is  $\exp(\beta(g) \cdot \log \theta - \zeta(\theta)) = e^{-\zeta(\theta)} \prod_{uv \in E(g)} \theta_u \theta_v$ . The parameter  $\theta_v$  represents the attractiveness of vertex  $v$  (Petrović, 2015). The log-partition function is  $\zeta(\theta) = \prod_{u=1}^n \prod_{v=1}^{u-1} (\theta_u \theta_v + 1)$  (Chatterjee et al., 2011, § 1.2). Chatterjee et al. (2011, Thm. 1.5) gave an algorithm for the MLE of  $\theta$ .  $\square$

**2.5.1 Literature Review.** Since the 1980s exponential random graph models (ERGMs) have proved successful for modeling single observations of random networks whose probabilities are parameterized functions of certain observables, called sufficient statistics, such as the number of edges, the degree sequence, or the counts of specified subgraphs. These



are generative statistical models, allowing scientists to estimate parameters that weight the importance of different sufficient statistics in explaining why a network has a certain topology. In the succeeding decades, statisticians have developed sophisticated techniques for estimating these parameters and testing goodness of fit for these models. For surveys of this literature, see Goldenberg et al. (2010), Kolaczyk (2017), and Kolaczyk and Csárdi (2014/2020b, pp. 88–97).

Stochastic models of change over time in social networks started with the discrete-time, dyadically independent model of Katz and Proctor (1959). Later models viewed the underlying social process as a continuous-time Markov chain that we observe at discrete time points. The earliest of these studies, such as Sørensen and Hallinan (1976), did not model the entire social network, focusing instead on triads (three mutual friends). By assuming dyadic independence—that friendships between two people only depend on the people in question—Holland and Leinhardt (1977) and Wasserman (1980) derived the equilibrium behavior of the Markov chain separately for each dyad in terms of the model’s parameters. We will use the assumption of dyadic independence in subsection 2.5.3. Time-evolving *stochastic actor-oriented models* introduced in Snijders (2001) have a game theoretic flavor, allowing optimizations at each node of the network to drive the stochastic evolution.

Time series of ERGMS first appeared in Wasserman and Iacobucci (1988). In that paper’s model, time is discrete and all time periods’ distributions are mutually independent. Each time period’s ERGM has the same sufficient statistics as all the others, but parameters are allowed to change over time. Grindrod and Higham (2010) focused on *range-dependent random graphs* where nodes are numbers and transition probabilities depend on the distance between them. Grindrod and Parsons (2011) considered stochastic processes of graphs where each edge is independent and has memory longer than one step. Robins and Pattison (2001) introduced Markov chains of networks whose transition probabilities constitute exponential families. Hanneke et al. (2010) and Hanneke and Xing (2006) explored this family of models, calling them *temporal ERGMS* (TERGMS), and focused on the dyadic independence case. Krivitsky and Handcock (2013) introduced *separable* TERGMS, which are TERGMS in which the edges formed and edges dissolved at each time step are parameter-

ized separately. Rastelli et al. (2018) extended the stochastic block model to a time series of networks where the block allocations form independent Markov chains for each node. In particular Hanneke et al. explored a couple methods for parameter estimation of this class of Markov chains and discussed degeneracy issues in the selection of sufficient statistics.

Frank and Strauss (1986) investigate how dependence among dyads determines which subgraph counts should be the sufficient statistics, concluding that triangles and stars are sufficient for shared-neighbor-only dependence. More generally, let  $d$  be the *dependence graph* of a  $\mathcal{G}_{n,1}$ -valued random variable  $G$ , that is,  $d$ 's vertex set is  $\mathcal{D}_n$  and it has edges between dyads  $f, h \in \mathcal{D}_n$  if  $G(f)$  and  $G(h)$  are not independent conditional on the rest of  $G$ . Then the probability distribution of  $G$  is a log-linear model whose sufficient statistics are the indicator functions for the presence of the cliques of  $d$  in  $G$  (Frank & Strauss, 1986, Thm. 1 and Eq. 3.3). Thus every distribution over  $\mathcal{G}_{n,1}$  is an ERGM, albeit one with a potentially very high dimensional sufficient statistic.

Grindrod and Higham (2013) and Grindrod et al. (2011) proposed a way of summarizing time-respecting communicability within a sequence of (possibly directed) graphs. If  $A_{t_1}, \dots, A_{t_k} \in 0, 1^{n \times n}$  is a sequence of  $k$  adjacency matrices of graphs with no loops, define  $S_0 = 0 \in \mathbb{R}^{n \times n}$  and  $S_i = (I + e^{-b(t_i - t_{i-1})} S_{i-1})(I - aA_{t_i})^{-1} - I$ . For this to work,  $a$  must be less than the reciprocal of the largest spectral radius of any of the  $A_i$ s, and  $b$  must be positive. The  $S_i$ s count the number of time-respecting walks from each node to each other node, exponentially discounting old edges ( $e^{b \times \text{time}}$ ) and distant edges ( $a^{\text{distance}}$ ). Communication from or to a node can then be summarized as  $S_k \mathbf{1}$  or  $S_k^T \mathbf{1}$ .

**2.5.2 Finite Exchangeability.** We say that  $\mathbf{b}, \mathbf{c} \in \mathcal{G}_{n,t}$  are *isomorphic* and write  $\mathbf{b} \sim \mathbf{c}$  if there exists a bijection  $\phi$  on  $[n]$  such that  $b(\phi(i)\phi(j)) = c(\phi(i)\phi(j))$  for all  $i, j \in [n]$ . Isomorphism is an equivalence relation. If  $\mu$  is any PMF (or any other function) on  $\mathcal{G}_{n,t}$ , we say that  $\mu$  is *finitely exchangeable* if  $\mathbf{b} \sim \mathbf{c} \implies \mu(\mathbf{b}) = \mu(\mathbf{c})$  (cf. lemma 2.4.4's condition (a)). Lauritzen et al. (2019) and Lauritzen et al. (2018) defined and analyzed finitely exchangeable PMFs on  $\mathcal{G}_{n,1}$ . The former showed that the set of all such PMFs on  $\mathcal{G}_{n,1}$  form an exponential family whose sufficient statistic counts subgraphs of the random network by isomorphism class. The latter related the finitely exchangeable distributions of random networks to the marginal

distributions of their subgraphs, and gave a de Finetti-like theorem for representing those distributions.

We can use lemma 2.4.4 to extend finite exchangeability from a transition matrix  $P$  on  $\mathcal{G}_{n,t} \times \mathcal{G}_{n,t}$  p-uniform under  $\pi$  to or from a PMF  $\mu$  on  $\mathcal{G}_{n,t}$  such that  $P(\mathbf{a}, \mathbf{b}) = \mu(\pi_{\mathbf{a}}\mathbf{b})$ . To do so requires that  $\sim$  be *invariant under*  $\pi$  in the sense of lemma 2.4.4's condition (c). Since  $\mathcal{G}_{n,t}$  is finite,  $\sim$  is invariant under  $\pi$  if and only if  $\sim$  is invariant under  $\pi^{-1} := \{\pi_{\mathbf{a}}^{-1}\}_{\mathbf{a} \in \mathcal{G}_{n,t}}$ , which is lemma 2.4.4's condition (d). We summarize these considerations as follows.

**Corollary 2.5.4.** *Suppose  $P$  and  $\mu$  are as above for  $\mathcal{S} = \mathcal{G}_{n,t}$ . If either of lemma 2.4.4's conditions (c) and (d) or their converses hold, then  $\mu$  is finitely exchangeable if and only if every row of  $P$  is, and every row of  $P$  is finitely exchangeable if and only if some row of  $P$  is.*

*Proof.* In lemma 2.4.4, replace  $\mathcal{S}$  by  $\mathcal{G}_{n,t}$  (which is finite),  $f$  by  $P$ , and  $g$  by  $\mu$ . □

One permutation set  $\pi$  preserving isomorphism classes has  $\pi_{\mathbf{a}}\mathbf{b} = \bar{\mathbf{b}}$  for all  $\mathbf{a}, \mathbf{b} \in \mathcal{G}_{n,t}$ . Finitely exchangeable distributions include any in exponential families for which the carrier measure and the sufficient statistic are finitely exchangeable. Edge counts (example 2.5.2) are constant on isomorphism classes. Degree sequence (example 2.5.3) is not constant on isomorphism classes, but degree distribution, a contingency table of degree counts (Lauritzen et al., 2018, § 5), the degree sequence after sorting, or any other of unlabeled subgraphs are.

**2.5.3 Dyadic Independence.** Suppose  $G$  is a  $\mathcal{G}_{n,t}$ -valued random variable. If  $\{G(f) \mid f \in \mathcal{D}_n\}$  are mutually independent then  $G$  (or equivalently, its distribution) is *dyadically independent* (Goodreau, 2007, "Methods"). Imposing dyadic independence on a model can be appropriate in "settings where the drivers of link formation are predominately bilateral in nature, as may be true in some types of friendship and trade networks as well as in models of (some types of) conflict between nation-states." (Graham, 2017, p. 1039). Further, dyadically dependent ERGMS are known to be prone to *model degeneracy*, wherein nearly all the probability mass of the model lies on a small subset of the sample space, such as the empty and complete graphs (Handcock, 2003; Rinaldo et al., 2009, §§ 3.3–4). The remainder of this subsection describes conditions under which  $G$  is dyadically independent and how

dyadic independence interacts with  $p$ -uniformity. The main result, theorem 2.5.9, will allow us quickly to convert  $t + 1$  samples from a  $p$ -uniform MEF of simple graphs into a single observation of a  $t$ -ERMGM.

**2.5.3.1 Exponential Random Multigraph Models.** In this sub-subsection, the main result is lemma 2.5.6, which characterizes a  $t$ -ERMGM's dyadic independence in terms of its sufficient statistic and carrier measure. When the carrier measure is constant and  $t = 1$ —the ERGM case—it is already known that an ERGM is dyadically independent if and only if its sufficient statistic is dyadditive (D. R. Hunter et al., 2008, § 3.2; Shalizi & Rinaldo, 2013, Eq. 20): We say that a function  $\tau: \mathcal{G}_{n,t} \rightarrow \mathbb{R}^\ell$  is *dyadditive*, or *factors over edges* (Hanneke et al., 2010, Eq. 2), if there are functions  $\tau_f: \{0, 1, \dots, t\} \rightarrow \mathbb{R}^\ell$  for each dyad  $f \in \mathcal{D}_n$  such that

$$\tau(\mathbf{g}) = \sum_{f \in \mathcal{D}_n} \tau_f(\mathbf{g}(f)). \quad (2.50)$$

When  $t = 1$ ,  $\tau$  is dyadditive if and only if there is a real,  $\ell \times \mathcal{D}_n$  matrix  $Q$  such that  $Q\mathbf{g} = \tau(\mathbf{g}) - \tau(\mathbf{0})$ ; then the  $f$ th column of  $Q$  is  $\tau_f(1) - \tau_f(0)$ .

**Example 2.5.5.** Let  $\mathbf{g} \in \mathcal{G}_{n,1}$ . The sufficient statistic in the Erdős-Rényi graph model is the number  $|E(\mathbf{g})|$  of edges. The number of edges is dyadditive because  $|E(\mathbf{g})| = \mathbf{1} \cdot \mathbf{g}$ .

The  $\beta$  model's sufficient statistic, the degree sequence  $\beta(\mathbf{g})$ , is dyadditive. To see this, let  $K \in \{0, 1\}^{n \times \mathcal{D}_n}$  be the complete graph's *incidence matrix*, whose columns are the indicator  $n$ -vectors of the two-element sets constituting  $\mathcal{D}_n$ . Then  $\beta(\mathbf{g}) = K\mathbf{g}$ .

Statistics that rely on cliques larger than single edges are not dyadditive. □

When an ERGM's carrier measure is not constant, lemma 2.5.6 still specifies some cases where dyadditivity of the sufficient statistic implies dyadic independence. To describe those cases, we say that a function  $\kappa: \mathcal{G}_{n,t} \rightarrow \mathbb{R}$  is *dyadically multiplicative* if there are functions  $\kappa_f: \{0, \dots, t\} \rightarrow \mathbb{R}$  for each dyad  $f \in \mathcal{D}_n$  such that

$$\kappa(\mathbf{g}) = \prod_{f \in \mathcal{D}_n} \kappa_f(\mathbf{g}(f)). \quad (2.51)$$

Dyadically multiplicative carrier measures in ERGMs include the common cases in which the carrier measure is constant.

**Lemma 2.5.6.**  $G$  is a  $\mathcal{G}_{n,t}$ -valued random variable with the PMF in eq. (2.5) such that  $\tau$  is dyadditive as in eq. (2.50) and  $\kappa$  is dyadically multiplicative as in eq. (2.51) if and only if, for each  $f \in \mathcal{D}_n$ ,  $G(f)$  is a  $\{0, \dots, t\}$ -valued random variable with the PMF

$$\mu_{\theta}^f(m) = \frac{\kappa_f(m) \exp(\eta(\theta) \cdot \tau_f(m))}{\sum_{r=0}^t \kappa_f(r) \exp(\eta(\theta) \cdot \tau_f(r))}, \quad (2.52)$$

and  $G$  is dyadically independent.

*Proof.* That  $G$  is a  $\mathcal{G}_{n,t}$ -valued random variable if and only if  $G(f)$  is a  $\{0, \dots, t\}$ -valued random variable follows directly from the definition of  $\mathcal{G}_{n,t}$ . For some  $\theta \in \Theta$ , define  $\mu_{\theta}$  as in eq. (2.5) and  $\mu_{\theta}^f$  as in eq. (2.52). The backward implication will follow if we can show that, for all  $g \in \mathcal{G}_{n,t}$ ,  $\prod_{f \in \mathcal{D}_n} \mu_{\theta}^f(g(f)) = \mu_{\theta}(g)$ , where eqs. (2.50) and (2.51) define  $\tau$  and  $\kappa$  in terms of  $\{\tau_f\}_{f \in \mathcal{D}_n}$  and  $\{\kappa_f\}_{f \in \mathcal{D}_n}$ , respectively. The forward implication will follow if we can show that, for all  $f \in \mathcal{D}_n$ ,

$$\sum_{\substack{g \in \mathcal{S} \\ g(f)=m}} \mu_{\theta}(g) = \mu_{\theta}^f(m),$$

where eqs. (2.50) and (2.51) define  $\{\tau_f\}_{f \in \mathcal{D}_n}$  and  $\{\kappa_f\}_{f \in \mathcal{D}_n}$  in terms of  $\tau$  and  $\kappa$ , respectively. Then the dyadic independence of  $G$  will follow from the equality established when proving the backward implication.

**Forward Implication.** Fix an arbitrary dyad  $f \in \mathcal{D}_n$ . For tidiness we write  $\mathcal{S} = \mathcal{G}_{n,t}$ .

The probability that  $G(f) = m \in \{0, \dots, t\}$ , is

$$\sum_{\substack{g \in \mathcal{S} \\ g(f)=m}} \mu_{\theta}(g) = \frac{\sum_{\substack{g \in \mathcal{S} \\ g(f)=m}} \kappa(g) \exp(\eta(\theta) \cdot \tau(g))}{\sum_{r=0}^t \sum_{\substack{x \in \mathcal{S} \\ x(f)=r}} \kappa(x) \exp(\eta(\theta) \cdot \tau(x))},$$

where we have used eq. (2.4). Using eqs. (2.50) and (2.51), define

$$u(g) := \prod_{\substack{h \in \mathcal{D}_n \\ h \neq f}} \kappa_h(g(h)) \exp(\eta(\theta) \cdot \tau_h(g(h))).$$

$\tau$  is dyadditive and  $\kappa$  is dyadically multiplicative, so

$$\kappa(g) \exp(\eta(\theta) \cdot \tau(g)) = \kappa_f(g(f)) \exp(\eta(\theta) \cdot \tau_f(g(f))) u(g).$$

The expression for  $u(\mathbf{g})$  does not involve  $g(f)$ , so  $u(\mathbf{g}) = u(\mathbf{x})$  regardless of whether  $\mathbf{g}, \mathbf{x} \in \mathcal{S}$  have edge  $f$  the same number of times. Consequently,

$$\sum_{\substack{\mathbf{g} \in \mathcal{S} \\ g(f)=m}} u(\mathbf{g}) = \sum_{\substack{\mathbf{x} \in \mathcal{S} \\ x(f)=r}} u(\mathbf{x}) \quad (2.53)$$

for each  $r \in \{0, \dots, t\}$ . Factoring out this sum gives

$$\frac{\sum_{\substack{\mathbf{g} \in \mathcal{S} \\ g(f)=m}} \kappa(\mathbf{g}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}(\mathbf{g}))}{\sum_{r=0}^t \sum_{\substack{\mathbf{x} \in \mathcal{S} \\ x(f)=r}} \kappa(\mathbf{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}(\mathbf{x}))} = \frac{\kappa_f(m) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}_f(m)) \sum_{\substack{\mathbf{g} \in \mathcal{S} \\ g(f)=m}} u(\mathbf{g})}{\sum_{r=0}^t \kappa_f(r) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}_f(r)) \sum_{\substack{\mathbf{x} \in \mathcal{S} \\ x(f)=r}} u(\mathbf{x})},$$

which, after canceling out eq. (2.53) in the numerator and denominator, equals eq. (2.52).

**Backward Implication.** First off, for any  $\mathbf{x} \in \mathcal{G}_{n,t}$ , we have

$$\begin{aligned} \prod_{f \in \mathcal{D}_n} \kappa_f(x(f)) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}_f(x(f))) &= \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \sum_{f \in \mathcal{D}_n} \boldsymbol{\tau}_f(x(f))\right) \prod_{f \in \mathcal{D}_n} \kappa_f(x(f)) \\ &= \kappa(\mathbf{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}(\mathbf{x})), \end{aligned} \quad (2.54)$$

where we have defined  $\boldsymbol{\tau}$  and  $\kappa$  via eqs. (2.50) and (2.51).

Fix an arbitrary graph  $\mathbf{g} \in \mathcal{G}_{n,t}$ . Then, using eq. (2.54), the (joint) probability that  $\mathbf{G} = \mathbf{g}$  is

$$\prod_{h \in \mathcal{D}_n} \mu_{\boldsymbol{\theta}}^h(\mathbf{g}(h)) = \frac{\kappa(\mathbf{g}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}(\mathbf{g}))}{\prod_{h \in \mathcal{D}_n} \sum_{r=0}^t \kappa_h(r) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}_h(r))}.$$

This matches eq. (2.5) if we can show that the denominator equals  $e^{\zeta(\boldsymbol{\theta})}$ . To that end, we exchange  $\prod_{f \in \mathcal{D}_n}$  and  $\sum_{r=0}^t$  using lemma 2.3.12. In the language of lemma, set  $\mathcal{T} := \mathcal{D}_n$ , and  $B_h := \{0, \dots, t\}$  and  $f_h(r) := \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}_h(r))$  for each  $r \in B_h$  and each  $h \in \mathcal{T}$ . Then  $\mathcal{F}_{\mathcal{T}} = \mathcal{G}_{n,t}$ , and

$$\begin{aligned} \prod_{h \in \mathcal{D}_n} \sum_{r=0}^t \kappa_h(r) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}_h(r)) &= \sum_{\mathbf{x} \in \mathcal{G}_{n,t}} \prod_{h \in \mathcal{D}_n} \kappa_h(x(h)) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}_h(x(h))) \\ &= \sum_{\mathbf{x} \in \mathcal{G}_{n,t}} \kappa(\mathbf{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau}(\mathbf{x})) = \exp \zeta(\boldsymbol{\theta}), \end{aligned} \quad (2.55)$$

where the second equality follows from eq. (2.54) and the last from eq. (2.4).  $\square$

Calculating the log-partition function, and thus PMF, of an ERGM is computationally intractable for large  $n$ : it is  $\#\mathcal{P}$ -hard and inapproximable in polynomial time (Bannister et al., 2014). However, for the special case of dyadically independent ERGMs, computing the log-partition function requires a number of multiplications merely quadratic in  $n$  (Frank & Strauss, 1986, Example 1). The following corollary of lemma 2.5.6, which follows directly from expanding eq. (2.55), generalizes this existing result to  $t$ -ERMGMs.

*Observation 2.5.7.* Under the conditions of lemma 2.5.6, the partition function is

$$e^{\zeta(\theta)} = \prod_{u=1}^n \prod_{v=1}^{u-1} \sum_{r=0}^t \kappa_{uv}(r) \exp(\eta(\theta) \cdot \tau_{uv}(r)).$$

There are tractable solutions to the partition function for some other sufficient statistics. When  $\tau(a)$  is a vector of the number of different sized stars in graph  $a$ , Chatterjee and Diaconis (2013) approximated the partition function as the solution to a variational problem. Park and Newman (2004) used techniques from physics to solve the partition function if the sufficient statistic is the vector of the number of edges and the number of two-stars.

**2.5.3.2 Independent Sequences and Multigraphs.** Because of theorem 2.4.5, we may be able to derive an IID sequence of simple graphs from a Markov chain of simple graphs. In this sub-subsection, we consider how to turn an IID sequence of  $t$  ERGMs into a single observation of a  $t$ -ERMGM. Define the *multigraph union* of  $z_1, \dots, z_t \in \mathcal{G}_{n,s}$  to be their vector sum  $z_1 + \dots + z_t \in \mathcal{G}_{n,st}$ . Let  $Z = (Z_1, \dots, Z_t)$  be an IID sequence of dyadically independent,  $\mathcal{G}_{n,1}$ -valued random variables with multigraph union  $\mathbf{W}$ . Fix  $z = (z_1, \dots, z_t) \in \mathcal{G}_{n,1}^t$  with multigraph union  $w$ .

The following theorem roughly says that the order of the appearance of edges in  $Z$  does not matter to  $\mathbf{W}$ .

**Theorem 2.5.8** (Dyadically Independent Multigraphs). *With the notation above,  $\mathbf{W}$  is dyadically independent and*

$$\mathbb{P}(\mathbf{W} = w) = \mathbb{P}(Z = z) \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)}. \quad (2.56)$$

*Proof.* First we show dyadic independence. By the law of total probability,

$$\mathbb{P}(\mathbf{W} = \mathbf{w}) = \sum_{\substack{x \in \mathcal{G}_{n,1}^t \\ \sum_i x_i = \mathbf{w}}} \mathbb{P}(\mathbf{Z}_1 = \mathbf{x}_1, \dots, \mathbf{Z}_t = \mathbf{x}_t) = \sum_{\substack{x \in \mathcal{G}_{n,1}^t \\ \sum_i x_i = \mathbf{w}}} \prod_{f \in \mathcal{D}_n} \prod_{i=1}^t \mathbb{P}(Z_i(f) = x_i(f))$$

since  $Z$  is IID and dyadically independent.

To apply lemma 2.3.12 on page 31 to swap the sum and product above, notice that  $\mathcal{T} := \mathcal{D}_n$  is a finite set. For each  $f \in \mathcal{D}_n$ , take  $B_f$  to be the set of indicator vectors for the time periods  $\leq t$  at which  $f$  could enter the multigraph union:  $B_f := \{\mathbf{b} \in \{0,1\}^t \mid \sum_{i=1}^t b_i = w(f)\}$ . Further, take  $h_f: B_f \rightarrow \mathbb{R}$  such that  $h_f(\mathbf{b}) = \prod_{i=1}^t \mathbb{P}(Z_i(f) = b_i)$ . In the notation of lemma 2.3.12, this makes  $\mathcal{F}_{\mathcal{D}_n} = \{x \in \mathcal{G}_{n,1}^t \mid \sum_{i=1}^t x_i = \mathbf{w}\}$ , so we may interchange the sum and the product as follows.

$$\begin{aligned} \mathbb{P}(\mathbf{W} = \mathbf{w}) &= \prod_{f \in \mathcal{D}_n} \sum_{\substack{\mathbf{b} \in \{0,1\}^t \\ \sum_i b_i = w(f)}} \prod_{i=1}^t \mathbb{P}(Z_i(f) = b_i) && (2.57) \\ &= \prod_{f \in \mathcal{D}_n} \sum_{\substack{\mathbf{b} \in \{0,1\}^t \\ \sum_i b_i = w(f)}} \mathbb{P}(Z_1(f) = b_1, \dots, Z_t(f) = b_t) && (Z \text{ is IID}) \\ &= \prod_{f \in \mathcal{D}_n} \mathbb{P}\left(\sum_{i=1}^t Z_i(f) = w(f)\right) = \prod_{f \in \mathcal{D}_n} \mathbb{P}(W(f) = w(f)). && (\text{Law of total prob.}) \end{aligned}$$

Therefore the multiplicities of each dyad in  $\mathbf{W}$  are independent of each other.

To prove eq. (2.56), let  $\mu_f := \mathbb{P}(Z_i(f) = 1)$  for each  $f \in \mathcal{D}_n$ . Since  $Z$  is IID and dyadically independent,

$$\mathbb{P}(Z = z) = \prod_{f \in \mathcal{D}_n} \prod_{i=1}^t \mathbb{P}(Z_i(f) = z_i(f)) = \prod_{f \in \mathcal{D}_n} \mu_f^{w(f)} (1 - \mu_f)^{t-w(f)}.$$

Likewise, from eq. (2.57) and the combinatorial definition of the binomial coefficient, we have

$$\begin{aligned} \mathbb{P}(\mathbf{W} = \mathbf{w}) &= \prod_{f \in \mathcal{D}_n} \sum_{\substack{\mathbf{b} \in \{0,1\}^t \\ \sum_i b_i = w(f)}} \prod_{i=1}^t \mathbb{P}(Z_i(f) = b_i) = \prod_{f \in \mathcal{D}_n} \sum_{\substack{\mathbf{b} \in \{0,1\}^t \\ \sum_i b_i = w(f)}} \mu_f^{w(f)} (1 - \mu_f)^{t-w(f)} \\ &= \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)} \mu_f^{w(f)} (1 - \mu_f)^{t-w(f)} = \mathbb{P}(Z = z) \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)}. \quad \square \end{aligned}$$



We close this subsection by showing that taking multigraph unions preserves exponential family structure. The proof of the result is a straightforward combination of Casella and Berger (2002, Thm. 5.2.11) with eqs. (2.5) and (2.56), using the facts that  $\tau$  is dyadditive and that  $\kappa$  is dyadically multiplicative. Notice that  $Z$  does not appear in eq. (2.58).

**Theorem 2.5.9.** *Suppose the common PMF of the components of  $Z$  is  $\mu_\theta$  from eq. (2.5) on page 11 such that  $\tau$  is dyadditive as in eq. (2.50) and  $\kappa$  is dyadically multiplicative as in eq. (2.51). Further, assume that  $\eta(\Theta)$  contains an open,  $\ell$ -dimensional set. Then the PMF of  $\mathbf{W}$  also has an exponential family representation with the same parameter and parameter function, and with the sufficient statistic and carrier measure respectively equal to*

$$\sum_{f \in \mathcal{D}_n} \tau_f(1)W(f) - \tau_f(0)[t - W(f)] \quad \text{and} \quad \prod_{f \in \mathcal{D}_n} \binom{t}{W(f)} \kappa_f(1)^{W(f)} \kappa_f(0)^{t-W(f)}. \quad (2.58)$$

*Proof.* By Casella and Berger (2002, Thm. 5.2.11) and eq. (2.5),

$$\mathbb{P}(Z = z) = \prod_{i=1}^t \mu_\theta(z_i) = \exp\left(\eta(\theta) \cdot \sum_{i=1}^t \tau(z_i) - t\zeta(\theta)\right) \prod_{i=1}^t \kappa(z_i).$$

Next we plug this into theorem 2.5.8's eq. (2.56):

$$\mathbb{P}(\mathbf{W} = w) = \exp\left(\eta(\theta) \cdot \sum_{i=1}^t \tau(z_i) - t\zeta(\theta)\right) \left[ \prod_{i=1}^t \kappa(z_i) \right] \left[ \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)} \right].$$

Since  $\tau$  is dyadditive,

$$\begin{aligned} \sum_{i=1}^t \tau(z_i) &= \sum_{i=1}^t \sum_{f \in \mathcal{D}_n} \tau_f(z_i(f)) = \sum_{i=1}^t \sum_{f \in \mathcal{D}_n} z_i(f) \tau_f(1) + (1 - z_i(f)) \tau_f(0) \\ &= \sum_{f \in \mathcal{D}_n} \sum_{i=1}^t z_i(f) \tau_f(1) + (1 - z_i(f)) \tau_f(0) \\ &= \sum_{f \in \mathcal{D}_n} \left[ \tau_f(1) \sum_{i=1}^t z_i(f) + \tau_f(0) \sum_{i=1}^t (1 - z_i(f)) \right] \\ &= \sum_{f \in \mathcal{D}_n} \tau_f(1)w(f) + \tau_f(0)(t - w(f)). \end{aligned}$$

Since  $\kappa$  is dyadically multiplicative,

$$\begin{aligned}
\left[ \prod_{i=1}^t \kappa(\mathbf{z}_i) \right] \left[ \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)} \right] &= \left[ \prod_{i=1}^t \prod_{f \in \mathcal{D}_n} \kappa_f(\mathbf{z}_i(f)) \right] \left[ \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)} \right] \\
&= \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)} \prod_{i=1}^t \kappa_f(\mathbf{z}_i(f)) \\
&= \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)} \left[ \prod_{\substack{i=1 \\ \mathbf{z}_i(f)=1}}^t \kappa_f(1) \right] \left[ \prod_{\substack{i=1 \\ \mathbf{z}_i(f)=0}}^t \kappa_f(0) \right] \\
&= \prod_{f \in \mathcal{D}_n} \binom{t}{w(f)} \kappa_f(1)^{w(f)} \kappa_f(0)^{t-w(f)}. \quad \square
\end{aligned}$$

**2.5.3.3 Permutation Uniformity.** Suppose  $\tau: \mathcal{G}_{n,t} \times \mathcal{G}_{n,t} \rightarrow \mathbb{R}^\ell$  is  $\mathfrak{p}$ -uniform under  $\pi$ , so that, by lemma 2.4.2, there is a function  $\nu: \mathcal{G}_{n,t} \rightarrow \mathbb{R}^\ell$  such that  $\tau(a, b) = \nu(\pi_a b)$  for all  $a, b \in \mathcal{G}_{n,t}$ . We say that this two-argument function  $\tau$  is *dyadditive* if  $\mathbf{b} \mapsto \tau(a, \mathbf{b})$  is dyadditive for all  $a \in \mathcal{G}_{n,t}$ . Can we extend dyadditivity from  $\tau$  to  $\nu$  or from  $\nu$  to  $\tau$ ? Generally no.

**Example 2.5.10.** For the case of simple graphs ( $t = 1$ ) and scalar sufficient statistics ( $\ell = 1$ ), we show that just because  $\tau$  is dyadditive, it is not necessary that  $\nu$  is. Suppose  $n = 3$  and  $\tau(a, b) = |E(\mathbf{b})|$ , the number  $\mathbf{1} \cdot \mathbf{b}$  of edges of  $\mathbf{b}$ , which is dyadditive (cf. example 2.5.5). Set  $\nu$  and  $\pi$  so that

$$\nu(\mathbf{b}) := \begin{cases} |E(\mathbf{b})| & \text{if } \mathbf{b} \in \{\mathbf{0}, \mathbf{1}\} \\ |E(\bar{\mathbf{b}})| & \text{else,} \end{cases} \quad \pi_a \mathbf{b} := \begin{cases} \mathbf{b} & \text{if } \mathbf{b} \in \{\mathbf{0}, \mathbf{1}\} \\ \bar{\mathbf{b}} & \text{else,} \end{cases}$$

for all  $a, b \in \mathcal{G}_{3,1}$ . Hence  $\tau(a, b) = \nu(\pi_a b)$ .

Suppose by way of contradiction that  $\nu$  is dyadditive. Per the comment after eq. (2.50), there are real numbers  $q_1, q_2$ , and  $q_3$  such that  $\nu(\mathbf{b}) = \sum_{f \in E(\mathbf{b})} q_f$ . When  $\mathbf{b}$  has one edge,  $\nu(\mathbf{b}) = 2$ , so  $q_1 = q_2 = q_3 = 2$ . But when  $\mathbf{b}$  is the complete graph, we have  $3 = \nu(\mathbf{b}) = q_1 + q_2 + q_3 = 2 + 2 + 2$ , a contradiction.  $\square$

There is one case where we can guarantee that dyadditivity passes from  $\tau$  to  $\nu$ : when one of the permutations is the identity permutation, such as in example 2.4.19.

**Proposition 2.5.11.** *If  $\tau$  is dyadditive and, for some  $\mathbf{x} \in \mathcal{G}_{n,t}$ ,  $\pi_{\mathbf{x}}$  is the identity permutation, then  $\nu$  is dyadditive.*

*Proof.* For some  $\tau_f$ s and all  $\mathbf{b}$ s,  $\nu(\mathbf{b}) = \tau(\mathbf{x}, \pi_{\mathbf{x}}^{-1}\mathbf{b}) = \tau(\mathbf{x}, \mathbf{b}) = \sum_{f \in \mathcal{D}_n} \tau_f(\mathbf{x}, b(f))$ .  $\square$

**2.5.4 Examples from the Literature.** In this subsection, we run through the graph-valued Markov chains proposed in Hanneke et al. (2010), applying theory where fruitful. Throughout,  $\mathcal{S} = \mathcal{G}_{n,1}$ , and  $G = \{\mathbf{G}_i\}_{i \in \mathbb{N}}$  is a  $\mathcal{G}_{n,1}$ -valued stochastic process that is a Markov chain under the CEF  $\mathcal{P}$  of transition matrices  $P_{\theta}$  from eq. (2.7) on page 14. However, we set  $\kappa(\mathbf{a}, \mathbf{b}) = 1$  for all  $\mathbf{a}, \mathbf{b} \in \mathcal{G}_{n,1}$ , so  $P_{\theta}(\mathbf{a}, \mathbf{b}) = \exp(\eta(\theta) \cdot \tau(\mathbf{a}, \mathbf{b}) - \zeta(\theta))$ . Hanneke et al. name each example after natural parameter  $\gamma = \eta(\theta)$ , but we equivalently use the name for the sufficient statistic  $\tau$ . All examples in this subsection have scalar sufficient statistics, so  $\ell = d = 1$  and we just write  $\tau$  instead of  $\tau$  and  $\gamma$  instead of  $\gamma$ . Notationally,  $\mathbf{a}, \mathbf{b}, \mathbf{c}$  are generic, simple graphs in  $\mathcal{G}_{n,1}$ .

Scaling a sufficient statistic  $\tau$  by a constant  $c$ , often  $c = n$  or  $c = \frac{1}{n-1}$  so that  $\tau$  lies in  $[0, n]$ , occasionally improves interpretability while being transparent to the probability model because  $(\frac{1}{c}\gamma) \cdot (c\tau) = \gamma \cdot \tau$ . For example, in the Erdős-Rényi graph model of example 2.5.2 in which the sufficient statistic is the number of edges  $m$ , the natural parameter  $\gamma$  is the *log odds ratio*, or *logit* (Casella & Berger, 2002, p. 591),  $\eta(p) = \log \frac{p}{1-p}$  of the edge probability  $p$ . If the sufficient statistic is changed to  $cm$ , we obtain exactly the same model by taking the natural parameter to be  $\frac{1}{c}\eta(p) = \log \left( \frac{p}{1-p} \right)^{1/c}$ .

**Example 2.5.12 (Density).**  $\tau(\mathbf{a}, \mathbf{b}) := \frac{1}{n-1} \sum_{ij \in \mathcal{D}_n} b(ij)$  is called *density* (Hanneke et al., 2010, § 2.1). It is dyadditive because

$$\tau(\mathbf{a}, \mathbf{b}) = \frac{1}{n-1} |E(\mathbf{b})| = \frac{1}{n-1} \mathbf{1} \cdot \mathbf{b},$$

similar to the Erdős-Rényi graph model of example 2.5.2. In fact, since  $\tau(\mathbf{a}, \mathbf{b}) = \tau(\mathbf{c}, \mathbf{b})$ , so  $G$  is actually an IID sequence of Erdős-Rényi graphs, except that the natural parameter is  $\eta(p) = (n-1) \log \frac{p}{1-p}$  for some  $p \in [0, 1]$ . P-uniformity is more complicated than we need here. Applying theorem 2.5.9, let  $\mathbf{W} = \sum_{i=1}^t \mathbf{G}_i$ , which is a single  $t$ -ERMGM random variable with sufficient statistic equal to the total edge count divided by  $n-1$ . Then the

joint probability of  $G$  is proportional, per theorem 2.5.8, to the probability of  $W$ , in which every edge is added to the multigraph by flipping a  $p$ -weighted coin for each dyad  $t$  times. Put differently, for each dyad  $f$ ,  $W(f)$  is an independent, binomially distributed random variable with parameters  $t$  and  $p$ .  $\square$

For subsequent examples, table 2.1 reminds the reader of some standard notation from graph theory in terms of the edge-indicator vector notation we have been using.

Table 2.1. Graph theoretic operations on edge-indicator vectors

Operation	Definition
<i>Intersection</i>	$\mathbf{a} \cap \mathbf{b} := \min\{\mathbf{a}, \mathbf{b}\} = (a(f)b(f) \mid f \in \mathcal{D}_n)$
<i>Union</i>	$\mathbf{a} \cup \mathbf{b} := \max\{\mathbf{a}, \mathbf{b}\}$
<i>Complement</i>	$\bar{\mathbf{a}} := \mathbf{1} - \mathbf{a}$
<i>Relative Complement</i>	$\mathbf{a} \setminus \mathbf{b} := \mathbf{a} \cap \bar{\mathbf{b}}$
<i>Symmetric Difference</i>	$\mathbf{a} \Delta \mathbf{b} := (\mathbf{a} \setminus \mathbf{b}) \cup (\mathbf{b} \setminus \mathbf{a}) = \mathbf{a} + \mathbf{b} \pmod{2}$

$\mathbf{a}, \mathbf{b} \in \mathcal{G}_{n,1}$  are edge-indicator vectors, on which all operations are entry-wise. See example 2.4.19 for more on symmetric differences.

**Example 2.5.13** (Stability).  $\tau(\mathbf{a}, \mathbf{b}) := \frac{1}{n-1} \sum_{f \in \mathcal{D}_n} [b(f)a(f) + (1-b(f))(1-a(f))]$  is called *stability* (Hanneke et al., 2010, § 2.1). It measures the tendency of an edge to continue existing or not existing at time  $i+1$  if it was doing so at time  $i$  by counting the dyads that are not switching on or off. Mathematically, we are saying that

$$(n-1)\tau(\mathbf{a}, \mathbf{b}) = \mathbf{1} \cdot (\min\{\mathbf{a}, \mathbf{b}\} + \min\{\mathbf{1} - \mathbf{a}, \mathbf{1} - \mathbf{b}\}) = \left| E(\overline{\mathbf{a} \Delta \mathbf{b}}) \right|,$$

which shows that it is dyadditive. Stability is  $p$ -uniform because  $\mathbf{b} \mapsto \overline{\mathbf{a} \Delta \mathbf{b}}$  is bijective. As we shall see, the canonical parameter is  $p \in (0, 1)$ , and the natural parameter is  $\eta(p) = (n-1) \log \frac{p}{1-p}$ . Fix some value of  $p$  and write the transition matrix for  $G$  as  $P$ .

By observation 2.4.25, the transition matrix  $P$  is  $p$ -uniform. By corollary 2.4.24,  $P$  is drawn from a Markovian exponential family (MEF). By theorem 2.3.11, the joint distribution of  $G$  is drawn from an exponential family.

By theorem 2.4.5,  $\mathbf{Z}_{i+1} := \overline{\mathbf{G}_i \Delta \mathbf{G}_{i+1}}$ ,  $i \in [t-1]$ , is an IID sequence of  $\mathcal{G}_{n,1}$ -valued random variables. By proposition 2.4.23 and example 2.5.2, the common probability vector

$\mu$  of  $Z_{i+1}$  is the Erdős-Rényi graph model with sufficient statistic  $\frac{1}{n-1}|E(Z_{i+1})|$ , which is dyadditive, and natural parameter  $\eta(p) = (n-1) \log \frac{p}{1-p}$ . Thus lemma 2.5.6 implies that  $Z_{i+1}$  is dyadically independent.

By theorem 2.4.9, the mean parameter equals, for any  $i \in \mathbb{N}$ ,

$$\mathbb{E}[\tau(\mathbf{G}_i, \mathbf{G}_{i+1})] = \sum_{\mathbf{b} \in \mathcal{G}_{n,1}} \tau(\mathbf{0}, \mathbf{b})P(\mathbf{0}, \mathbf{b}) = \sum_{\mathbf{b} \in \mathcal{G}_{n,1}} \tau(\mathbf{0}, \bar{\mathbf{b}})P(\mathbf{0}, \bar{\mathbf{b}}) = \sum_{\mathbf{b} \in \mathcal{G}_{n,1}} \frac{1}{n-1} |E(\mathbf{b})| \mu(\mathbf{b}) = p \frac{N}{n-1},$$

where  $N = |\mathcal{D}_n|$ . That is, the average transition has the number of edges in the average Erdős-Rényi graph with parameter  $p$ .

Symmetric differences commute, so lemma 2.4.17 implies that  $P$  is symmetric.  $P$  is therefore doubly stochastic. Probabilistically, the most interesting conclusion is that the unique stationary distribution is uniform because  $\kappa \equiv 1$  implies that all entries of  $P$  are positive. See example 2.4.19 and corollary 2.4.21. Linear algebraically, we can say much more.

Since  $\overline{\mathbf{a}\Delta\mathbf{a}}$  is the complete graph with  $N = |\mathcal{D}_n|$  edges for any  $\mathbf{a} \in \mathcal{G}_{n,1}$ , the diagonal elements of  $P$  are  $P_{\mathbf{a}\mathbf{a}} = p^N$ . (The factors of  $n-1$  in the sufficient statistic and parameter function cancel out.) There are  $|\mathcal{G}_{n,1}| = 2^N$  diagonal entries, so the trace of  $P$  is  $2^N p^N$ . Every other off-diagonal entry of  $P$  has a  $(1-p)$  factor, so  $P \rightarrow I$  as  $p \rightarrow 1^-$  and  $P \rightarrow \frac{1}{2^N-1}(\mathbf{1}\mathbf{1}^\top - I)$  as  $p \rightarrow 0^+$ . At  $p = \frac{1}{2}$ , every row of  $P = \frac{1}{2^N} \mathbf{1}\mathbf{1}^\top$  is the uniform distribution.

The rank  $r$  of  $P$  approaches  $2^N$  as  $p$  approaches the boundary of the parameter space and is one in the middle, but we can also bound the subset of the parameter space in which  $r > 1$ . Since  $P$  is doubly stochastic, lemma 2.4.16 says that  $\|\mu\|_\infty 2^{N/2} / \sqrt{r} \leq \|P\|_2 = 1$ . Thus  $2^N \|\mu\|_\infty^2 \leq r$ . From example 2.5.2,  $\mu$  is monotone as a function of the number of edges in the graph: if  $p < \frac{1}{2}$ , then  $\|\mu\|_\infty = (1-p)^N$ , and if  $p \geq \frac{1}{2}$ , then  $\|\mu\|_\infty = p^N$ . Therefore,  $2^N \|\mu\|_\infty^2 = 2^N \max\{p, 1-p\}^{2N} \leq r$ . The lower bound is strictly decreasing in  $p$  when  $p < \frac{1}{2}$  and strictly increasing in  $p$  when  $p > \frac{1}{2}$ . If  $p < \frac{1}{2}$ , then  $r \geq 2^N (1-p)^{2N} > 1$  when  $p < 1 - 1/\sqrt{2} < 0.2929$ , and, if  $p \geq \frac{1}{2}$ , then  $r \geq 2^N p^{2N} > 1$  when  $p > 1/\sqrt{2} > 0.7071$ .

Applying theorem 2.5.9, let  $\mathbf{W} = \sum_{i=1}^t \mathbf{Z}_i$ , which is a single, dyadically independent  $t$ -ERMGM random variable with sufficient statistic equal to the total edge count divided by  $n-1$ . Then the joint probability of  $G$  is proportional, per theorem 2.5.8, to the probability of

*W.*

□

The next two examples are neither p-uniform nor MEFs.

**Example 2.5.14** (Reciprocity). Our focus has been on undirected graphs, but we briefly mention a statistic with applications to directed graphs. Defining  $0/0$  as zero,  $\tau(\mathbf{a}, \mathbf{b}) := n \left[ \sum_{ij \in \mathcal{D}_n} a(ij) \right]^{-1} \sum_{ij \in \mathcal{D}_n} b(ji) a(ij)$  is called *reciprocity* (Hanneke et al., 2010, § 2.1). Since the dyad  $ij$  is the same as the dyad  $ji$  for undirected graphs, we can write

$$\tau(\mathbf{a}, \mathbf{b}) = n \frac{|E(\mathbf{a} \cap \mathbf{b})|}{|E(\mathbf{a})|}.$$

This is dyadditive, and, by lemma 2.5.6,  $G_{i+1}$  is dyadically independent conditional on the value of  $G_i$ . Suppose  $n \geq 3$  so that  $N = |\mathcal{D}_n| \geq 2$ . To see that  $\tau$  is not p-uniform, let  $e_1, \dots, e_N$  be the standard basis of  $\mathcal{G}_{n,1}$ ; these are the graphs containing exactly one edge. Then  $\tau(e_1, \mathbf{b}) \in \{0, n\}$  whereas  $\tau(e_1 \cup e_2, e_1) = n/2$ . In fact, by violating the conclusion of proposition 2.3.16, this also shows that  $P$  is not drawn from an MEF. That proposition makes no reference to  $\kappa$  and its only reference to  $\eta$  is that  $\eta(\Theta)$  contains a number other than zero, so different choices of parameterization or carrier measure do not make  $P$  an MEF. Since it's still drawn from a CEF, theorem 2.3.13 gives its likelihood function. □

**Example 2.5.15** (Transitivity). Interpreting  $0/0$  as zero and taking sums over triples of vertices  $i < j < k$ ,  $\tau(\mathbf{a}, \mathbf{b}) := n \left[ \sum_{ijk} a(ij) a(jk) \right]^{-1} \sum_{ijk} b(ik) a(ij) a(jk)$  is called *transitivity* (Hanneke et al., 2010, § 2.1). Transitivity measures the tendency of edge  $ik$  to come into existence in the future (graph  $\mathbf{b}$ ) if edges  $ij$  and  $jk$  exist in the present (graph  $\mathbf{a}$ ). That is the fraction of  $\mathbf{a}$ 's paths of length two (counted in the denominator) that try to close the triangle in graph  $\mathbf{b}$ . We say “try” because edges  $ij$  and  $jk$  may or may not exist in  $\mathbf{b}$ ; we're only measuring if the edge  $ik$  exists in  $\mathbf{b}$ .

Transitivity is dyadditive:  $\mathbf{b}$  only shows up in each summand evaluated at one edge. By lemma 2.5.6,  $G_{i+1}$  is dyadically independent conditional on the value of  $G_i$ .

However, transitivity is not p-uniform and  $P$  is not drawn from an MEF when we suppose  $n \geq 3$  and that  $\eta(\Theta)$  contains a number other than zero. When  $\mathbf{a}$  is a graph containing exactly two edges and those edges are adjacent, forming a path of length two,

$\tau(\mathbf{a}, \mathbf{b})$  is either zero or  $n$ .  $\tau(\mathbf{0}, \mathbf{b}) = 0$  since we're interpreting  $0/0 = 0$ . The non-p-uniformity is not an artifact of this convention. Suppose  $c$  contains exactly one path of length three and no other edges;  $c$  then contains two paths of length two. If  $b$  contains an edge completing just one of those two triangles, then  $\tau(c, b) = n/2$ . Thus  $\tau$  is not p-uniform, and, by proposition 2.3.16 and the fact that  $\ell = 1$ ,  $P$  is not an MEF.<sup>17</sup> As in the reciprocity example above, different choices of parameterization or carrier measure do not make  $P$  an MEF. Finally, when  $\eta(\theta) \neq 0$ ,  $P_\theta$  is not p-uniform: the  $\mathbf{0}$ th row of  $P_\theta$  is the uniform distribution, but the  $a$ th row contains two different values.<sup>18</sup>  $\square$

**2.5.5 A Model of Loyalty.** Motivated by the theoretical developments of the previous sections, we introduce a new TERGM.

**Example 2.5.16 ( $\beta$  TERGM).** By analogy with example 2.5.3 on page 70, we may define an MEF using the degree sequences of functions of the current and next graph. Using the same notation as the previous subsection, replace  $d = \ell := n$ ;  $\Theta := \mathbb{R}_{>0}^n$ , the positive orthant;  $\eta := \log$ ; and  $\tau(\mathbf{a}, \mathbf{b}) := \beta(\pi_a \mathbf{b})$ , where  $\beta$  is the degree sequence. With these choices,  $\mathcal{P}$  is an MEF p-uniform under any permutations  $\pi$  on  $\mathcal{G}_{n,1}$ . We call the model the  $\beta$  TERGM under  $\pi$ .

For example, we may take use the symmetric difference operator as the permutations:

$$P_\theta(\mathbf{a}, \mathbf{b}) = \exp(\beta(\mathbf{a} \Delta \mathbf{b}) \cdot \log \theta - \zeta(\theta)) = \frac{\prod_{uv \in E(\mathbf{a} \Delta \mathbf{b})} \theta_u \theta_v}{\prod_{u=1}^n \prod_{v=1}^{u-1} (\theta_u \theta_v + 1)}.$$

$\mathbf{Z}_i := \mathbf{G}_{i-1} \Delta \mathbf{G}_i$  is an IID sequence of  $\beta$  ERGMS as in example 2.5.3. For a given vertex  $v \in [n]$ , as  $\theta_v$  increases, so does the probability edges lying on it appear in one or the other, but not both, of the current and next iteration of  $G$ . If an edge lying on  $v$  is in  $G_t$  and it appears in  $G_t \Delta G_{t+1}$ , then that dyad will not appear in  $G_{t+1}$ . Likewise, if a dyad lying on  $v$  is not in  $G_t$  and it appears in  $G_t \Delta G_{t+1}$ , then that edge will appear in  $G_{t+1}$ . Thus  $G$  changes the most in the neighborhoods of vertices  $v$  with high values of  $\theta_v$ . We might say that  $v$  is *loyal* if it has a low value of  $\theta_v$  (near zero) and is *disloyal* if it has a high value of  $\theta_v$ .

<sup>17</sup>↑ Since  $P$  is still drawn from a CEF, theorem 2.3.13 gives its likelihood function.

<sup>18</sup>↑ We cannot use theorem 2.4.26 to show that  $P_\theta$  is not p-uniform because we already have concluded that  $P$  is not an MEF.

## 2.6 Conclusion

The theorems of this thesis point to a new technique for analyzing time series of networks.

1. Verify that
  - the set of edges but not the set of vertices changes over time,
  - time can be meaningfully discretized so that changes that occur on the network between snapshots can be summarized by the state of the network at each snapshot, and
  - the network has no long-term memory so that the Markov property holds.
2. Find an invariant of the transitions between snapshots of the network that allows you to identify the transitions with some network-valued function of the current and next states of the network. For example, imagine observing that each dyad is as likely to go from on to off as it is to go from off to on. This suggests that transitions can be identified with a random process on the symmetric differences of the current and next states of the network.
3. If the invariant is in terms of dyads, carry on. If it is in terms of larger structures such as triads, the model will lack dyadic independence and thus not be subject to theorem 2.5.8, which we need.
4. Restate the research hypothesis in terms of this invariant to identify a dyadically independent  $p$ -uniform MEF. For example, if the research hypothesis is that certain types of nodes exhibit less frequent edge flip-flops than others, we might wish to regress the parameters from the  $\beta$  TERGM with respect to symmetric differences of example 2.5.16 on node type.
5. Transform the time-discretized network data according to the set of bijections from the chosen model. (This converts from  $X$  to  $Z$  in the language of theorem 2.4.5, and will reduce the number of network snapshots by one.)



6. Count the number of times each dyad appears in any of the post-transformation network snapshots and record the resulting  $t$ -ERMGM. (This converts from  $Z$  to  $W$  in the language of theorem 2.5.8, and will reduce the number of network snapshots to one.)
7. Estimate the parameters of the model on the  $t$ -ERMGM.
8. Test for goodness of fit on the  $t$ -ERMGM.

An extensive literature covers items 7 and 8, so we have not discussed it in this thesis so far. Fienberg and Rinaldo (2012a) thoroughly describes parameter estimation for ERGMS, but its theorems are easily extended to multigraphs. Petrović (2015) gives a high-level overview and references to the literature for goodness-of-fit testing for ERGMS, and, again, its discussion can be easily extended to multigraphs.

## Chapter 3

HYPOTHESIS TESTS FOR MIXED MEMBERSHIP STOCHASTIC BLOCK MODELS<sup>19</sup>**3.1 Model**

In this chapter we define a statistical network model, the *mixed-membership stochastic block models* (MMSBMs), which postulate that individuals in a network have different probabilities of forming relationships with other nodes based on some form of community membership that may differ depending on which individual initiates the relationship. It is an exponential random graph model in the sense introduced above. Rather than consider a time series of such graphs, we focus on observing just one snapshot of the network's topology. Our model broadly follows Airoldi et al. (2008) where it was first introduced. Those authors introduced the model from a Bayesian perspective and discussed posterior inference. Our goal is to offer an algorithm for (frequentist) hypothesis testing under the assumption that we observe the block assignments of all nodes. Testing the the *simple* null hypothesis that a particular parameter estimate is true for the observed network within an MMSBM is relatively straightforward, relying on a test statistic that we introduce and standard Monte Carlo simulation. To our knowledge, algorithms for testing a simple null hypothesis for fixed block assignments in an MMSBM have not yet been offered. We show results for some simulated data. However, testing the *general* null hypothesis that the true distribution lies in an MMSBM in the first place poses serious challenges. In particular, there is a divide in the statistics literature about how to test hypotheses like that. One position requires rejecting all simple null hypotheses to reject the general null hypothesis, which is equivalent to a certain optimization problem. The other position requires rejecting a specific simple null hypothesis to reject the general null hypothesis. The former is more conservative than the latter, but the latter is more typical in the discrete-statistics literature.

Section 3.1 describes the MMSBMs and introduces our notation, and section 3.2 develops an algorithm for testing it. Subsection 3.2.1 introduces our test statistic. Subsection 3.2.2

---

<sup>19</sup>This chapter includes joint work with Sonja Petrović, Debdeep Pati, and Vishesh Karwa: Karwa et al. (2021–present).

defines a p-value from that test statistic for both the simple and general null hypotheses. Subsection 3.2.3 presents an algorithm for computing the simple-hypothesis p-value and discusses the algorithm's convergence rate. Finally, subsection 3.2.4 discusses the two sides to the literature on hypothesis testing for general null hypotheses like ours.

**3.1.1 Basic Notation.** First we dispose of some bookkeeping notation and conventions. If  $\mathbf{a}$  and  $\mathbf{b}$  are matrices or vectors of the same shape, then  $\mathbf{a} \cdot \mathbf{b}$  denotes the sum product of their entries, i.e., the standard dot product for vectors or the Frobenius inner product for matrices. When we apply a function, such as  $\log$ , defined on scalars to a vector or matrix, we apply the function element-wise.  $\mathbb{N}$  is the set of nonnegative integers and  $\mathbb{R} \equiv (-\infty, \infty)$  is the set of real numbers. If  $n \in \mathbb{N}$ , then  $[n] := \{1, \dots, n\}$  and  $[n]_0 := \{0, 1, \dots, n\}$ . If  $X$  is a random variable and  $\mathcal{X}$  is a set, then  $\mathbb{1}(X \in \mathcal{X})$  is a random variable whose value is one when  $X \in \mathcal{X}$  and zero otherwise.

To simplify the statements of certain results, we commit to the following arithmetic conventions. We define  $0^0 := 1$ ,  $\log 0 := -\infty$ , and  $e^{-\infty} := 0$ . The usual rules of arithmetic in the extended reals  $\overline{\mathbb{R}} \equiv [-\infty, \infty]$  apply (Jacod & Protter, 2004, p. 24):  $x \pm \infty = \pm\infty$  for all  $x \in \mathbb{R} \cup \{\pm\infty\}$ ;  $0 \times \pm\infty = 0$ ;  $x \times \pm\infty = \mp\infty$  for all  $x \in [-\infty, 0)$ ; and  $x \times \pm\infty = \pm\infty$  for all  $x \in (0, \infty]$ .

Implicit in our statistical analysis throughout is a discrete measurable space  $(\mathcal{X}, 2^{\mathcal{X}})$ . Suppose  $\mathcal{Z}$  is a state space that is finite—we will only consider finite spaces. When we say that  $Z$  is a  *$\mathcal{Z}$ -valued random variable*, we mean that it is a function  $Z: \mathcal{X} \rightarrow \mathcal{Z}$ . That is, all random variables implicitly have a common domain. A probability mass function (PMF)  $\theta: \mathcal{Z} \rightarrow [0, 1]$  uniquely determines a probability measure, which we denote  $\mathbb{P}_\theta$ , on  $\mathcal{X}$  (Jacod & Protter, 2004, p. 27). Likewise, we denote the expectation and variance operators corresponding to  $\mathbb{P}_\theta$  as  $\mathbb{E}_\theta$  and  $\text{Var}_\theta$ , respectively. (We have no further cause to discuss  $\mathcal{X}$ .)

**3.1.2 State Space.** For the remainder of this chapter, fix positive  $k, n \in \mathbb{N}$  such that  $1 \leq k \leq n$  and  $n \geq 2$ .  $n$  is the *number of nodes* and  $k$  is the *number of blocks*, objects we define presently.

We now turn to notation and terminology for describing networks, which we also call *graphs*. Name the *nodes* of a network by numbers  $i \in [n]$ . For networks on the node set

$[n]$ , choose some  $\mathcal{D}_n \subseteq [n]^2$  to be the set of *allowed edges*. We write  $(i, j) \in [n]^2$  as  $ij$  as long as context clarifies that that we mean an allowed edge or an index of a matrix rather than  $i \times j$ . An allowed edge  $ii$  from a node  $i$  to itself is a *self-loop*. If we want to consider only *simple* graphs, then we stipulate that  $\mathcal{D}_n$  contains no self-loops. For *directed* graphs allowing self-loops,  $\mathcal{D}_n = [n]^2$ . From the perspective of a given node  $i$ ,  $i$  *sends*  $ij \in \mathcal{D}_n$ , which goes *out*, and  $i$  *receives*  $ji \in \mathcal{D}_n$ , which comes *in*. For *undirected* graphs, nodes both send and receive allowed edges, which go both out and in. To avoid ambiguity we use upper triangular indices: the set of undirected allowed edges including self-loops is  $\{ij \in [n]^2 \mid i \leq j\}$ . The set of simple, undirected allowed edges is  $\{ij \in [n]^2 \mid i < j\}$ , which corresponds one-to-one with the set of *dyads*, or node pairs  $\{i, j\} \subseteq [n]$  such that  $i \neq j$ .<sup>20</sup> We can in like manner form any other such restrictions, the collection of which that apply to  $\mathcal{D}_n$  in a given context we call the *sense* of  $\mathcal{D}_n$ . Table 3.1 summarizes the common cases. We describe  $\mathcal{D}_n$ 's sense by saying that  $\mathcal{D}_n$  is itself (*non-*)*simple* or (*un*)*directed*, as applicable.

Table 3.1. The set  $\mathcal{D}_n$  of allowed edges

$\mathcal{D}_n$	Non-Simple	Simple
<b>Directed</b>	$[n]^2$	$\{ij \in [n]^2 \mid i \neq j\}$
<b>Undirected</b>	$\{ij \in [n]^2 \mid i \leq j\}$	$\{ij \in [n]^2 \mid i < j\}$

The set  $\mathcal{D}_n$  of allowed edges on the node set  $[n]$  depends on the sense in which we mean *networks*.

The set of *adjacency matrices* of networks on the sets  $[n]$  of nodes and  $\mathcal{D}_n$  of allowed edges is  $\mathcal{G}_n \subseteq \{0, 1\}^{n \times n}$ , the set of  $n \times n$  zero-one matrices. We generically denote adjacency matrices  $\mathbf{a}$  throughout this chapter. If  $ij \in \mathcal{D}_n$ , then  $a_{ij} = 1$  means that  $ij$  is an *edge* in the graph corresponding to  $\mathbf{a}$ , and  $a_{ij} = 0$  means that  $ij$  is not an edge. We describe  $\mathcal{G}_n$ 's *sense* by saying that  $\mathcal{G}_n$  is itself (*non-*)*simple* or (*un*)*directed* whenever  $\mathcal{D}_n$  is. When  $\mathcal{G}_n$  is simple, we stipulate that all matrices in  $\mathcal{G}_n$  have zeros along the main diagonal. When  $\mathcal{G}_n$  is undirected, we stipulate that all matrices in  $\mathcal{G}_n$  are symmetric. Table 3.2 on the following page summarizes these restrictions.

<sup>20</sup>↑“Dyads are pairs of vertices and, in directed graphs, may take on three possible states: null (no directed edges), asymmetric (one directed edge), or mutual (two directed edges).” (Kolaczyk & Csárdi, 2014/2020a, pp. 53–54). “Holland and Leinhardt’s  $p_1$  model focuses on dyadic pairings and keeps track of whether node  $i$  links to  $j$ ,  $j$  to  $i$ , neither, or both.” (Goldenberg et al., 2010, p. 162; referring to Holland and Leinhardt, 1981).

Table 3.2. The set  $\mathcal{G}_n$  of adjacency matrices

$\mathcal{D}_n$	Non-Simple	Simple
<b>Directed</b>	$\{0, 1\}^{n \times n}$	$\{\mathbf{a} \in \{0, 1\}^{n \times n} \mid a_{ii} = 0 \text{ for } i \in [n]\}$
<b>Undirected</b>	$\{\mathbf{a} \in \{0, 1\}^{n \times n} \mid \mathbf{a} = \mathbf{a}^\top\}$	$\{\mathbf{a} \in \{0, 1\}^{n \times n} \mid a_{ii} = 0 \text{ for } i \in [n], \mathbf{a} = \mathbf{a}^\top\}$

The set  $\mathcal{G}_n$  of adjacency matrices of networks on the sets  $[n]$  of nodes and  $\mathcal{D}_n$  of allowed edges. The sense of  $\mathcal{G}_n$  determines the symmetry and sparsity patterns of its member matrices.

We assume that each node  $i$  belongs to one of  $k$  communities called *blocks* per allowed edge  $ij$  that it sends and per allowed edge  $ji$  that it receives. Equivalently each allowed edge  $ij$  belongs to two blocks, one named  $\sigma \in [k]$  from  $i$ 's perspective as a sender and one named  $\rho \in [k]$  from  $j$ 's perspective as a receiver. A *sender-block matrix*  $z_{\rightarrow}$  and a *receiver-block matrix*  $z_{\leftarrow}$  record these associations. When considering the allowed edge  $ij$  that node  $i$  sends to a receiver node  $j$ , node  $i$  identifies as a member of block  $z_{\rightarrow j}$  and node  $j$  identifies as a member of block  $z_{\leftarrow i}$ .<sup>21</sup> Thus  $z_{\rightarrow}$  and  $z_{\leftarrow}$  are  $[k]_0$ -valued  $n \times n$  matrices with the following sparsity and symmetry requirements depending on the sense of  $\mathcal{G}_n$ . If  $\mathcal{G}_n$  permits self-loops, then no entry of  $z_{\rightarrow}$  or  $z_{\leftarrow}$  may be zero; if  $\mathcal{G}_n$  is simple, then off-diagonal entries are non-zero and the diagonal is zero. If  $\mathcal{G}_n$  is undirected, then we require  $z_{\rightarrow}$  and  $z_{\leftarrow}$  to be symmetric (after all, every allowed edge in an undirected graph goes both out and in).<sup>22</sup> For other constraints  $ij \notin \mathcal{D}_n$ , zeros pad the remaining entries  $z_{\rightarrow j}$  and  $z_{\leftarrow i}$ . Denote the set of these matrices  $\mathcal{B}_{n,k}$ . Table 3.3 on the next page summarizes these definitions.

A *block-assignments* array, which we generically denote  $z$  throughout this chapter, comprises a sender-block matrix  $z_{\rightarrow}$  and a receiver-block matrix  $z_{\leftarrow}$ . When  $\mathcal{G}_n$  is undirected, those nodes cannot tell whether they are sending or receiving along  $ij$ , so  $z_{\rightarrow j} = z_{\leftarrow i} = z_{\leftarrow j} = z_{\rightarrow i}$ . Thus  $z \in \{\rightarrow, \leftarrow\} \times \mathcal{B}_{n,k}$ , but when  $\mathcal{G}_n$  is undirected we require  $z_{\rightarrow} = z_{\leftarrow}^\top = z_{\leftarrow} = z_{\rightarrow}^\top$ .

<sup>21</sup>↑For readers already accustomed to the notation of Airoldi et al. (2008, p. 1984), our sender- and receiver-block matrix notation perhaps looks confusingly similar to Airoldi et al.'s subtly different "latent group indicator" notation. They considered only directed graphs that permitted self-loops (but see footnote 22). Our  $z_{\rightarrow j}$  and  $z_{\leftarrow i}$  denote the names in  $[k]$  of the respective sender and receiver blocks. Airoldi et al.'s  $\vec{z}_{i \rightarrow j}$  and  $\vec{z}_{i \leftarrow j}$  denoted indicator vectors in  $\{0, 1\}^k$ , each containing a one at the index corresponding to the name in  $[k]$  of the respective block and zeros elsewhere. If  $e_1, \dots, e_k$  is the standard basis of  $\mathbb{R}^k$ , then the two notations relate to each other as  $\vec{z}_{i \rightarrow j} = e_{z_{\rightarrow j}}$  and  $\vec{z}_{i \leftarrow j} = e_{z_{\leftarrow i}}$ . See also footnote 23.

<sup>22</sup>↑"Also note that the pairs of group memberships that underlie interactions need not be equal; this fact is useful for characterizing asymmetric interaction networks. Equality may be enforced when modeling symmetric interactions." (Airoldi et al., 2008, p. 1985).

Table 3.3. The set  $\mathcal{B}_{n,k}$  of sender- or receiver-block matrices

$\mathcal{D}_n$	Non-Simple	Simple
<b>Directed</b>	$[k]^{n \times n}$	$\{\mathbf{b} \in [k]_0^{n \times n} \mid b_{ij} = 0 \iff i = j\}$
<b>Undirected</b>	$\{\mathbf{b} \in [k]^{n \times n} \mid \mathbf{b} = \mathbf{b}^\top\}$	$\{\mathbf{b} \in [k]_0^{n \times n} \mid (b_{ij} = 0 \iff i = j), \mathbf{b} = \mathbf{b}^\top\}$

The sense of  $\mathcal{G}_n$  determines the symmetry and sparsity patterns of  $\mathcal{B}_{n,k}$ 's members. The table builds sets with a dummy variable  $\mathbf{b}$ , but we generally denote sender- and receiver-block matrices with  $\mathbf{z}_\rightarrow$  and  $\mathbf{z}_\leftarrow$  respectively.

Denote the set of block-assignments arrays as  $\mathcal{B}_{n,k}$ . Table 3.4 summarizes this definition.

Table 3.4. The set  $\mathcal{B}_{n,k}$  of block-assignments arrays

$\mathcal{D}_n$	Non-Simple	Simple
<b>Directed</b>	$\{\rightarrow, \leftarrow\} \times \mathcal{B}_{n,k}$	$\{\rightarrow, \leftarrow\} \times \mathcal{B}_{n,k}$
<b>Undirected</b>	$\{z \in \{\rightarrow, \leftarrow\} \times \mathcal{B}_{n,k} \mid z_\rightarrow = z_\leftarrow\}$	$\{z \in \{\rightarrow, \leftarrow\} \times \mathcal{B}_{n,k} \mid z_\rightarrow = z_\leftarrow\}$

A block-assignments array  $z$  comprises a sender-block matrix  $z_\rightarrow$  and a receiver-block matrix  $z_\leftarrow$ . What matrices  $\mathcal{B}_{n,k}$  allows depends only on whether  $\mathcal{G}_n$  is directed, not on whether it permits self-loops.

The random block-assignments array  $\mathbf{Z}$  is a  $\mathcal{B}_{n,k}$ -valued random variable whose sender- and receiver-block matrices we write as  $\mathbf{Z}_\rightarrow$  and  $\mathbf{Z}_\leftarrow$  respectively.<sup>23</sup> The random *network* is the  $\mathcal{G}_n$ -valued random variable that we denote as  $\mathbf{A}$ . As we will often consider the random pair  $(\mathbf{A}, \mathbf{Z})$ , we define the *state space*  $\mathcal{Y}_{n,k} := \mathcal{G}_n \times \mathcal{B}_{n,k}$ .

**3.1.3 Probability Distributions.** In this subsection, we establish notation for the distribution of  $(\mathbf{A}, \mathbf{Z})$ . Let  $\Theta$  be the set of all PMFS on  $\mathcal{Y}_{n,k}$ :

$$\Theta := \left\{ \theta: \mathcal{Y}_{n,k} \rightarrow [0, 1] \mid \sum_{(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}} \theta(\mathbf{a}, \mathbf{z}) = 1 \right\}.$$

$(\mathbf{A}, \mathbf{Z})$  has an unknown, *true PMF*  $\theta^* \in \Theta$ . We also think of  $\theta^*$  as the *unknown parameter* of the parameterized probability measure  $\mathbb{P}_{\theta^*}$  of  $(\mathbf{A}, \mathbf{Z})$  from among the parameterized set of probability measures  $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ .

We are primarily interested in situations when  $\theta^*$  lies in a particular, parameterized subset of  $\Theta$ , which we now describe starting with its parameter space. It depends on the

<sup>23</sup>↑ Another (see footnote 21) subtle distinction between the notation here versus in Airolidi et al. (2008, pp. 1984–1985) is that here we name random variables with capital letters. We write arbitrary sender- and receiver-block matrices as  $z_\rightarrow$  and  $z_\leftarrow$  with entries  $z_{\rightarrow ij}$  and  $z_{\leftarrow ij}$ . We write the two sender- and receiver-block-matrix-valued random variables under consideration here as  $\mathbf{Z}_\rightarrow$  and  $\mathbf{Z}_\leftarrow$  with entries  $Z_{\rightarrow ij}$  and  $Z_{\leftarrow ij}$ . In contrast Airolidi et al. wrote the “latent group indicator”-valued random variables they considered as  $\vec{z}_{i \rightarrow j}$  and  $\vec{z}_{i \leftarrow j}$ . They wrote the collections of those random indicator vectors as  $Z_\rightarrow$  and  $Z_\leftarrow$ .

sense of  $\mathcal{G}_n$ . If  $\mathcal{G}_n$  is directed, let  $\Psi_k$  be the set  $[0, 1)^{k \times k}$  of  $k \times k$  matrices with nonnegative entries less than one; if  $\mathcal{G}_n$  is undirected, let  $\Psi_k$  be the subset of  $[0, 1)^{k \times k}$  whose matrices are symmetric. We generically denote  $\Psi_k$ 's matrices  $\psi$  and call them *edge probabilities*.<sup>24</sup> An  $n \times k$  *stochastic matrix* is a matrix in  $[0, 1]^{n \times k}$  whose rows each sum to one. If  $\mathcal{G}_n$  is directed, let  $\Pi_{n,k}$  be the set of  $n \times k$  stochastic matrices; if  $\mathcal{G}_n$  is undirected, let  $\Pi_{n,k}$  be the subset of  $n \times k$  stochastic matrices  $\pi$  such that every row  $\pi_i := (\pi_{i1}, \dots, \pi_{ik})$  of  $\pi$  is the same. We generically denote  $\Pi_{n,k}$ 's matrices  $\pi$  and call them *block probabilities*.

**Definition 3.1.1.** The *latent mixed-membership stochastic block model* (MMSBM) is  $\Theta$ 's subset  $\mathcal{M}$  of PMFS  $\mu_{\psi\pi}$  on  $\mathcal{Y}_{n,k}$  parameterized by edge probabilities  $\psi \in \Psi_k$  and block probabilities  $\pi \in \Pi_{n,k}$  such that the following conditions hold for all PMFS  $\mu \in \mathcal{M}$ , all allowed edges  $ij \in \mathcal{D}_n$ , and all states  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ .

(a)  $\mathbf{Z}$ 's entries are mutually independent:      (d)  $\mathbf{A}$ 's entries are mutually independent conditional on  $\mathbf{Z}$ :

$$\mathbb{P}_\mu(\mathbf{Z} = \mathbf{z}) = \prod_{ij \in \mathcal{D}_n} \mathbb{P}_\mu(Z_{i\leftarrow} = z_{i\leftarrow}) \mathbb{P}_\mu(Z_{i\rightarrow} = z_{i\rightarrow}).$$

(b)  $\pi_i$  is the PMF for  $Z_{i\rightarrow}$  in row  $i$  of  $\mathbf{Z}_{\rightarrow}$ :

$$\mathbb{P}_\mu(Z_{i\rightarrow} = z_{i\rightarrow}) = \pi_{iz_{i\rightarrow}}.$$

(c)  $\pi_j$  is the PMF for  $Z_{i\leftarrow}$  in column  $j$  of  $\mathbf{Z}_{\leftarrow}$ :

$$\mathbb{P}_\mu(Z_{i\leftarrow} = z_{i\leftarrow}) = \pi_{jz_{i\leftarrow}}.$$

$$\mathbb{P}_\mu(\mathbf{A} = \mathbf{a} \mid \mathbf{Z} = \mathbf{z}) = \prod_{ij \in \mathcal{D}_n} \mathbb{P}_\mu(A_{ij} = a_{ij} \mid \mathbf{Z} = \mathbf{z}).$$

(e) Conditional on  $\mathbf{Z}$ ,  $A_{ij}$  is Bernoulli distributed with parameter  $\psi_{Z_{i\rightarrow}Z_{i\leftarrow}}$ :

$$\mathbb{P}_\mu(A_{ij} = a_{ij} \mid \mathbf{Z} = \mathbf{z}) = \begin{cases} \psi_{z_{i\rightarrow}z_{i\leftarrow}} & \text{if } a_{ij} = 1 \\ 1 - \psi_{z_{i\rightarrow}z_{i\leftarrow}} & \text{if } a_{ij} = 0. \end{cases}$$

Saying “for all  $ij \in \mathcal{D}_n$ ” takes care of the simple- $\mathcal{G}_n$  case in which  $A_{ii}$ ,  $Z_{i\leftarrow}$ , and  $Z_{i\rightarrow}$  are deterministically zero for all  $i \in [n]$  as well as the undirected case in which their lower triangles of  $\mathbf{A}$  and of  $\mathbf{Z}_{\leftarrow} = \mathbf{Z}_{\rightarrow}$  are equal to the respective upper triangles. Saying “for all  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ ” takes care of the undirected case in which  $z_{\leftarrow} = z_{\rightarrow}$  is symmetric.

Conditions (b) and (c) say that  $\pi_{i\sigma}$  is the propensity of node  $i$  to send or receiver

<sup>24</sup>↑ Airolidi et al. (2008, p. 1984) called the edge probabilities the “matrix of Bernoulli rates” and denoted it as  $B$  rather than  $\psi$ .

in block  $\sigma$ , so the probability that an allowed edge  $ij$  sends in  $\sigma$  and receives in block  $\rho$  is  $\pi_{i\sigma}\pi_{j\rho}$  by condition (a). In Airolidi et al. (2008, p. 1984), the rows  $\pi_i$  are themselves random variables with a Dirichlet distribution. We eschew this Bayesian view in favor of the frequentist’s fixed-but-unknown-parameter approach. Other than that and the precise notation (see footnotes 21, 23 and 24), our latent MMSBM matches Airolidi et al.’s “mixed membership stochastic blockmodel (MMB)” (Airolidi et al., 2008, p. 1984).

Definition 3.1.1 motivates the restrictions on  $\Psi_k$  and  $\Pi_{n,k}$  when  $\mathcal{G}_n$  is undirected. We require  $\psi$  to be symmetric because  $\mathbf{Z}_{\rightarrow} = \mathbf{Z}_{\leftarrow}$ , so  $\psi_{Z_{\overrightarrow{ij}}Z_{\overrightarrow{ij}}} = \psi_{Z_{\overleftarrow{ij}}Z_{\overleftarrow{ij}}}$  perforce. Additionally,  $\mathbf{Z}_{\leftarrow} = \mathbf{Z}_{\overleftarrow{\quad}}$  implies that  $\mathbb{P}_{\mu}(Z_{\overleftarrow{ij}} = \rho) = \mathbb{P}_{\mu}(Z_{\overrightarrow{ji}} = \rho)$  for all  $ij \in \mathcal{D}_n$  and  $\rho \in [k]$ . By definition 3.1.1’s conditions (b) and (c),  $\pi_{i\rho} = \mathbb{P}_{\mu}(Z_{\overleftarrow{ij}} = \rho)$  and  $\pi_{j\rho} = \mathbb{P}_{\mu}(Z_{\overrightarrow{ji}} = \rho)$ , so  $\pi_{i\rho} = \pi_{j\rho}$ . Hence every row of  $\pi$  is the same.

**Example 3.1.2.** The *Erdős-Rényi graph model* is the family of PMFs  $\mu_p$  over simple, undirected networks and parameterized by an edge probability  $p \in [0, 1)$  such that  $\mu_p$  chooses edges of the random network  $\mathbf{A}$  independently each with probability  $p$ . It is a sub-model of the latent MMSBM  $\mathcal{M}$  in the sense that  $\mu_p(\mathbf{a}) = \mathbb{P}_{\mu_{\psi}\pi}(\mathbf{A} = \mathbf{a})$  for any block probabilities  $\pi$  and any network  $\mathbf{a}$  if the edge probabilities  $\psi$  is the matrix of all  $ps$ .

To facilitate situations in which we observe block assignments  $\mathbf{Z} = z$ , definition 3.1.3 on the next page sets up notation for conditional models and their PMFs. Assuming that  $(\mathbf{A}, \mathbf{Z})$ ’s true PMF  $\theta^*$  is in a latent MMSBM, we can observe  $\mathbf{Z} = z$  only if its probability is positive under a latent MMSBM PMF  $\mu_{\psi}\pi$  for some block probabilities  $\pi$  and some edge probabilities  $\psi$ :

$$\mathbb{P}_{\mu_{\psi}\pi}(\mathbf{Z} = z) \neq 0. \quad (3.1)$$

Definition 3.1.1’s conditions (a) to (c) guarantee that whether eq. (3.1) holds doesn’t depend on  $\psi$ . By choosing each row of  $\pi$  to be, say, the uniform distribution on  $[k]$ , we may always find a  $\pi$  that achieves eq. (3.1) for any  $z$ . This allows us to write down latent MMSBM probabilities conditional on block assignments without dependence of the the conditional probability’s *existence* on the latent MMSBM parameters. Moreover, conditioning on  $\mathbf{Z} = z$  completely removes the dependence of  $\mathbb{P}_{\mu_{\psi}\pi}$  on  $\pi$  as long as eq. (3.1) holds because of



definition 3.1.1's conditions (d) and (e). This ensures that eq. (3.2) below makes sense.

**Definition 3.1.3.** Fix block assignments  $z$  for which there is some network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . For each PMF  $\theta \in \Theta$  such that  $\mathbb{P}_\theta(\mathbf{Z} = z) \neq 0$ , define the *conditional PMF*  $\theta^z: \mathcal{G}_n \rightarrow [0, 1]$  by  $\theta^z(\mathbf{a}) := \mathbb{P}_\theta(\mathbf{A} = \mathbf{a} \mid \mathbf{Z} = z)$  for each  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . Denote the set of all such conditional PMFs as  $\Theta^z$ . For all edge probabilities  $\psi \in \Psi_k$  and any block probabilities  $\pi \in \Pi_{n,k}$  for which eq. (3.1) holds, we define the conditional PMF corresponding to the latent MMSBM PMF  $\mu_{\psi\pi} \in \mathcal{M}$  as  $\gamma_\psi^z: \mathcal{G}_n \rightarrow [0, 1]$  such that

$$\gamma_\psi^z(\mathbf{a}) := \mathbb{P}_{\mu_{\psi\pi}}(\mathbf{A} = \mathbf{a} \mid \mathbf{Z} = z) \quad (3.2)$$

for each  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . We define the *z-conditional MMSBM* as the set  $\mathcal{M}^z$  of conditional PMFs  $\gamma_\psi^z$  parameterized by  $\psi$ . When clear from context that  $\mathbf{Z} = z$ , we write  $\mathbb{P}_\psi$ ,  $\mathbb{E}_\psi$ , and  $\text{Var}_\psi$  in place of  $\mathbb{P}_{\gamma_\psi^z}$ ,  $\mathbb{E}_{\gamma_\psi^z}$ , and  $\text{Var}_{\gamma_\psi^z}$ .

In the notation of definition 3.1.3, the random network  $\mathbf{A}$ 's *true conditional PMF*, conditional on  $\mathbf{Z} = z$ , is  $\theta^{*z} \in \Theta^z$ .

*Remark 3.1.4.* We can rewrite the  $z$ -conditional MMSBM as a multinomial model as follows. For each  $i, j \in [n]$  such that  $ij$  or  $ji$  are allowed edges, let  $m_{\{i,j\}} \in \{2, 3, 4\}$  be the number of possible states, called *cells* (Fienberg & Rinaldo, 2012b, p. 1000), that that the vector  $(a_{ij}, a_{ji})$  may take across all networks  $\mathbf{a}$  for which  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . In the directed case with  $i \neq j$ , then  $m_{\{i,j\}} = 4$ : the possible states are  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ , and  $(1, 1)$ . In the non-simple case with  $i = j$  or the undirected case with  $i < j$ ,  $m_{\{i,j\}} = 2$ : the possible states are  $(0, 0)$  and  $(1, 1)$ . Define  $\mathbf{X}_{\{i,j\}}$  to be an  $m_{\{i,j\}}$ -length indicator vector, called a *contingency table* (Fienberg & Rinaldo, 2012b, p. 1000), for the states in lexicographic order containing exactly a single one in the entry corresponding to the state of  $(A_{ij}, A_{ji})$ . For example in the directed case with  $i \neq j$ , if  $(A_{ij}, A_{ji}) = (1, 0)$ , then  $\mathbf{X}_{\{i,j\}} = (0, 1, 0, 0)$ . Under the  $z$ -conditional MMSBM with edge probabilities  $\psi$ ,  $\mathbf{X}_{\{i,j\}}$  has a *multinomial* distribution with one trial (Casella & Berger, 2002, Def. 4.6.2 on p. 180). The *cell probabilities*  $p_{\{i,j\}}$ , vectors indexed by cell, depend on the sense of the networks and the cell. For example in the directed case with  $i \neq j$ , the  $(1, 1)$  cell, i.e.,  $\mathbf{X}_{\{i,j\}} = (0, 0, 0, 1)$ , has probability  $p_{\{i,j\}}(1, 1) := \psi_{z_{ij}z_{ji}}\psi_{z_{ji}z_{ij}}$ . In the directed case

with  $i < j$ , the  $(1, 1)$  cell, i.e.,  $\mathbf{X}_{\{i,j\}} = (0, 1)$ , has probability  $p_{\{i,j\}}(1, 1) := \psi_{z_{\overrightarrow{ij}} z_{\overleftarrow{ij}}}$ . The sum of  $\mathbf{p}_{\{i,j\}}$ 's entries is one, so  $\mathbf{p}_{\{i,j\}}$  lies in the standard simplex  $\Delta_{m_{\{i,j\}}-1} \subseteq \mathbb{R}^{m_{\{i,j\}}}$  (Petrović et al., 2010, p. 263).

**3.1.4 Sufficient Statistics.** We now find sufficient statistics (the definition of which we review below) for  $\mathcal{M}$  (proposition 3.1.11) and  $\mathcal{M}^z$  (lemma 3.1.8). To this end, definition 3.1.5 lays out statistics that will be our mainstays throughout the remainder of this report.

**Definition 3.1.5.** Define matrix-valued functions  $N: \mathcal{B}_{n,k} \rightarrow \mathbb{N}^{k \times k}$  and  $M: \mathcal{Y}_{n,k} \rightarrow \mathbb{N}^{k \times k}$  to give the number of allowed edges or edges, respectively, that are sending and receiving in each sending/receiving block pair  $\sigma, \rho \in [k]$ :

$$N_{\sigma\rho}(z) := \sum_{ij \in \mathcal{D}_n} \mathbb{1}(z_{\overrightarrow{ij}} = \sigma) \mathbb{1}(z_{\overleftarrow{ij}} = \rho),$$

$$M_{\sigma\rho}(\mathbf{a}, z) := \sum_{ij \in \mathcal{D}_n} \mathbb{1}(z_{\overrightarrow{ij}} = \sigma) \mathbb{1}(z_{\overleftarrow{ij}} = \rho) a_{ij}.$$

In the conditional model,  $M$  and  $N$  count the elements of block-membership equivalence classes. We will need this fact (and some notation) frequently enough that we present it as lemma 3.1.6.

**Lemma 3.1.6.** Fix block assignments  $z$  for which there is some  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . The set  $\mathcal{D}_n$  of allowed edges is the disjoint union over sending/receiving block pairs  $\sigma, \rho \in [k]$  of the **block-assignment equivalence classes**  $\mathcal{E}_{\sigma\rho}^z := \{ij \in \mathcal{D}_n \mid z_{\overrightarrow{ij}} = \sigma, z_{\overleftarrow{ij}} = \rho\}$ .  $N_{\sigma\rho}(z)$  counts the number  $|\mathcal{E}_{\sigma\rho}^z|$  of allowed edges  $ij \in \mathcal{E}_{\sigma\rho}^z$ .  $M_{\sigma\rho}(\mathbf{a}, z)$  counts the number of them that are edges in  $\mathbf{a}$ :  $M_{\sigma\rho}(\mathbf{a}, z) = |\{ij \in \mathcal{E}_{\sigma\rho}^z \mid a_{ij} = 1\}|$ .

*Proof.* That  $\mathcal{D}_n$  equals the disjoint union  $\sqcup_{\sigma=1}^k \sqcup_{\rho=1}^k \mathcal{E}_{\sigma\rho}^z$  follows from the fact that  $ij \mapsto z_{\overrightarrow{ij}}$  and  $ij \mapsto z_{\overleftarrow{ij}}$  are functions, so each allowed edge  $ij$  is in exactly one sending/receiving block pair  $z_{\overrightarrow{ij}}, z_{\overleftarrow{ij}} \in [k]$ . The counting formulas follow directly from definition 3.1.5 and the fact that  $\mathbb{1}(Z_{\overrightarrow{ij}} = \sigma) \mathbb{1}(Z_{\overleftarrow{ij}} = \rho) = 1$  if and only if  $ij \in \mathcal{E}_{\sigma\rho}^z$ .  $\square$

*Remark 3.1.7.* One takeaway from lemma 3.1.6 is that, for cells of the form  $(\sigma, \rho) \in [k]^2$ ,  $N$  is a contingency table. For the moment, focus on undirected networks, and suppose that  $Z$  is distributed according to a latent MMSBM distribution. The probability that an allowed

edge  $ij$  sends in block  $\sigma$  and receives in block  $\rho$  is  $\pi_{i\sigma}\pi_{j\rho}$  per definition 3.1.1's conditions (a) to (c). Every row of the block probabilities matrix  $\pi$  is the same because  $\mathcal{D}_n$  is undirected, so  $\pi_{i\sigma}\pi_{j\rho} = \pi_{1\sigma}\pi_{1\rho}$ . Since lemma 3.1.6 says that block pairs, which index  $N$ , name parts of a partition of  $\mathcal{D}_n$ ,  $N(\mathbf{Z})$  has a multinomial distribution with  $|\mathcal{D}_n|$  trials and cell probabilities  $\pi_{1\sigma}\pi_{1\rho}$ . In particular,  $N$  is the contingency table under a multinomial sampling scheme with the constraint that  $\sum_{\sigma\rho} N_{\sigma\rho}(z) = |\mathcal{D}_n|$  (Fienberg & Rinaldo, 2012b, p. 1001).

For the next lemma, which is the main result of this subsection, we will need the following terminology. Just to set up the definitions, suppose  $\emptyset \neq \mathcal{L} \subseteq \mathbb{R}^d$  for some positive integer  $d$ . A parameterized set  $\mathcal{P} := \{p_\lambda\}_{\lambda \in \mathcal{L}}$  of PMFs (or probability densities) on a set  $\mathcal{X}$  is an *exponential family* if there exist a positive integer  $\ell \geq d$  and functions  $\eta: \mathcal{L} \rightarrow \mathbb{R}^\ell$ ,  $\tau: \mathcal{X} \rightarrow \mathbb{R}^\ell$ ,  $\zeta: \mathcal{L} \rightarrow \mathbb{R}$ , and  $\kappa: \mathcal{X} \rightarrow [0, \infty)$ , such that the probability density of any state  $x \in \mathcal{X}$  under  $p_\lambda$  for any  $\lambda \in \mathcal{L}$  is

$$p_\lambda(x) = \kappa(x) \exp(\eta(\lambda) \cdot \tau(x) - \zeta(\lambda)). \quad (3.3)$$

Equation (3.3) is the exponential-family *representation* of  $\mathcal{P}$  (a quick introduction to exponential families is available from standard textbooks such as Casella & Berger, 2002, § 3.4; see Barndorff-Nielsen, 1978, Ch. 8, for a detailed investigation).  $\tau$  satisfies (Casella & Berger, 2002, Thm. 6.2.10) the definition of being a *sufficient statistic* for  $\lambda$  in  $\mathcal{P}$ :  $p_\lambda(x)$  does not depend on  $\lambda$  as long as  $\tau(x)$  is held constant (Casella & Berger, 2002, Def. 6.2.1). Finally, the *logit* function  $\text{logit}: [0, 1) \rightarrow \overline{\mathbb{R}}$  is defined by  $\text{logit } p := \log \frac{p}{1-p}$  for  $p \in (0, 1)$  (Casella & Berger, 2002, p. 591), and we take the convention that  $\text{logit } 0 := -\infty$  to match our convention that  $\log 0 = -\infty$ .

**Lemma 3.1.8.** *Fix block assignments  $z$  for which there is some  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . Then the  $z$ -conditional MMSBM  $\mathcal{M}^z$  is an exponential family with the representation*

$$\gamma_\psi^z(\mathbf{a}) = \exp(\text{logit}(\psi) \cdot \mathbf{M}(\mathbf{a}, z) + \log(\mathbf{1}\mathbf{1}^\top - \psi) \cdot \mathbf{N}(z)) \quad (3.4)$$

*for all edge probabilities  $\psi \in \Psi_k$  and all networks  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . In particular,  $\mathbf{M}(\mathbf{A}, z)$  is a sufficient statistic for  $\psi$  in  $\mathcal{M}^z$ .*

*Proof.* Fix  $\psi \in \Psi_k$  and  $\mathbf{a}$  such that  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ . For some  $\pi \in \Pi_{n,k}$ , eq. (3.2) holds, to the right side of which we may then apply definition 3.1.1's condition (d):

$$\gamma_{\psi}^{\mathbf{z}}(\mathbf{a}) = \mathbb{P}_{\mu_{\psi}\pi}(\mathbf{A} = \mathbf{a} \mid \mathbf{Z} = \mathbf{z}) \quad (3.2)$$

$$= \prod_{ij \in \mathcal{D}_n} \mathbb{P}_{\mu_{\psi}\pi}(A_{ij} = a_{ij} \mid \mathbf{Z} = \mathbf{z}). \quad (3.5)$$

The multiplicand in eq. (3.5) is the subject of definition 3.1.1's condition (e), which we rearrange as follows:

$$\mathbb{P}_{\mu_{\psi}\pi}(A_{ij} = a_{ij} \mid \mathbf{Z} = \mathbf{z}) = \psi_{z_{\uparrow j} z_{\uparrow i}}^{a_{ij}} \left(1 - \psi_{z_{\uparrow j} z_{\uparrow i}}\right)^{1-a_{ij}} \quad (3.6)$$

$$= \left(\frac{\psi_{z_{\uparrow j} z_{\uparrow i}}}{1 - \psi_{z_{\uparrow j} z_{\uparrow i}}}\right)^{a_{ij}} \left(1 - \psi_{z_{\uparrow j} z_{\uparrow i}}\right) \quad (3.7)$$

$$= \exp(\text{logit}(\psi_{z_{\uparrow j} z_{\uparrow i}})a_{ij} + \log(1 - \psi_{z_{\uparrow j} z_{\uparrow i}})). \quad (3.8)$$

The above equations avoid undefined terms by relying in eq. (3.6) on our convention that  $0^0 = 1$ ; in eq. (3.7) on  $\psi_{z_{\uparrow j} z_{\uparrow i}} \neq 1$ , per the definition of the set  $\Psi_k$  of edge probabilities; and in eq. (3.8) on our conventions that  $\text{logit} 0 = -\infty$  and  $e^{-\infty} = 0$ . We plug the latter equation back into eq. (3.5):

$$\begin{aligned} \gamma_{\psi}^{\mathbf{z}}(\mathbf{a}) &= \prod_{ij \in \mathcal{D}_n} \exp(\text{logit}(\psi_{z_{\uparrow j} z_{\uparrow i}})a_{ij} + \log(1 - \psi_{z_{\uparrow j} z_{\uparrow i}})) \\ &= \exp \sum_{ij \in \mathcal{D}_n} (\text{logit}(\psi_{z_{\uparrow j} z_{\uparrow i}})a_{ij} + \log(1 - \psi_{z_{\uparrow j} z_{\uparrow i}})). \end{aligned} \quad (3.9)$$

Lemma 3.1.6 permits us to break up the sum in eq. (3.9) as follows, so eq. (3.9) equals

$$\begin{aligned} &\exp \sum_{\sigma, \rho=1}^k \sum_{ij \in \mathcal{E}_{\sigma\rho}^{\mathbf{z}}} (\text{logit}(\psi_{z_{\uparrow j} z_{\uparrow i}})a_{ij} + \log(1 - \psi_{z_{\uparrow j} z_{\uparrow i}})) \\ &= \exp \sum_{\sigma, \rho=1}^k \left( \text{logit}(\psi_{\sigma\rho}) \sum_{ij \in \mathcal{E}_{\sigma\rho}^{\mathbf{z}}} a_{ij} + \log(1 - \psi_{\sigma\rho}) \sum_{ij \in \mathcal{E}_{\sigma\rho}^{\mathbf{z}}} 1 \right) \\ &= \exp \sum_{\sigma, \rho=1}^k \left( \text{logit}(\psi_{\sigma\rho}) M_{\sigma\rho}(\mathbf{a}, \mathbf{z}) + \log(1 - \psi_{\sigma\rho}) N_{\sigma\rho}(\mathbf{z}) \right). \end{aligned} \quad (3.10)$$

Equation (3.4) is nothing but eq. (3.10) with dot products denoting sum products and  $\mathbf{1}\mathbf{1}^{\top}$  denoting a matrix of all ones.

Equation (3.4) is an exponential-family representation of  $\mathcal{M}^z$  because  $M$  depends on  $\mathbf{a}$  but  $N$  does not. In particular,  $M(\mathbf{A}, z)$  is a sufficient statistic for  $\psi$  by Casella and Berger (2002, Thm. 6.2.10).  $\square$

The definition we gave above for *sufficient statistic* is that the PMF doesn't depend on the parameter when we hold the statistic constant. Lemma 3.1.8 says that  $M$  is a sufficient statistic in the conditional model. In section 3.2, we will use this ability of  $M$  to sideline the edge probabilities parameter  $\psi$ . Thus we will often need to hold  $M$  constant. This leads us to the following definition.

**Definition 3.1.9 (Fiber).** Fix a state  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . Network  $\mathbf{a}$ 's *fiber* conditional on the block assignments  $z$  is  $\mathbf{a}$ 's pre-image under  $M$  (cf. Petrović et al., 2010, Def. of  $\mathcal{T}_t$  on p. 265):  $\mathcal{F}(\mathbf{a}, z) := \{\mathbf{b} \in \mathcal{G}_n \mid M(\mathbf{b}, z) = M(\mathbf{a}, z)\}$ .

If the distribution of the random network  $\mathbf{A}$  does not depend on the edge probabilities  $\psi$  when we hold  $M$  constant, then what is the distribution? It is uniform according to the following lemma, which algorithm 3.2.1 uses to sample from  $\mathcal{F}$ .

**Lemma 3.1.10.** Fix a state  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . For any  $z$ -conditional MMSBM PMF  $\gamma \in \mathcal{M}^z$ , the distribution of the random network  $\mathbf{A}$  conditional on  $\mathbf{a}$ 's fiber, i.e.,  $M(\mathbf{A}, z) = M(\mathbf{a}, z)$ , is uniform:

$$\mathbb{P}_\gamma(\mathbf{A} = \mathbf{b} \mid \mathbf{A} \in \mathcal{F}(\mathbf{a}, z)) = \frac{1}{|\mathcal{F}(\mathbf{a}, z)|} \quad (3.11)$$

for all  $\mathbf{b} \in \mathcal{F}(\mathbf{a}, z)$ , where

$$|\mathcal{F}(\mathbf{a}, z)| = \prod_{\sigma=1}^k \prod_{\rho=1}^k \binom{N_{\sigma\rho}(z)}{M_{\sigma\rho}(\mathbf{a}, z)}. \quad (3.12)$$

*Proof.* That the conditional model  $\mathcal{M}^z$  has a uniform distribution over the fiber follows from the exponential-family representation eq. (3.4) and the discussion in Diaconis and Sturmfels (1998, p. 365). In particular,  $\gamma(\mathbf{b})$  depends on the network  $\mathbf{b}$  only through  $M(\mathbf{b}, z)$ , per eq. (3.4), so  $\gamma$  is constant on the fiber, which directly implies eq. (3.11).

It remains to prove eq. (3.12). By lemma 3.1.6 we may consider the block-assignment equivalence class  $\mathcal{E}_{\sigma\rho}^z$  of one sending/receiving block pair  $\sigma, \rho \in [k]$  at a time. The number of networks  $\mathbf{b} \in \mathcal{G}_n$  for which  $M_{\sigma\rho}(\mathbf{b}, z) = M_{\sigma\rho}(\mathbf{a}, z)$  is the number of ways we can choose

$M_{\sigma\rho}(\mathbf{a}, \mathbf{z})$  edges from  $\mathcal{E}_{\sigma\rho}^z$ . That number is the multiplicand in eq. (3.12). These choices are independent across equivalence classes, hence eq. (3.12).  $\square$

Proposition 3.1.11 offers a result analogous to lemma 3.1.8 but for the latent MMSBM.

**Proposition 3.1.11.** *Define the function  $\mathbf{J}: \mathcal{B}_{n,k} \rightarrow \mathbb{N}^{n \times k}$  to give the number of allowed edges that a node  $i \in [n]$  sends in each block  $\sigma \in [k]$ :*

$$J_{i\sigma}(\mathbf{z}_{\rightarrow}) := \sum_{\substack{j \in [n] \\ ij \in \mathcal{D}_n}} \mathbb{1}(z_{ij} = \sigma).$$

Then the latent MMSBM  $\mathcal{M}$  is an exponential family with the representation

$$\mu_{\psi\pi}(\mathbf{a}, \mathbf{z}) = \exp[\text{logit}(\psi) \cdot \mathbf{M}(\mathbf{a}, \mathbf{z}) + \log(\mathbf{11}^\top - \psi) \cdot \mathbf{N}(\mathbf{z}) + \log(\pi) \cdot (\mathbf{J}(\mathbf{z}_{\rightarrow}) + \mathbf{J}(\mathbf{z}_{\leftarrow}^\top))]$$

for all edge probabilities  $\psi \in \Psi_k$ , all block probabilities  $\pi \in \Pi_{n,k}$ , and all states  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ .

*Proof.* Fix  $\psi \in \Psi_k$ ,  $\pi \in \Pi_{n,k}$ , and  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ . Denote  $\mu := \mu_{\psi\pi}$ . We start by showing that the marginal distribution of  $\mathbf{Z}$  under  $\mathbb{P}_\mu$  has an exponential family representation. By definition 3.1.1's conditions (a) and (b), we have

$$\mathbb{P}_\mu(\mathbf{Z}_{\rightarrow} = \mathbf{z}_{\rightarrow}) = \prod_{ij \in \mathcal{D}_n} \mathbb{P}_\mu(Z_{ij} = z_{ij}) = \prod_{ij \in \mathcal{D}_n} \pi_{iz_{ij}}.$$

Rearranging this yields

$$\mathbb{P}_\mu(\mathbf{Z}_{\rightarrow} = \mathbf{z}_{\rightarrow}) = \prod_{i \in [n]} \prod_{\sigma=1}^k \pi_{i\sigma}^{J_{i\sigma}(\mathbf{z}_{\rightarrow})} = \exp(\log(\pi) \cdot \mathbf{J}(\mathbf{z}_{\rightarrow})).$$

Likewise,  $\mathbb{P}_\mu(\mathbf{Z}_{\leftarrow} = \mathbf{z}_{\leftarrow}) = \exp(\log(\pi) \cdot \mathbf{J}(\mathbf{z}_{\leftarrow}^\top))$  (via definition 3.1.1's condition (c)). The result follows from combining these distributions with eq. (3.4) via the law of total probability,

$$\mu(\mathbf{a}, \mathbf{z}) = \mathbb{P}_\mu(\mathbf{A} = \mathbf{a} \mid \mathbf{Z} = \mathbf{z}) \mathbb{P}_\mu(\mathbf{Z}_{\rightarrow} = \mathbf{z}_{\rightarrow}) \mathbb{P}_\mu(\mathbf{Z}_{\leftarrow} = \mathbf{z}_{\leftarrow}). \quad \square$$

### 3.2 Goodness of Fit

We now extend to conditional MMSBMs the work of Karwa et al. (2016) in hypothesis testing for the classical, non-mixed-membership *stochastic block models* (SBMs), in which  $\mathbf{Z}_{\rightarrow} = \mathbf{Z}_{\leftarrow}^\top$  and every row of  $\mathbf{Z}_{\rightarrow}$  is the same. Throughout this section, we assume the

following statistical scenario. Recall that  $\theta^*$  is the unknown, unknowable, true PMF of the random network and block assignments  $(\mathbf{A}, \mathbf{Z})$ .  $\theta^*$  is the distribution from which all observations of those random variables are drawn.

**Definition 3.2.1.** Assume that we have observed the event

$$\mathbf{A} = \mathbf{a}^*, \quad \mathbf{Z} = \mathbf{z}^* \quad (3.13)$$

for some *observed* network  $\mathbf{a}^*$  and block assignments  $\mathbf{z}^*$  such that  $(\mathbf{a}^*, \mathbf{z}^*)$  is a state in  $\mathcal{Y}_{n,k}$ .

We seek to test the *general null hypothesis*  $H_0$  that the random network's true distribution conditional on the observed block assignments is from the  $\mathbf{z}^*$ -conditional MMSBM  $\mathcal{M}$ , which is the *general null model*, against the alternative hypothesis  $H_1$  that it isn't:

$$H_0: \theta^{*\mathbf{z}^*} \in \mathcal{M}^{\mathbf{z}^*} \quad \text{against} \quad H_1: \theta^{*\mathbf{z}^*} \in \Theta^{\mathbf{z}^*} \setminus \mathcal{M}^{\mathbf{z}^*}. \quad (3.14)$$

Put differently, the general null hypothesis is

$$H_0: \text{there are } \mathbf{true\ edge\ probabilities\ } \psi^* \in \Psi_k \text{ such that } \theta^{*\mathbf{z}^*} = \gamma_{\psi^*}^{\mathbf{z}^*} \quad (3.15)$$

against the alternative hypothesis that no such  $\psi^*$  exists (see Bishop et al., 2007, § 14.7.1 esp. pp. 502–503). In particular, by conditioning on  $\mathbf{Z} = \mathbf{z}^*$ , we obviate the need to consider the existence of block probabilities  $\pi \in \Pi_{n,k}$ .

To break up the problem in definition 3.2.1 into smaller pieces, we also define the following *simple* hypothesis asserting the truth of a single PMF (Casella & Berger, 2002, p. 388 in § 8.3.2).

**Definition 3.2.2.** Fix some edge probabilities  $\psi \in \Psi_k$ . The *simple null hypothesis*  $H_{0\psi}$  asserts that the random network's true distribution conditional on the observed block assignments  $\mathbf{z}^*$  is the  $\mathbf{z}^*$ -conditional MMSBM PMF  $\gamma_{\psi}^{\mathbf{z}^*}$ . The alternative hypothesis  $H_{1\psi}$  is that the true conditional distribution could be any other PMF. In other words, the test is of

$$H_{0\psi}: \theta^{*\mathbf{z}^*} = \gamma_{\psi}^{\mathbf{z}^*} \quad \text{against} \quad H_{1\psi}: \theta^{*\mathbf{z}^*} \in \Theta^{\mathbf{z}^*} \setminus \{\gamma_{\psi}^{\mathbf{z}^*}\}. \quad (3.16)$$

Because we have only one observed network, our test cannot use asymptotic methods: the numbers of observed networks (one), nodes ( $n$ ), and blocks ( $k$ ) are all fixed.

Rather we will employ an *exact test*, which approximates the p-value without relying on asymptotic arguments. This is in contrast to, for example, Lei (2016), which presented a goodness-of-fit test valid as  $n \rightarrow \infty$ . Lei tested the null hypothesis that the number  $k$  of blocks equals some given number under an SBM for simple, undirected networks. Proving the validity of the test relied on the convergence in distribution of a test statistic as  $n \rightarrow \infty$ . Having observed only  $\mathbf{A} = \mathbf{a}^*$ , we cannot profit from allowing  $n \rightarrow \infty$ .

Another asymptotic analysis of SBMs in goodness-of-fit testing came from Latouche et al. (2018). The null hypothesis was that each edge indicator  $A_{ij}$  is distributed according to a logistic regression model with a residual term that in the absence of any covariates would make the model an SBM with  $k = 1$  block. The alternative hypothesis was that the residual term has  $k > 1$  blocks. The Bayesian goodness-of-fit test relied on  $k \rightarrow \infty$ . Our observational scenario assumes a fixed number of blocks. See Haberman (1981) for more on the challenges of asymptotics in statistical models of networks.

In addition to choosing an exact test, the other facet of our strategy is to use a *conditional test*: In subsection 3.2.2 we will construct a p-value as a test statistic's tail probability *conditional* on the observed sufficient statistic  $M(\mathbf{a}^*, \mathbf{z}^*)$ 's fiber  $\mathcal{F}(\mathbf{a}^*, \mathbf{z}^*)$ . Subsection 3.2.1 will define the particular test statistic. Conditioning on a sufficient statistic is a standard technique when the statistic is discrete (see, e.g., Casella & Berger, 2002, p. 399). The chief advantage of conditioning is that it reduces the p-value's dependence on parameters of the null model by removing the probability's dependence on the parameters (*cf.* Casella & Berger, 2002, Thm. 8.3.27 on p. 397). The edge probabilities still appear in the test statistic even after conditioning. In the case of our simple null hypothesis, that's no impediment: computing the p-value comes down to approximating the number of networks in a subset of the fiber, which subsection 3.2.3 does with a Monte Carlo (MC) algorithm.

However we have no algorithm for the p-value that subsection 3.2.2 develops for the general null hypothesis. The edge probabilities appearance in the test statistic is the impediment because the general-hypothesis p-value is the maximum over the simple-hypothesis p-values, and we don't know how to solve the optimization problem. Oddly the statistics literature appears to be split on the necessity of the optimization. On balance it seems



to us to be necessary, so we view the optimization as an open problem. Subsection 3.2.4 discusses the two schools of thought in the literature and details the options for avoiding the optimization problem.

**3.2.1 Test Statistic.** In this subsection definition 3.2.5 on page 106 introduces a new test statistic, which we call *discrepancy*, and we present some results to demonstrate its suitability for conditional goodness-of-fit testing of the general null hypothesis in definition 3.2.1.

In view of remark 3.1.4, we draw inspiration for our test statistic from goodness-of-fit testing for multinomial models. In that context, “[a] goodness-of-fit statistic is a measure of the ‘distance’ from” (Bishop et al., 2007, p. 507) the null model to the unconstrained MLE for the multinomial cell probabilities, which is the observed cell proportions (Bishop et al., 2007, p. 503). A *distance function*, which defines the goodness-of-fit statistic, is a nonnegative function on pairs of cell-probability vectors that maps pairs to zero if and only if they’re equal, and is at least vaguely, eventually monotonic in the Euclidean distance (Bishop et al., 2007, p. 504). Two of the most popular test statistics for discrete data in general and multinomial data in particular are twice the log-likelihood ratio, called the *likelihood statistic*  $G^2$ , and its second-order Taylor-series approximation, Pearson’s  $\hat{\chi}^2$  (Bishop et al., 2007, pp. 124–126, and pp. 513–514 regarding the relationship between  $G^2$  and  $\hat{\chi}^2$ ; for further evidence of their popularity, see Petrović et al., 2010, pp. 264–265; see also Diaconis & Sturmfels, 1998). Both are based on distance functions (Bishop et al., 2007, pp. 507–508), and they have the same asymptotic distribution (Bishop et al., 2007, pp. 513–514). Either provides a reasonable starting point for testing a graphical model such as a conditional MMSBM. Our discrepancy statistic is similar to Pearson’s  $\hat{\chi}^2$  statistic, which, in our context, is

$$\hat{\chi}^2(\mathbf{a}, \mathbf{z}) := \sum_{ij \in \mathcal{D}_n} \frac{(a_{ij} - \hat{\psi}_{z_{\overline{ij}} z_{\overline{ij}}}(\mathbf{a}, \mathbf{z}))^2}{\hat{\psi}_{z_{\overline{ij}} z_{\overline{ij}}}(\mathbf{a}, \mathbf{z})}, \quad (3.17)$$

for any state  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ , where  $\hat{\psi}(\mathbf{a}, \mathbf{z}) \in \Psi_k$  is the maximum likelihood estimator (MLE) of the edge probabilities  $\psi$  under the general null hypothesis in definition 3.2.1 (Bishop et al., 2007, pp. 57–58).

Actually  $\hat{\psi}$  could be another estimator of the edge probabilities, but the MLE is the

most popular (Petrović et al., 2010, p. 264), and it also minimizes goodness-of-fit statistics that are based on a distance function—at least in multinomial models that replace  $a_{ij}$  in eq. (3.17) with observed cell proportions (Bishop et al., 2007, p. 505). We return to this issue in subsection 3.2.4. For now the next lemma shows that computing the MLE is straightforward under definition 3.2.1’s assumption that we have observed the block assignments.

**Lemma 3.2.3.** *Fix block assignments  $z$  for which there is a network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$  and for which  $N(z)$  is non-zero in every entry. The MLE of the edge probabilities  $\psi$  in the  $z$ -conditional MMSBM is the function  $\hat{\psi}: \mathcal{Y}_{n,k} \rightarrow [0, 1]^{k \times k}$  such that, for each block pair  $\sigma, \rho \in [k]$  and each network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ ,*

$$\hat{\psi}_{\sigma\rho}(\mathbf{a}, z) := \frac{M_{\sigma\rho}(\mathbf{a}, z)}{N_{\sigma\rho}(z)}. \quad (3.18)$$

*Proof.* Fix a network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$  and a block pair  $\sigma, \rho \in [k]$ . Lemma 3.1.8 says that  $\mathcal{M}^z$  is an exponential family with sufficient statistic  $\mathbf{M}$ . Thus, by E. L. Lehmann and Casella (1998, p. 470), the MLE of edge probabilities given the observation  $\mathbf{A} = \mathbf{a}$  from a PMF in  $\mathcal{M}^z$  is the unique  $\hat{\psi} \in [0, 1]^{k \times k}$  that satisfies  $M_{\sigma\rho}(\mathbf{a}, z) = \mathbb{E}_{\hat{\psi}}[M_{\sigma\rho}(\mathbf{A}, z)]$ . Expanding the right side using the cardinality formulas in lemma 3.1.6 shows that  $M_{\sigma\rho}(\mathbf{a}, z)$  equals

$$\mathbb{E}_{\hat{\psi}}[M_{\sigma\rho}(\mathbf{A}, z)] = \mathbb{E}_{\hat{\psi}}\left[\sum_{ij \in \mathcal{E}_{\sigma\rho}^z} A_{ij}\right] = \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} \mathbb{E}_{\hat{\psi}}[A_{ij}] = \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} \hat{\psi}_{\sigma\rho} = \hat{\psi}_{\sigma\rho} \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} 1 = \hat{\psi}_{\sigma\rho} N_{\sigma\rho}(z).$$

As long as  $N_{\sigma\rho}(z) \neq 0$ , we may divide through to obtain eq. (3.18).  $\square$

Lemma 3.2.3 requires that  $N(z)$  contain no zeros for the MLE to exist. If the true PMF  $\theta^*$  lies in the latent MMSBM  $\mathcal{M}$ , the no-zeros condition necessitates that the block probabilities  $\pi$  contain no zeros, too. We could have stipulated such a constraint in the definition of  $\Pi_{n,k}$  (just as we stipulated that edge probabilities  $\psi \in \Psi_k$  contain no ones) (cf. the assumption for MLE that the parameter space contain an open set of which the true parameter is an interior point: E. L. Lehmann & Casella, 1998, Assumption (A3) on p. 444). Instead we chose to emphasize the role of observing no zeros in  $N$ . At the same time, it is entirely possible for eq. (3.18) to yield an estimate of zero or one even if the truth it is estimating is strictly between zero and one. For more on what happens to exponential

families' MLEs when zeros appear in their contingency tables, see Fienberg and Rinaldo (2012b) in conjunction with remarks 3.1.4 and 3.1.7.

Estimators other than maximum likelihood have been used in the literature outside the context of goodness of fit. Under the assumption that only the network but not the block assignments have been observed, Airolodi et al. (2008, pp. 1988–1992) gave an expectation-maximization algorithm for MLE and a Bayesian, posterior-inference algorithm under a Dirichlet prior for the rows of  $\pi$ .

The ease of MLE aside,  $\hat{\chi}^2$  is not a suitable statistic for a conditional test using the sufficient statistic  $M$  because we cannot use  $\hat{\chi}^2$  to distinguish distributions of observed graphs that are in the same fiber of the sufficient statistic. Proposition 3.2.4 makes this assertion more precise.

**Proposition 3.2.4.** *Fix some state  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ . Pearson's  $\hat{\chi}^2$  is constant on  $M$ 's fiber, i.e.,  $\hat{\chi}^2(\mathbf{b}) = \hat{\chi}^2(\mathbf{a})$  for all  $\mathbf{b} \in \mathcal{F}(\mathbf{a}, \mathbf{z})$ , whenever  $\hat{\psi}$  is. When  $\hat{\psi}$  is the MLE in eq. (3.18),  $\hat{\chi}^2(\mathbf{a}) = n$ .*

*Proof.* Using lemma 3.1.6, we break up the sum in eq. (3.17) by equivalence class  $\mathcal{E}_{\sigma\rho}^z$  for each block pair  $\sigma, \rho \in [k]$ . We also expand the square in the numerator. (Let's assume that  $\hat{\psi}(\mathbf{a}, \mathbf{z})$  is non-zero in every entry to avoid trivialities.)

$$\begin{aligned} \hat{\chi}^2(\mathbf{a}, \mathbf{z}) &= \sum_{\sigma, \rho} \left[ \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} \frac{a_{ij}^2 - a_{ij} \hat{\psi}_{z_{\bar{\sigma}} z_{\bar{\rho}}}(\mathbf{a}, \mathbf{z}) + \hat{\psi}_{z_{\bar{\sigma}} z_{\bar{\rho}}}^2(\mathbf{a}, \mathbf{z})}{\hat{\psi}_{z_{\bar{\sigma}} z_{\bar{\rho}}}(\mathbf{a}, \mathbf{z})} \right] \\ &= \sum_{\sigma, \rho} \frac{1}{\hat{\psi}_{\sigma\rho}(\mathbf{a}, \mathbf{z})} \left[ \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} (a_{ij}^2 - a_{ij} \hat{\psi}_{\sigma\rho}(\mathbf{a}, \mathbf{z})) + \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} \hat{\psi}_{\sigma\rho}^2(\mathbf{a}, \mathbf{z}) \right] \end{aligned}$$

$a_{ij} \in \{0, 1\}$ , so  $a_{ij} = a_{ij}^2$ :

$$= \sum_{\sigma, \rho} \frac{1}{\hat{\psi}_{\sigma\rho}(\mathbf{a}, \mathbf{z})} \left[ (1 - \hat{\psi}_{\sigma\rho}(\mathbf{a}, \mathbf{z})) \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} a_{ij} + \hat{\psi}_{\sigma\rho}^2(\mathbf{a}, \mathbf{z}) \sum_{ij \in \mathcal{E}_{\sigma\rho}^z} 1 \right]$$

Using the cardinality formulas in lemma 3.1.6:

$$= \sum_{\sigma, \rho} \frac{1}{\hat{\psi}_{\sigma\rho}(\mathbf{a}, \mathbf{z})} \left[ (1 - \hat{\psi}_{\sigma\rho}(\mathbf{a}, \mathbf{z})) M_{\sigma\rho}(\mathbf{a}, \mathbf{z}) + \hat{\psi}_{\sigma\rho}^2(\mathbf{a}, \mathbf{z}) N_{\sigma\rho}(\mathbf{z}) \right].$$

Plugging eq. (3.18) into that last expression simplifies it to  $\sum_{\sigma, \rho} N_{\sigma\rho}(\mathbf{z}) = n$ .  $\square$

Instead of  $\chi^2$ , we use the following definition.

**Definition 3.2.5.** Define  $L: \mathcal{Y}_{n,k} \rightarrow \mathbb{N}^{n \times k \times k}$  to give the number of edges along which a node  $i$  sends or receives in a block  $\sigma$  and  $i$ 's neighbor receives or sends in a block  $\rho$ :

$$L_{i\sigma\rho}(\mathbf{a}, \mathbf{z}) := \sum_{\substack{j=1 \\ ij \in \mathcal{D}_n}}^n \left[ \mathbb{1}(z_{\overrightarrow{ij}} = \sigma) \mathbb{1}(z_{\overleftarrow{ij}} = \rho) a_{ij} \right] + \sum_{\substack{j=1 \\ ji \in \mathcal{D}_n, j \neq i}}^n \left[ \mathbb{1}(z_{\overrightarrow{ji}} = \rho) \mathbb{1}(z_{\overleftarrow{ji}} = \sigma) a_{ji} \right]. \quad (3.19)$$

Fix a state  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ . Let  $\theta$  be a PMF either on  $\mathcal{Y}_{n,k}$  ( $\theta \in \Theta$ ) or on the set  $\mathcal{G}_n$  of networks and conditional on  $\mathbf{Z} = \mathbf{z}$  ( $\theta \in \Theta^{\mathbf{z}}$ ). Define the *discrepancy*  $W_\theta: \mathcal{Y}_{n,k} \rightarrow \mathbb{R}$  as

$$W_\theta(\mathbf{a}, \mathbf{z}) := \sum_{i=1}^n \sum_{\sigma=1}^k \sum_{\rho=1}^k \frac{(L_{i\sigma\rho}(\mathbf{a}, \mathbf{z}) - \mathbb{E}_\theta[L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})])^2}{\text{Var}_\theta[L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})]}. \quad (3.20)$$

$\text{Var}_\theta[L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})] \neq 0$

Where context implies  $\psi$  is a matrix of edge probabilities, we write  $W_\psi$  in place of  $W_{\gamma, \mathbf{z}}$ .

The discrepancy gives the a measure of the relative difference between  $L(\mathbf{a}, \mathbf{z})$  and the expected value of  $L(\mathbf{A}, \mathbf{Z})$  under the distribution  $\theta$ . Larger values of  $W_\theta(\mathbf{a}, \mathbf{z})$  indicate that the state  $(\mathbf{a}, \mathbf{z})$  is less typical of samples from  $\theta$  and thus testify against the state's being a sample from  $\theta$ .

Under a conditional MMSBM,  $L$ 's distribution, expected value, and variance are straightforward to compute. Proposition 3.2.6 on the next page relies on how eq. (3.19) breaks up the sum to ensure all the summands are mutually independent. To help the proposition describe that, let  $\mathcal{S}_{i\sigma\rho}^{\mathbf{z}}$  be the set of allowed edges that node  $i$  sends in block  $\sigma$  to a receiver in block  $\rho$ , and let  $\mathcal{R}_{i\rho\sigma}^{\mathbf{z}}$  be the set of allowed edges that node  $i$  receives in block  $\sigma$  from a sender in block  $\rho$  (excluding self-loops even if  $\mathcal{G}_n$  is non-simple). Equation (3.21) formalizes these definitions using lemma 3.1.6:

$$\mathcal{S}_{i\sigma\rho}^{\mathbf{z}} := \left\{ uv \in \mathcal{E}_{\sigma\rho}^{\mathbf{z}} \mid u = i \right\}, \quad \mathcal{R}_{i\rho\sigma}^{\mathbf{z}} := \left\{ uv \in \mathcal{E}_{\rho\sigma}^{\mathbf{z}} \mid u \neq v = i \right\}, \quad (3.21)$$

which implies

$$\left| \mathcal{S}_{i\sigma\rho}^{\mathbf{z}} \right| = \sum_{\substack{j=1 \\ ij \in \mathcal{D}_n}}^n \mathbb{1}(z_{\overrightarrow{ij}} = \sigma) \mathbb{1}(z_{\overleftarrow{ij}} = \rho), \quad \left| \mathcal{R}_{i\rho\sigma}^{\mathbf{z}} \right| = \sum_{\substack{j=1 \\ ji \in \mathcal{D}_n, j \neq i}}^n \mathbb{1}(z_{\overrightarrow{ji}} = \rho) \mathbb{1}(z_{\overleftarrow{ji}} = \sigma), \quad (3.22)$$

for all nodes  $i \in [n]$ , block pairs  $\sigma, \rho \in [k]$ , and block assignments  $\mathbf{z}$  for which there is some network  $\mathbf{a}$  such that  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ . As is perhaps easier to see from eq. (3.22), eq. (3.21)

defines sets that correspond to summands in eq. (3.19):

$$L_{i\sigma\rho}(\mathbf{a}, \mathbf{z}) = \sum_{ij \in \mathcal{S}_{i\sigma\rho}^z} a_{ij} + \sum_{ji \in \mathcal{R}_{i\rho\sigma}^z} a_{ji}.$$

**Proposition 3.2.6.** *Fix block assignments  $\mathbf{z}$  for which there is some  $\mathbf{a}$  such that  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ . Fix a node  $i \in [n]$  and a block pair  $\sigma, \rho \in [k]$ . Under a  $\mathbf{z}$ -conditional MMSBM PMF with some edge probabilities  $\psi \in \Psi_k$ ,  $L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})$  has Poisson's binomial distribution (Poisson's binomial distribution is that of the sum of a finite number of independent Bernoulli trials not necessarily with the same success probabilities. For a thorough survey, see Wang, 1993), the sum of two, independent, binomial, random variables: one with success probability  $\psi_{\sigma\rho}$  and  $|\mathcal{S}_{i\sigma\rho}^z|$  trials; and the other with success probability  $\psi_{\rho\sigma}$  and  $|\mathcal{R}_{i\rho\sigma}^z|$  trials. The expected value and variance are*

$$\mathbb{E}_\psi(L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})) = \psi_{\sigma\rho} |\mathcal{S}_{i\sigma\rho}^z| + \psi_{\rho\sigma} |\mathcal{R}_{i\rho\sigma}^z|, \quad (3.23)$$

$$\text{Var}_\psi(L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})) = \psi_{\sigma\rho}(1 - \psi_{\sigma\rho}) |\mathcal{S}_{i\sigma\rho}^z| + \psi_{\rho\sigma}(1 - \psi_{\rho\sigma}) |\mathcal{R}_{i\rho\sigma}^z|. \quad (3.24)$$

*Proof.* By eq. (3.2), the only events with positive probability are those with  $\mathbf{Z} = \mathbf{z}$ . That event conditions everything that follows. For brevity, let  $\mathcal{S} := \mathcal{S}_{i\sigma\rho}^z$  and  $\mathcal{R} := \mathcal{R}_{i\rho\sigma}^z$ .

First we prove the independence of the summands in eq. (3.19) when  $\mathbf{a} = \mathbf{A}$ . In particular, the summands are the  $A_{uv}$ s for  $uv \in \mathcal{S} \cup \mathcal{R}$ . Therefore our task is to show that there's no double counting, i.e., that  $\mathcal{S} \cap \mathcal{R} = \emptyset$ . For  $j \in [n]$ ,  $\mathcal{S}$  contains allowed edges of the form  $ij$  and  $\mathcal{R}$  of the form  $ji$ . Recall from subsection 3.1.2 that  $ij$  is an abbreviation for the ordered pair  $(i, j)$ . If  $ij \in \mathcal{S} \cap \mathcal{R}$ , then  $ij = ji$ , so  $j = i$ , i.e.,  $ij$  is the self-loop  $ii$ . (This holds even if  $\mathcal{G}_n$  is undirected; see table 3.1.) But eq. (3.21) defines  $\mathcal{R}$  to exclude self-loops in case  $\mathcal{G}_n$  is non-simple. Therefore  $\mathcal{S} \cap \mathcal{R} = \emptyset$ .

By definition 3.1.3 and definition 3.1.1's conditions (d) and (e), each summand  $A_{uv}$  in eq. (3.19) when  $\mathbf{a} = \mathbf{A}$  is an independent Bernoulli random variable with success probability  $\psi_{z_{\overline{u}\sigma} z_{\overline{v}\rho}}$ . If  $uv \in \mathcal{S}$ , then  $\psi_{z_{\overline{u}\sigma} z_{\overline{v}\rho}} = \psi_{\sigma\rho}$ ; if  $uv \in \mathcal{R}$ , then  $\psi_{z_{\overline{u}\sigma} z_{\overline{v}\rho}} = \psi_{\rho\sigma}$ . Hence  $L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})$  is the sum of  $|\mathcal{S}|$  independently and identically distributed (IID) Bernoulli trials with success probability  $\psi_{\sigma\rho}$ , plus  $|\mathcal{R}|$  IID Bernoulli trials with success probability  $\psi_{\rho\sigma}$ , and the two sets of trials are mutually independent. Therefore  $L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})$  has Poisson's binomial distribution

with PMF (cf. Wang, 1993, Eq. (7) on p. 298)  $f_{\psi}^z: \mathbb{N} \rightarrow [0, 1]$  such that

$$f_{\psi}^z(x) = \sum_{\ell=0}^x \binom{|\mathcal{S}|}{\ell} \psi_{\sigma\rho}^{\ell} (1 - \psi_{\sigma\rho})^{|\mathcal{S}|-\ell} \binom{|\mathcal{R}|}{x-\ell} \psi_{\rho\sigma}^{x-\ell} (1 - \psi_{\rho\sigma})^{|\mathcal{R}|-(x-\ell)}. \quad (3.25)$$

(For a proof of this formula, think of  $\ell$  as choosing the number of the  $x$  Bernoulli-trial successes coming from  $\mathcal{S}$ , so that the other  $x - \ell$  successes come from  $\mathcal{R}$ .)

The expected-value and variance formulas in eqs. (3.23) and (3.24) follow from the linearity of  $\mathbb{E}$  and  $\text{Var}$  over independent random variables (Jacod & Protter, 2004, Thms. 9.1(a) and 15.4 on pp. 52, 119; cf. the expected-value and variance formulas for a general Poisson's binomial distribution in Wang, 1993, Eq. (15) on p. 301).  $\square$

It would be nice to sample  $L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})$  directly from eq. (3.25) independently for each  $i\sigma\rho$  index, rather than sampling  $\mathbf{A}$ . Under the proposition's hypotheses,  $\{L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})\}_{\sigma,\rho=1}^k$  is a mutually independent set for each node  $i$  because of lemma 3.1.6. However,  $L_{i\sigma\rho}(\mathbf{A}, \mathbf{Z})$  and  $L_{j\sigma\rho}(\mathbf{A}, \mathbf{Z})$  are mutually dependent because

$$ij \in \mathcal{S}_{i\sigma\rho}^z \text{ and } i \neq j \quad \text{if and only if} \quad ij \in \mathcal{R}_{j\sigma\rho}^z.$$

We conclude this subsection by returning to our discussion of Pearson's  $\chi^2$  above. There we mentioned that MLE and distance-minimization coincide in multinomial models; remark 3.1.4 showed how to transform a  $z$ -conditional MMSBM into a multinomial model. With the invariance property of MLEs in mind (Casella & Berger, 2002, Thm. 7.2.10 on p. 320), we offer the following conjecture.

**Conjecture 3.2.7.** *Fix a state  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . The MLE  $\hat{\psi}(\mathbf{a}, z) =: \psi$  from lemma 3.2.3 minimizes the discrepancy, i.e.,  $W_{\psi}(\mathbf{a}, z) = \inf_{\phi \in \Psi_k} W_{\phi}(\mathbf{a}, z)$ .*

**3.2.2 p-Value Definitions.** In this subsection definition 3.2.9 uses the discrepancy  $W$  from definition 3.2.5 to construct a p-value for an exact, conditional test of the general null hypothesis in definition 3.2.1. Karwa et al. (2016, § 3.2) was our starting point for the definitions, which we build up from a p-value for a the simple null hypothesis in definition 3.2.2.

A statistic  $p: \mathcal{Y}_{n,k} \rightarrow [0, 1]$  is a *p-value* for testing a null hypothesis  $H'_0$  against an alternative hypothesis  $H'_1$  if  $p$ 's being small is evidence against  $H'_0$  in favor of  $H'_1$ .  $p$  is a *valid*

p-value if additionally

$$\mathbb{P}_\theta(p(\mathbf{A}, \mathbf{Z}) \leq \alpha) \leq \alpha \quad \text{under any PMF } \theta \text{ in the null model} \quad (3.26)$$

and for any given *significance level*  $\alpha \in [0, 1]$  (Casella & Berger, 2002, Def. 8.3.26 on p. 397). The resulting test that rejects  $H'_0$  if  $p(\mathbf{A}, \mathbf{Z}) \leq \alpha$  is a *level- $\alpha$  test* because eq. (3.26) means that the probability when  $H'_0$  is true of rejecting  $H'_0$ , a *type I error*, is at most  $\alpha$  (Casella & Berger, 2002, pp. 382–383, 385, 397). The situation is worse for *invalid* p-values. If we compare one to  $\alpha$  to test  $H'_0$ , we risk a type I error with some probability that we know nothing about, but it might exceed  $\alpha$ . This is especially problematic because many statisticians treat rejecting a null hypothesis as a hypothesis test's only epistemically valid inference (Wooldridge, 1999, August/2012, p. 135; Casella & Berger, 2002, pp. 373–374).

A tool we will use to derive results about our p-value for the general null hypothesis is a function that gives the p-value except for choosing how to compute the expected-value and variance terms in eq. (3.20). Since we are focusing on the hypotheses in definitions 3.2.1 and 3.2.2, in which we assume we have observed the block assignments, we will restrict our attention here to computing those expected-value and variance terms under a conditional MMSBM PMF as in proposition 3.2.6. Fix block assignments  $z$  for which there is some network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . In particular we will focus on the general null model  $\mathcal{M}^z$ . For each matrix of edge probabilities  $\psi \in \Psi_k$ , define the *simple-hypothesis p-value*  $p_\psi^z: \mathcal{G}_n \rightarrow [0, 1]$ , for each network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ , by

$$p_\psi^z(\mathbf{a}) := \mathbb{P}_\psi(W_\psi(\mathbf{A}, z) \geq W_\psi(\mathbf{a}, z) \mid \mathbf{A} \in \mathcal{F}(\mathbf{a}, z)). \quad (3.27)$$

To be explicit, the events upon which the probability distribution in eq. (3.27) is conditioned are the block assignments and the fiber  $\mathcal{F}$ :

$$\mathbf{Z} = z, \quad \mathbf{M}(\mathbf{A}, \mathbf{Z}) = \mathbf{M}(\mathbf{a}, z). \quad (3.28)$$

$\mathbf{M}$ 's sufficiency in the general null model means that  $\mathbb{P}_\psi(\cdot \mid \mathbf{A} \in \mathcal{F}(\mathbf{a}, z))$  does not depend on  $\psi$ . From lemma 3.1.10, we have (cf. Petrović et al., 2010, Eq. (2.4) on p. 265)

$$p_\psi^z(\mathbf{a}) = \frac{|\{\mathbf{b} \in \mathcal{F}(\mathbf{a}, z) \mid W_\psi(\mathbf{b}, z) \geq W_\psi(\mathbf{a}, z)\}|}{|\mathcal{F}(\mathbf{a}, z)|}. \quad (3.29)$$

$\psi$  still appears in eq. (3.29) because it determines how to compute the expected-value and variance terms in eq. (3.20) via eqs. (3.23) and (3.24).

**Lemma 3.2.8.** *For edge probabilities  $\psi$ , the simple-hypothesis  $p$ -value  $p_{\psi}^{z^*}$  is a valid  $p$ -value for the simple null hypothesis  $H_{0\psi}$  from definition 3.2.2.*

*Proof.* As discussed after definition 3.2.5, a large value of  $W_{\psi}(\mathbf{a}^*, z^*)$  is evidence against the observed network's being a sample from  $\gamma_{\psi}^{z^*}$ . Since eq. (3.27) conditions on the fiber of a sufficient statistic, the result then follows from Casella and Berger (2002, p. 399).  $\square$

Next our main definition constructs the test of the hypotheses in definition 3.2.1 using the *intersection-union method* based on the simple null hypothesis  $H_{0\psi}$  of eq. (3.16) (Casella & Berger, 2002, § 8.2.3 on pp. 380–381).

**Definition 3.2.9.** Fix block assignments  $z$  for which there is some network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . We define the *general-hypothesis  $p$ -value*  $p^z : \mathcal{G}_n \rightarrow [0, 1]$  as

$$p^z(\mathbf{a}) := \sup_{\psi \in \Psi_k} p_{\psi}^z(\mathbf{a}) \quad (3.30)$$

for each network  $\mathbf{a}$  such that  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ . The *conditional test* rejects definition 3.2.1's  $H_0$  in favor of its  $H_1$  at a given significance level  $\alpha \in [0, 1]$  when

$$p^z(\mathbf{a}^*) \leq \alpha. \quad (3.31)$$

Equation (3.29) shows that eq. (3.30) is a supremum over a finite set, so the “sup” is a “max”. Moreover, the denominator of eq. (3.29) doesn't depend on  $\psi$ . Hence

$$p^z(\mathbf{a}) = \frac{1}{|\mathcal{F}(\mathbf{a}, z)|} \max_{\psi \in \Psi_k} |\{\mathbf{b} \in \mathcal{F}(\mathbf{a}, z) \mid W_{\psi}(\mathbf{b}, z) \geq W_{\psi}(\mathbf{a}, z)\}|. \quad (3.32)$$

**Proposition 3.2.10.** *For the hypotheses in definition 3.2.1,  $p^{z^*}$  is a valid  $p$ -value.*

*Proof.* First we show that  $p^{z^*}$  is a  $p$ -value for the hypotheses in definition 3.2.1. For any edge probabilities  $\psi \in \Psi_k$ ,  $p_{\psi}^{z^*}(\mathbf{a}) \leq p^{z^*}(\mathbf{a})$ . Suppose the latter value is small, so that the former is as well. That is evidence, according to lemma 3.2.8, against the simple null hypothesis  $H_{0\psi}$  of eq. (3.16). The general null model is  $\mathcal{M}^{z^*} = \bigcup_{\psi \in \Psi_k} \{\gamma_{\psi}^{z^*}\}$ , and each of the singleton sets in



that union is the null model of the corresponding simple null hypothesis  $H_{0\psi}$ . Therefore the small value of  $p^{z^*}(\mathbf{a})$  is evidence against the general null hypothesis  $H_0$ .

Second, to prove validity, let  $\alpha \in [0, 1]$  and observe that the rejection region that eq. (3.31) defines is the intersection of the rejection regions for the corresponding test of each  $H_{0\gamma}$ . In detail, if  $\psi \in \Psi_k$ , then

$$\mathbb{P}_\psi(p^{z^*}(\mathbf{A}) \leq \alpha) = \mathbb{P}_\psi\left(\sup_{\phi \in \Psi_k} p_\phi^{z^*}(\mathbf{A}) \leq \alpha\right) = \mathbb{P}_\psi\left(\bigcap_{\phi \in \Psi_k} \{p_\phi^{z^*}(\mathbf{A}) \leq \alpha\}\right) \leq \mathbb{P}_\psi(p_\psi^{z^*}(\mathbf{A}) \leq \alpha).$$

Since  $p_\psi^{z^*}$  is valid by lemma 3.2.8,  $\mathbb{P}_\psi(p_\psi^{z^*}(\mathbf{A}) \leq \alpha) \leq \alpha$ . Therefore  $\mathbb{P}_\psi(p^{z^*}(\mathbf{A}) \leq \alpha) \leq \alpha$ .  $\square$

**3.2.3 Estimation.** This subsection develops an algorithm for computing the simple-hypothesis p-value  $p_\psi^z(\mathbf{a})$ . Without a closed-form expression for it, we must approximate it by mc. We construct the algorithm in two steps. Sub-subsection 3.2.3.1 presents algorithm 3.2.1 for sampling from the fiber of the observed network. Sub-subsection 3.2.3.2 presents algorithm 3.2.2 for estimating the distribution of the discrepancy among networks in the fiber, calling algorithm 3.2.1 as a subroutine. Finally sub-subsection 3.2.3.3 goes into detail analyzing the running time of algorithm 3.2.2. As we have no algorithm to perform the optimization in eq. (3.30), this subsection does not discuss the general-hypothesis p-value.

**3.2.3.1 Fiber Sampling.** A uniformly random sample from definition 3.1.9's fiber  $\mathcal{F}$  of definition 3.1.5's statistic  $M$  is a necessary input to any mc algorithm that estimates p-values conditioned on eq. (3.28). The samples' target distribution is uniform because of eq. (3.11).

For this purpose we introduce the *direct fiber sampler* in algorithm 3.2.1 on the following page. The proof of the algorithm's correctness is precisely the same as lemma 3.1.10's proof of eq. (3.12). *Direct* contrasts with the Markov chain mc (MCMC) algorithms (for an introduction to MCMC, see Givens & Hoeting, 2012, Ch. 7 on pp. 201–235), such as the one Diaconis and Sturmfels (1998, Lem. 2.1) laid out, customary in the literature on goodness-of-fit for network models—e.g., Karwa et al. (2016, Thm. 4.3), Li et al. (2016), and Ogawa et al. (2013). Algorithm 3.2.1 is an *exact simulation*: the distribution of the sample exactly equals the target distribution (Givens & Hoeting, 2012, p. 153). Direct samplers beat MCMC in terms of speed: no waiting for the Markov chain to converge; accuracy: exact rather than

approximate simulation; and ease of implementation: no parameters to tune and many fewer steps per sample.

The key step is line 3.2.1.5's choosing  $M_{\sigma\rho}(\mathbf{a}, \mathbf{z})$  edges without replacement from the block-assignment equivalence class  $\mathcal{E}_{\sigma\rho}^z$  that lemma 3.1.6 defined. An implementation could use Python's `random.sample` or NumPy's `numpy.random.Generator.choice` functions, whose running times would be proportional to  $M_{\sigma\rho}(\mathbf{a}, \mathbf{z})$ . Line 3.2.1.5 runs once for each block pair  $\sigma, \rho$  and thus the algorithm spends time on the line proportional to the number  $\sum_{\sigma\rho} M_{\sigma\rho}(\mathbf{a}, \mathbf{z})$  of edges in the network  $\mathbf{a}$ . However, computing  $M(\mathbf{a}, \mathbf{z})$  and  $\{\mathcal{E}_{\sigma\rho}^z\}_{\sigma\rho}$  themselves requires inspecting each of the  $|\mathcal{D}_n| = O(n^2)$  entries of all three matrices. Hence algorithm 3.2.1's running time is quadratic in the number of nodes.

---

**Algorithm 3.2.1:** Direct fiber sampler

---

```

3.2.1.1 Function SampleFiber( $n, k, \text{sense}, (\mathbf{a}, \mathbf{z})$ ) where
    Input:  $n, k \in \mathbb{N}, 1 \leq k \leq n, 2 \leq n$ : number of nodes, blocks
    Input: sense: implicit constraints on the allowed-edges set  $\mathcal{D}_n$ 
    Input:  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ : the observed state
    Output:  $B \in \mathcal{F}(\mathbf{a}, \mathbf{z})$  sampled uniformly at random

3.2.1.2  $B \leftarrow \mathbf{0} \in \mathcal{G}_n$  // Start  $B$  as the  $n \times n$  zero matrix.
3.2.1.3 for  $\sigma \leftarrow 1$  to  $k$  do // Lemma 3.1.6:  $\{\mathcal{E}_{\sigma\rho}^z\}_{\sigma\rho}$  partitions  $\mathcal{D}_n$ .
3.2.1.4     for  $\rho \leftarrow 1$  to  $k$  do // Definition 3.1.5 gives  $M_{\sigma\rho}(\mathbf{a}, \mathbf{z})$ .
3.2.1.5          $\mathcal{E} \leftarrow$  choose  $M_{\sigma\rho}(\mathbf{a}, \mathbf{z})$  edges uniformly at random from  $\mathcal{E}_{\sigma\rho}^z$ 
3.2.1.6         for each  $ij \in \mathcal{E}$  do
3.2.1.7              $B_{ij} \leftarrow 1$ 
3.2.1.8             if sense is undirected then  $B_{ji} \leftarrow 1$ 
3.2.1.9     return  $B$ 

```

---

**3.2.3.2 Simple-Hypothesis p-Value.** We now develop mc algorithm 3.2.2 on page 114 to compute  $p_\psi^z$  for the simple null hypothesis  $H_{0\psi}$  in eq. (3.16) for some edge probabilities  $\psi$  that we pick. Fix some observed state  $(\mathbf{a}, \mathbf{z}) \in \mathcal{Y}_{n,k}$ . Sub-subsection 3.2.3.3 discusses how to choose the number of mc iterations  $m$ , but for now just let  $m$  be an arbitrary positive integer. Let  $Q$  be the fiber-conditional probability measure such that, for all networks  $\mathbf{b}$ ,  $Q(\mathbf{A} = \mathbf{b}) := \mathbb{P}_\psi(\mathbf{A} = \mathbf{b} \mid \mathbf{A} \in \mathcal{F}(\mathbf{a}, \mathbf{z}))$ , so that  $Q$  is conditional on eq. (3.28).

Our goal is to approximate eq. (3.29). Sample the  $\mathcal{F}(\mathbf{a}, \mathbf{z})$ -valued random variables  $X_1, \dots, X_m$  such that they are iid uniformly under  $Q$ —just what algorithm 3.2.1 is for. The

real-valued random variables  $-W_\psi(\mathbf{X}_1, z), \dots, -W_\psi(\mathbf{X}_m, z)$  are IID under  $Q$ . Denote their common cumulative distribution function (CDF) under  $Q$  (i.e., conditional on the fiber) as  $P^\psi$ . This is also the CDF of  $-W_\psi(\mathbf{A}, z)$  when  $\mathbf{A} \in \mathcal{F}(\mathbf{a}, z)$ . Hence estimating  $P^\psi$  is helpful here because  $P^\psi(-W_\psi(\mathbf{a}, z)) = p_\psi^z(\mathbf{a})$  by eq. (3.27).

We do so by adapting Shiryaev (2016, p. 452) and Jacod and Protter (2004, pp. 184–185). For each  $t \in [m]$ , define the random function  $P_t^\psi: \mathbb{R} \rightarrow [0, 1]$  as the running average count of  $\mathbf{X}_t$ s whose discrepancy is smaller than some given  $w \in \mathbb{R}$ :

$$P_t^\psi(w) := \frac{1}{t} \sum_{\ell=1}^t \mathbb{1}(W_\psi(\mathbf{X}_\ell, z) \leq w).$$

$P_t^\psi$  is the *empirical distribution function*. By the strong law of large numbers,  $P_m^\psi(w)$  converges  $Q$ -almost surely and in  $\mathcal{L}^2(Q)$  to  $P^\psi(w)$  as  $m \rightarrow \infty$ .<sup>25</sup> By the Glivenko-Cantelli theorem,  $P_m^\psi$  converges uniformly  $Q$ -almost surely to  $P^\psi$  as  $m \rightarrow \infty$ . Algorithm 3.2.2 on the following page translates the computation of  $P_m^\psi$  into pseudo-code.

An implication of eq. (3.26) for valid p-values is that under the null hypothesis they “tend to be bigger than” (Casella & Berger, 2002, Prob. 1.49 on p. 44) a uniform(0, 1) random variable (Casella & Berger, 2002, Proof of Thm. 8.3.27 on pp. 397–398). Estimates of those p-values should approximate the same tendency. To check that algorithm 3.2.2’s estimates behave accordingly, we selected block assignments  $z$  and edge probabilities  $\psi$ , sampled  $\mathbf{A} \sim \gamma_\psi^z$ , and plotted the estimated values of  $p_\psi^z(\mathbf{A})$  in fig. 3.1 on page 115.

**3.2.3.3 Convergence Rate.** We now analyze the convergence rate of algorithm 3.2.2. Adopting the notation from sub-subsection 3.2.3.2, abbreviate  $p := p_\psi^z(\mathbf{a})$ , which is a number in  $[0, 1]$ , and  $P_m := P_m^\psi(-W_\psi(\mathbf{a}, z))$ , which is a  $[0, 1]$ -valued random variable. By the central limit theorem, the convergence rate of  $P_m$  to  $p$  as  $m \rightarrow \infty$  is  $\sqrt{m}$  in the sense that the random variable  $\sqrt{m}(P_m - p)$  converges in distribution to a normal random variable with mean zero and variance  $p(1 - p)$  (Jacod & Protter, 2004, pp. 184–185; Shiryaev, 2016, pp. 452–455). Convergence occurs under the probability measure  $Q$ , which accounts for the simple null

---

<sup>25</sup>↑The fiber is finite, so we will have sampled its entirety with high probability as  $m \rightarrow \infty$ . (If we sample the entire fiber with no repeats, then then our approximation becomes exact. The *coupon collector’s problem* asks what the minimum number of samples needed to cover the entire fiber is (O’Neill, 2020).) Equation (3.12) says that the fiber is very large, bounded above only by  $|\mathcal{G}_n| = O(2^{n^2})$ . We need not worry about the end of  $m$ ’s asymptotic runaway even for small  $n$  and small  $k$ .

**Algorithm 3.2.2:** Monte Carlo estimator of the simple-hypothesis p-value

---

$p_\psi^z$

---

```

3.2.2.1 Function ConditionalPValueForPMF( $n, k, \text{sense}, (\mathbf{a}, z), m, \psi$ ) where
    Input:  $n, k \in \mathbb{N}, 1 \leq k \leq n, 2 \leq n$ : number of nodes, blocks
    Input:  $\text{sense}$ : implicit constraints on the allowed-edges set  $\mathcal{D}_n$ 
    Input:  $(\mathbf{a}, z) \in \mathcal{Y}_{n,k}$ : the observed state
    Input:  $m \in \mathbb{N}, m > 0$ : number of samples to take
    Input:  $\psi \in \Psi_k$ : edge probabilities
    Output:  $P_m^\psi(-W_\psi(\mathbf{a}, z))$ 
3.2.2.2  $e, v \leftarrow \mathbb{E}_\psi(\mathbf{L}(\mathbf{A}, z)), \text{Var}_\psi(\mathbf{L}(\mathbf{A}, z))$  // Use eqs. (3.23) and (3.24).
3.2.2.3  $w \leftarrow W_\psi(\mathbf{a}, z)$  // Plug  $e, v$  into definition 3.2.5.
3.2.2.4  $P \leftarrow 0$  // Start recursive sum in line 3.2.2.8.
3.2.2.5 for 1 to  $m$  do
3.2.2.6      $\mathbf{X} \leftarrow \text{SampleFiber}(n, k, \text{sense}, (\mathbf{a}, z))$  // Call algorithm 3.2.1.
3.2.2.7      $w' \leftarrow W_\psi(\mathbf{X}, z)$  // Plug  $e, v$  into definition 3.2.5.
3.2.2.8     if  $w' \geq w$  then  $P \leftarrow P + 1$ 
3.2.2.9 return  $P/m$ 

```

---

hypothesis  $H_{0\psi}$  of eq. (3.16) and the conditions of eq. (3.28). That  $\text{Var}[\sqrt{m}(P_m - p)] \rightarrow p(1-p)$  means that  $P_m$  converges fastest when  $p$  is near zero or one and slowest when near  $\frac{1}{2}$ . (If  $p$  is too close to zero, we may no longer believe the null hypothesis.) Moreover, we can estimate the variance  $\text{Var}[P_m]$  of  $P_m$  itself using

$$V_m := \frac{1}{m^2} \sum_{t=1}^m [\mathbb{1}(W_\psi(\mathbf{X}_t, z) \leq -W_\psi(\mathbf{a}, z)) - P_m]^2;$$

in particular,  $(P_m - p)/\sqrt{V_m}$  converges in distribution to standard normal as  $m \rightarrow \infty$  (Robert & Casella, 2004, § 3.2 on pp. 83–84). See fig. 3.2 on page 120 for an example.

Moreover, for any  $\epsilon, \delta \in (0, 1)$ , we may ensure the probability is at least  $\epsilon$  of getting at least  $\log_{10} \delta$  digits of accuracy in  $P_m$  by solving for  $m$  in

$$\epsilon \leq Q(|P_m - p| \leq \delta) \approx \Phi\left(\frac{\delta\sqrt{m}}{p(1-p)}\right) - \Phi\left(\frac{-\delta\sqrt{m}}{p(1-p)}\right) = 2\Phi\left(\frac{\delta\sqrt{m}}{p(1-p)}\right) - 1,$$

where  $\Phi$  is the CDF of the standard normal distribution. We then get

$$m \gtrsim \left[ \frac{p(1-p)}{\delta} \Phi^{-1}\left(\frac{\epsilon+1}{2}\right) \right]^2.$$

When we do not have a reliable prior for the value of  $p$ , we may plug in  $p = \frac{1}{2}$  to maximize this lower bound. For the purpose of hypothesis testing, we really care about the number

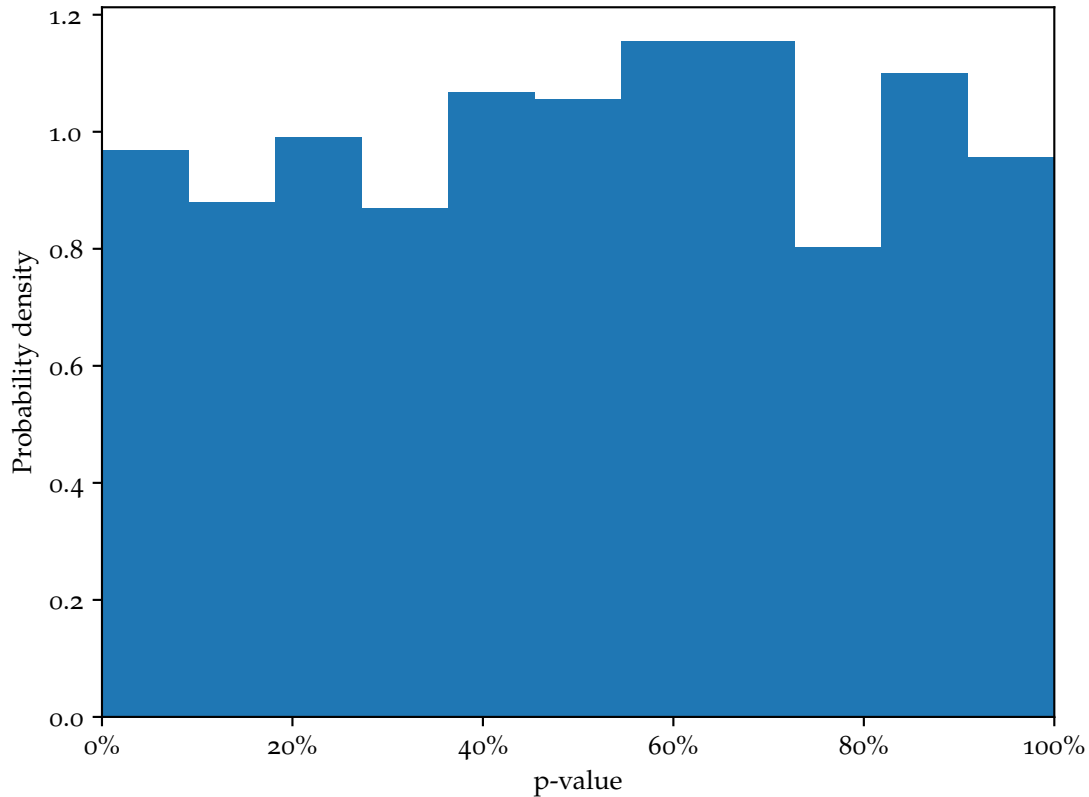


Figure 3.1. Histogram of a simple-hypothesis p-values from a single, conditional PMF. Histogram of 1,000 samples under the null hypothesis  $H_{0\psi}$  of  $P_{150}^{\psi}(-W_{\psi}(\mathbf{A}, z))$  estimating the simple-hypothesis p-value  $p_{\psi}^z(\mathbf{A})$ . All samples share a single, randomly chosen block assignments array  $z$  and a single matrix  $\psi$  of edge probabilities for  $n = 100$  nodes in  $k = 11$  blocks. Generated with version 0.3.0a0 of our `mmsbm` Python package with a pseudorandom-number-generator seed value of 0x73c84cc10c5a41a306f69107ee35c7b4.

of digits  $d \in \mathbb{N}$  of accuracy, so plugging in  $\delta = 10^{-d}$  is convenient. (Three digits of accuracy is probably enough for the social sciences.) Thus the number of iterations  $m$  to use in algorithm 3.2.2 should be

$$m \gtrsim \frac{100^d}{16} \left[ \Phi^{-1} \left( \frac{\epsilon + 1}{2} \right) \right]^2$$

For  $d = 3$  and  $\epsilon = .90$ ,  $m$  should exceed approximately 170 thousand. For  $d = 3$  and  $\epsilon = .99$ ,  $m$  should exceed approximately 420 thousand. Multiply by 100 for each additional digit. See fig. 3.3 on page 121. Keep in mind that this entire analysis assumes both that the null hypothesis is true (and that the number of graphs in the fiber is much larger than  $m$ , but see footnote 25 on page 113).

**3.2.4 Optimization.** The optimization problem in eq. (3.30) (or 3.32) is tricky. It defines the general-hypothesis p-value  $p^{z^*}(\mathbf{a}^*)$  and poses a serious impediment to estimating  $p^{z^*}(\mathbf{a}^*)$  for testing definition 3.2.1's general null hypothesis  $H_0$ . If we had at hand a maximizer  $\check{\psi}$ , then its simple-hypothesis p-value would equal the general-hypothesis p-value:

$$p_{\check{\psi}(\mathbf{a}^*, \mathbf{z}^*)}^{z^*}(\mathbf{a}^*) = p^{z^*}(\mathbf{a}^*). \quad (3.33)$$

We could pass  $\psi = \check{\psi}(\mathbf{a}^*, \mathbf{z}^*)$  to algorithm 3.2.2 to obtain an estimate of  $p^{z^*}(\mathbf{a}^*)$ . We could then plug that estimate into eq. (3.31) for a valid test of  $H_0$  (assuming algorithm 3.2.2 iterated long enough). The bad news is that we cannot offer even a suggestive reformulation of the optimization problem much less an algorithm to solve it.

The other news is that statisticians seem to disagree about whether we need to. If that's not good news, at least the statisticians who don't think we need to solve the optimization are not bothered by the "bad" news. Our development of the general-hypothesis p-value using the intersection-union method that led us to the optimization problem followed a school of thought that the textbook Casella and Berger (2002) exemplifies. Another school, which the textbook Bishop et al. (2007) exemplifies, might say that careful selection of a simple null hypothesis satisfies all the inferential needs of the general null hypothesis. That is, there may be edge probabilities  $\psi_0$  for which rejection of the simple null hypothesis  $H_{0\psi_0}$  convinces us to reject  $H_0$  as well even though we haven't tested any of the other PMFS in  $\mathcal{M}^{z^*}$ . In eq. (3.15)'s telling,  $H_0$  merely asserts the *existence* of some true edge probabilities  $\psi^*$ . We don't actually care which ones. In this sense the matrix of edge probabilities  $\psi$  is a *nuisance parameter*, one that complicates our computations but that is "not of direct inferential interest." (Casella & Berger, 2002, § 8.2.1 on p. 378)

Traditionally there have been five ways out of actually performing the optimization. Berger and Boos (1994, pp. 1012–1013) summarized the first three and introduced the fourth. The first way is to prove that some parameter value  $\bar{\psi}$  is the *least favorable configuration*. Here this means that  $\psi = \bar{\psi}$  maximizes  $p_{\psi}^{z^*}(\mathbf{a})$  over  $\psi \in \Psi_k$  for all  $\mathbf{a} \in \mathcal{G}_n$ . In some popular one-sided tests for continuous distributions, the entire boundary of the null model's parameter space is least favorable. We expect that neither latent MMSBMS nor conditional

MMSBMS have least favorable configurations. The second way is to choose a test statistic whose distribution does not depend on the parameter, which we already opted not to do. The third way is to condition on a sufficient statistic to remove the dependence of the probability distribution on the parameter, which we already did. Our optimization problem depends on  $\psi$  not because it appears in the probability distribution but because we are effectively using one test statistic  $W_\psi(\mathbf{A}, \mathbf{Z})$  per  $\psi$ . The fourth way, *confidence-set p-values*, was Berger and Boos's invention (Casella & Berger, 2002, § 8.5.4 on p. 415; Silvapulle, 1996, however, developed the same idea independently, having submitted his manuscript in June, 1994, three months before the publication of Berger and Boos's), wherein we would determine a parameter confidence set  $C(\mathbf{A}) \subseteq \Psi_k$  such that  $\mathbb{P}_\psi(\psi \in C(\mathbf{A})) \geq 1 - \beta$  for all  $\psi \in \Psi_k$  and some  $\beta \in [0, 1]$ . Then  $\mathbf{a} \mapsto \sup_{\psi \in C(\mathbf{a})} \mathbb{P}_\psi(W_\psi(\mathbf{A}, \mathbf{Z}) \geq W_\psi(\mathbf{a}, \mathbf{Z})) + \beta$  is a valid p-value. Unless we can offer a confidence set that is finite—and small at that—which we cannot, confidence-set p-values don't buy us out of our computational difficulties.

The fifth way around the optimization problem is to pick a  $\psi_0$  that is convincingly represents the general null model—the most representative configuration, as it were. This has usually meant picking  $\psi_0 = \hat{\psi}(\mathbf{a}^*, \mathbf{z}^*)$ , the MLE. Actually Bishop et al. (2007, § 14.7 esp. pp. 507–508) defined multinomial-model hypothesis testing without appeal to a supremum in the first place.<sup>26</sup> As subsection 3.2.1 mentioned, goodness-of-fit statistics based on distance functions measure the distance between the unconstrained MLE and an estimate constrained to be in the null model. Bishop et al. distinguished goodness-of-fit statistics' constrained estimators between those that minimize the distance function and those that do not (Bishop et al., 2007, pp. 507–508). In any event, they pointed out, MLE “for a multinomial model also corresponds to a minimum distance method of estimation” (Bishop et al., 2007, p. 505). Petrović et al. (2010, § 2.1 esp. pp. 264–265) went further in emphasizing the centrality of MLE to the three steps of “typical goodness-of-fit testing” they identified,<sup>27</sup> writing

<sup>26</sup>↑ Bishop et al. (2007, pp. 502–503) defined a null hypothesis analogous to our general null hypothesis in definition 3.2.1. Pick a null model  $\mathcal{A} \subseteq \Delta_{t-1}$ , the standard simplex in  $\mathbb{R}^t$ . Let  $\mathbf{X}$  be an  $\mathbb{N}^t$ -valued random vector whose distribution is multinomial with  $\sum_i X_i$  trials and unknown cell probabilities  $\phi \in \Delta_{t-1}$ . Their null hypothesis was  $H'_0: \phi \in \mathcal{A}$  against the alternative hypothesis  $H'_1: \phi \in \Delta_{t-1} \setminus \mathcal{A}$ .

<sup>27</sup>↑ The hypotheses that Petrović et al. (2010) tested were essentially the same as those of Bishop et al.

that the first step is computing the MLE and that, in the second step, “the goodness-of-fit statistic [. . .] measures how close the MLE is to the observed network[. . .]”. Petrović et al.’s final step mirrored Bishop et al.’s. The latter authors wrote, “Having computed the distance [. . .], the fit of the model is assessed as good or bad depending on the size of” (Bishop et al., 2007, p. 508) that value. This school of thought holds that the test statistic evaluated at one, well chosen parameter estimate  $\psi_0$  tells us everything we need to know about the entire null model.

Put differently, we choose  $\psi_0$  to be our best guess about how  $\mathbf{a}^*$  might have been sampled from a conditional MMSBM distribution. On the one hand, we know that  $p_{\psi_0}^{z^*}(\mathbf{a}^*) \leq p^{z^*}(\mathbf{a}^*)$ , so the former’s being large means we must not reject  $H_0$ . On the other hand, if  $p_{\psi_0}^{z^*}(\mathbf{a}^*)$  is small so that we don’t believe  $H_{0|\psi_0}$ , are we really going to believe  $H_{0|\psi}$  for any *worse* guesses  $\psi$  about how  $\mathbf{a}^*$  might have been sampled from a conditional MMSBM distribution?<sup>28</sup> Maybe we should! According to Berger and Boos (1994, p. 1013), who cited Storer and Kim (1990),  $p_{\psi_0}^{z^*}$  might not be valid for  $H_0$ , thereby rendering it prone to type I errors. Suppose for a moment that  $H_0$  is true and we chose  $\psi_0$  to be the MLE  $\hat{\psi}(\mathbf{a}^*, z^*)$ . In terms of eq. (3.15) the true conditional PMF  $\theta^{z^*}$  equals the  $z^*$ -conditional MMSBM PMF  $\gamma_{\psi^*}^{z^*}$ . It is *possible* that, for example,  $\psi^* \neq \psi_0$  so that  $\mathbb{P}_{\psi^*}(\mathbf{A} = \mathbf{a}^*) < \mathbb{P}_{\psi_0}(\mathbf{A} = \mathbf{a}^*)$ . If the likelihood function is relatively flat in a large enough neighborhood of  $\psi_0$  (see E. L. Lehmann & Casella, 1998, p. 470, for formulas for the derivatives of the log likelihood function of exponential families), it is possible that the likelihoods under  $\psi^*$  and  $\psi_0$  aren’t even that different. Using a simple-hypothesis p-value  $p_{\psi_0}^{z^*}$  to test the general null hypothesis ignores that possibility.

Unless that simple-hypothesis p-value equals the general-hypothesis p-value as in eq. (3.33). An estimator  $\check{\psi}$  of  $\psi$  that is easy to compute and also happens to maximize eq. (3.32) would make for the perfect most representative configuration  $\psi_0$ . Ideally the MLE

---

(2007) that we describe in footnote 26.

<sup>28</sup>↑Another advantage of plugging the MLE into the discrepancy is that every network in a fiber has the same MLE by lemma 3.2.3. Equation (3.32) says that the general-hypothesis p-value is just the proportion of networks in the observed fiber with a larger discrepancy. In some sense, using the MLE makes this an apples-to-apples comparison. “[A]ll the networks [] belonging to the same fiber will produce the same MLE and are, therefore, equivalent from the inferential standpoint.” (Petrović et al., 2010, p. 265).



$\hat{\psi}$  would fit the bill (but that's a taller order than conjecture 3.2.7). Finding such an estimator is the direction we suggest for future research and would open up a sixth way around the optimization problem.  $\check{\psi}$  differs from a least favorable configuration  $\bar{\psi}$  because  $\check{\psi}$  would be a statistic and thus depend on the random network  $\mathbf{A}$  whereas  $\bar{\psi}$  is a constant inherent to the null model.  $\check{\psi}$  differs from the most representative configuration  $\psi_0$  because  $\check{\psi}$  would provably solve the optimization problem. If such a  $\check{\psi}$  exists, it would solve the Casella et al. version of the problem by solving the optimization problem, and it would solve the Bishop et al. version of the problem by giving us a very convincingly representative  $\psi_0$  of the null model. If such a  $\check{\psi}$  exists, both schools of thought could be right.

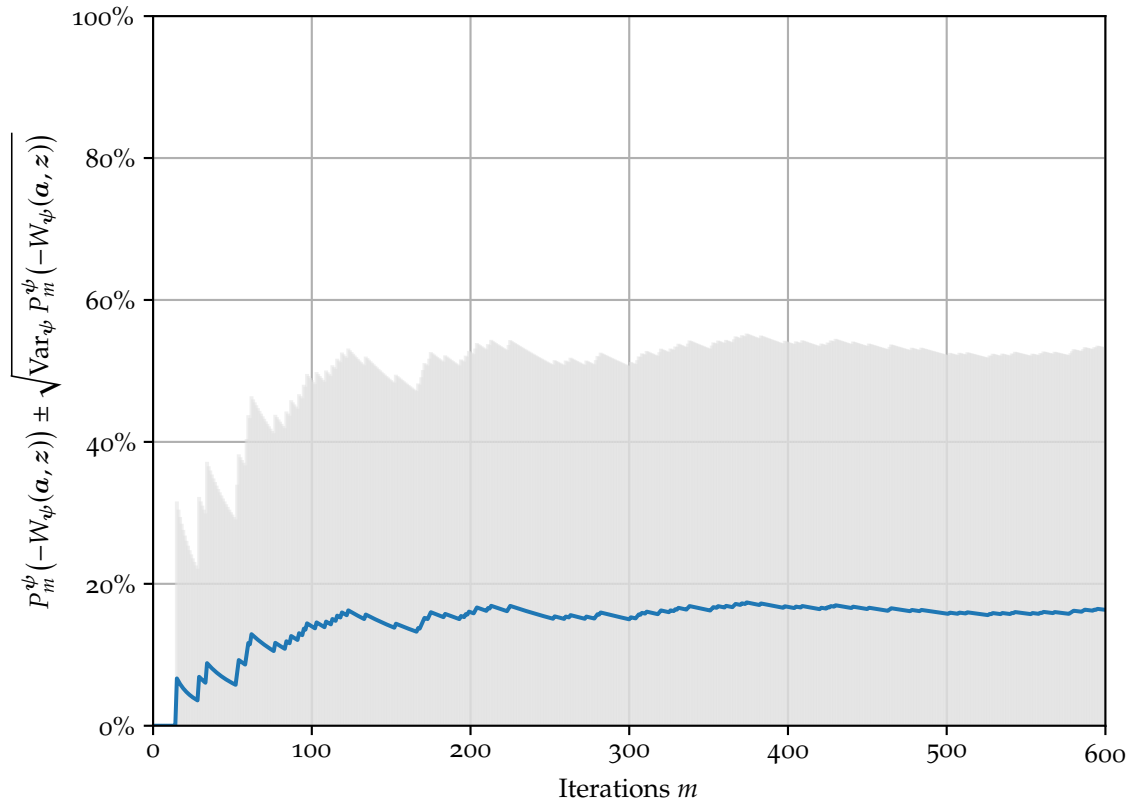


Figure 3.2. Convergence of p-value estimates. The simple-hypothesis p-value estimate  $P_m^\psi(-W_\psi(\mathbf{a}, z))$  at  $m$  iterations converges to  $p_\psi^z(\mathbf{a})$  after a few hundred iterations. It takes much longer for the standard deviations  $(\text{Var}_\psi P_m^\psi[-W_\psi(\mathbf{a}, z)])^{1/2}$  around the estimates (shown here in gray) to begin converging. All estimates share a single, randomly chosen block assignments array  $z$  on  $k = 11$  blocks, a single, randomly chosen edge probabilities matrix  $\psi$ , and a single network  $\mathbf{a}$  on  $n = 100$  nodes drawn from the corresponding  $z$ -conditional MMSBM distribution, i.e., drawn from the null model  $H_{0\psi}$ . Generated with version 0.3.0a0 of our `mmsbm` Python package with a pseudorandom-number-generator seed value of `0x174d3f66a8bd36300fbd1735b34ecfc3`.

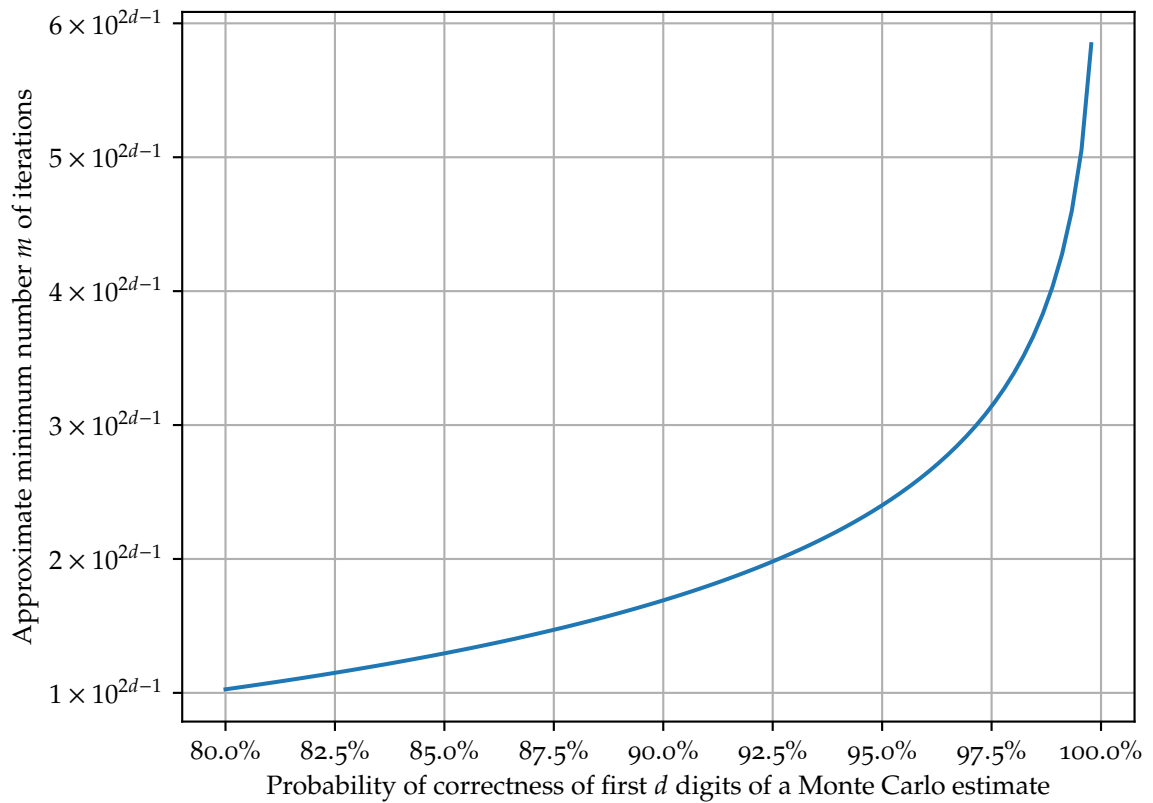


Figure 3.3. Approximate lower bound on the number of iterations needed in Monte Carlo integration. Approximate lower bound on the number  $m$  of iterations needed for a given probability of correctness of the first  $d$  digits of a p-value computed by Monte Carlo integration. Generated with version 0.3.0a0 of our `mmsbm Python` package with a pseudorandom-number-generator seed value of `0x8a6dde815ec68991cce6bdb1bb53620`.

## Chapter 4

## LEXICOGRAPHIC WINNER DETERMINATION

## 4.1 Model

Governments regularly auction publicly owned assets to private hands. More valuable assets have required increasingly sophisticated auction formats to facilitate bidders' complex demand schedules, allow heterogeneous supply, ensure efficiency and fairness, generate revenue, and deter collusion and fraud (Klemperer, 2004a, pp. 170–175; see also Day & Raghavan, 2007, pp. 1389–1390). Since 1994 when the United States switched from so-called “beauty contests”, in which a government panel selected winners based on contestants' submitted business plans, to auctions for allocating radio spectrum—licenses to the usufruct of specified ranges of radio frequencies—other governments have followed suit (Klemperer, 2004a, pp. 169–170). Among the most sophisticated auction formats in use today is the *combinatorial auction* (CA), which allows bidders to express demand that is non-linear across multiple units of multiple categories of goods or services in typically arbitrary combinations. Table 4.1 shows some recent examples to demonstrate their ongoing economic importance.

Whereas auction formats such as the “classic English auction of Sotheby's and Christie's” (Ausubel & Milgrom, 2006, p. 17) find a market-clearing combination of price and asset allocations *economically*, CAs clear the market *computationally*, and other mechanisms lie somewhere in between.<sup>29</sup> At the end of an English auction, the bidders who are still bidding are the winners, and they pay their last bid. At the end of a CA, the auctioneer has to feed the bids into an algorithm to find out who won and, in typical usage, how much they ought to pay. Bidders in a CA express demand for multiple combinations of discrete,

---

<sup>29</sup>↑“Thus, in the many real-world applications of CAs, the computational techniques of [operations research] facilitate more efficient economic outcomes in environments too complex for classical (i.e., noncomputational) economic theory.” (Day & Cramton, 2012, p. 588). “Some formal models show the equivalence between iterative CAs and decentralized optimization algorithms [ . . . ]” (Parkes, 2006, p. 42). Certain types of CAs are equivalent to certain LPS (Bikhchandani & Ostroy, 2006). Combinatorial auctions can vary in computational burden on the auctioneer, offloading some of it to bidders or to the economic mechanism itself (Pekeč & Rothkopf, 2006). See also Ausubel and Milgrom (2006, pp. 36–37) for a comparison of a type of CA and a type of multiunit, single-product ascending clock auction.

Table 4.1. Some recent combinatorial auctions for spectrum

Auctioneer	Year	Band	Revenue
United Kingdom <sup>a</sup>	2013	800 MHz & 2.6 GHz	£2.4 billion
Canada <sup>b</sup>	2014	700 MHz	£2.4 billion
United States <sup>c</sup>	2016–2017	600 MHz	\$19 billion
Ireland <sup>d</sup>	2017	3.6 GHz	€78 million
Canada <sup>e</sup>	2019	600 MHz	\$3.5 billion
Denmark <sup>f</sup>	2021	1.5 GHz, 2.1 GHz, 3.5 GHz, & 25 GHz	DKK2.1 billion
United States <sup>g</sup>	2021–	3.45–3.55 GHz	≥\$21 billion
Ireland <sup>h</sup>	upcoming	700 MHz; 2.1, 2.3, & 2.6 GHz	

Revenues are in local currency and not adjusted for inflation. <sup>a</sup>Office of Communications [Ofcom], 2013a. <sup>b</sup>Industry Canada, 2015. “The Commission adopts the assignment round bidding procedures proposed in the *Auction 1000 Comment PN* [ . . . ]” Federal Communications Commission [FCC], 2015, para. 2(j); this refers to FCC, 2014, app. H; only the assignment phase of Auction 1002 (the Incentive Auction’s forward auction) was combinatorial; the revenue includes the entire forward auction (FCC, 2017b, para. 2). <sup>c</sup>Commission for Communications Regulation [ComReg], 2017. <sup>d</sup>Bono et al., 2019; Innovation, Science and Economic Development Canada [ISED Canada], 2019b. <sup>e</sup>Danish Energy Agency [DEA], 2021a, 2021b. <sup>f</sup>FCC, 2021a, para. 238; only the assignment phase is combinatorial; the revenue is the “gross proceeds” as of round 124 on 2021, November 9 (FCC, 2021b). <sup>g</sup>ComReg, 2021.

indivisible goods—say, a *barrel* of water rather than *water*. Because the auctioneer can fit together multiple bidders’ bids, and the bids may be for product combinations that are not directly comparable, we cannot simply say, “highest bidder wins”. Table 4.2 on page 129 will show an example of a CA in which the highest bid loses. Instead, an algorithm solves the *winner determination problem* (WDP) to maximize the auctioneer’s revenue while awarding bidders only those packages they bid for and constrained by the auctioneer’s actual supply of those products.

We develop mathematical results to facilitate algorithm design for a problem formulation of WDP that we believe improves on the formulations available in the literature when it comes to modelling how governments use CAs today to sell radio spectrum. Spectrum is far from the only market using CAs (for a wide variety of applications of CAs specifically, see Cramton et al., 2006b, pt. v; and for other types of auction markets, see also Klemperer, 2004a, pp. 96–97), but it serves an outsize role in the multiunit-auction literature (“Much current work [on multiunit auctions] has been stimulated by the recent government auctions of radio spectrum licenses [for mobile telephony, etc.], and emphasises the problem of selling heterogeneous goods with complementarities between them, with common-value compo-

nents to bidders' valuations, and perhaps also externalities between bidders." Klemperer, 1999, endnote 78 on p. 271). We won't directly concern ourselves with bidders' valuations, economic efficiency, incentive compatibility, the desirability of auction features from an policy perspective, or other topics of the economics and policy literature, even as we use that literature to develop some mathematics.

The modern, game theoretic treatment of auctions first appeared in the economics literature with Vickrey (1961) (the paper "is still essential reading." Klemperer, 1999, p. 231). Economists have combined Vickrey's auction mechanism with Clarke (1971) and Groves (1973) to create the so-called *Vickrey-Clarke-Groves* (VCG) mechanism for multiunit auctions of heterogeneous items (Ausubel & Milgrom, 2006, p. 19), still the "lovely and elegant reference point" (Ausubel & Milgrom, 2006, p. 37) for judging other mechanisms if "not [...] a likely real-world auction design." (Ausubel & Milgrom, 2006, p. 37) Rassenti et al. (1982) first introduced the economics literature to specifically combinatorial auctions (Cramton et al., 2006c, endnote 4 on p. 12 from p. 10). (Both Smith and Vickrey went on to win Economics Nobel prizes (Cramton et al., 2006b, p. 631; Klemperer, 1999, p. 231).) However, the real explosion of interest in CAS began at the very end of the 1990s. Klemperer (1999), a lengthy survey of the auction literature as of 1999, did not mention CAS and included only a short section on multiunit auctions, but admitted that "this is probably the section of this survey that will become obsolete most quickly." (Klemperer, 1999, p. 241) Indeed by the time he republished the article as Klemperer (2004b), he added an afterword urging the reader to catch up on all that had happened in the fields of multiunit and combinatorial auctions in the intervening years (Klemperer, 2004b, pp. 62–65). Cramton et al. (2006b) collated that progress into a compendium of recent articles on the economics, computer science, and public policy dimensions of CAS (and related multiunit auctions). We lean heavily on several of its chapters throughout this report. Additionally auctioneer and regulatory documents (Hoffman et al., 2001; Power Auctions, LLC [Power Auctions], 2019) as well as professional consultants' white papers (Bono et al., 2019; Maldoom, 2007) have offered highly technical analysis, sometimes from the same academicians who write for the scholarly journals. Academic writing about CAS sometimes present a stylized WDP,

assuming away the complications that auctioneers add in the real world. Practitioners writing about CAS typically present a WDP specific to the auction at hand with an ad hoc formulation.

Subsections 4.1.2 and 4.1.3 aggregate the most popular of these complications into a unified and concise formulation called the *lexicographic winner determination problem*, which we abbreviate  $\text{WDP}^\times$ , and which subsumes classical WDP as a subset of instances. The formulation should better match the way that governments use CAS to allocate spectrum. Hopefully  $\text{WDP}^\times$  presents a shorter on-ramp to practitioners trying to implement auction solvers than traditional WDP literature.

The rest of the report proceeds as follows after subsection 4.1.1 establishes some notational conventions. Subsection 4.1.4 and sections 4.2 to 4.4 present a novel set of techniques for solving  $\text{WDP}^\times$ . Section 4.2 is a short aside on pre-solving techniques for simplifying input data. It explores the economic notion of *marginal value* for arbitrary, discrete, heterogeneous packages of products in a novel way with an application to finding dominated bids. The larger solution structure combines the branch-and-bound method with a top-down dynamic programming (DP) algorithm with memoization. Subsection 4.1.4 and section 4.3 define the branching and the bounding techniques, respectively. Subsection 4.3.1's main theorem, theorem 4.3.5, offers a new, heuristic explanation of why linear programming (LP) relaxations have been successful in helping solve classical WDP. Subsection 4.3.2's main theorem, theorem 4.3.7, extends LP relaxation techniques to  $\text{WDP}^\times$ , taking care of the complications that lexicographic optimization pose. We hope that other lexicographic combinatorial optimization problems can make use of many of the techniques it relies upon. Every branch-and-bound algorithm needs an actual optimization algorithm at its core. Often that algorithm is for solving LPS. Instead section 4.4 presents a DP algorithm combined with a novel memoization scheme adapted to lexicographic optimization. That section reviews the reasons for choosing DP, but one major advantage is that it supplies simple proofs for some results about  $\text{WDP}^\times$  (theorem 4.4.7 and proposition 4.4.8) that are not even specific to DP. Finally, section 4.5 explores some of the ways in which the DP solution speeds up computation of prices in certain CA mechanisms.

**4.1.1 Notation.** Denote the set of real numbers as  $\mathbb{R}$  and the set of nonnegative integers as  $\mathbb{N}$ . For  $p \in \mathbb{N}$ , abbreviate  $[p] := \{1, \dots, p\}$ ; in particular  $[0] = \emptyset$ . Define the *extended reals* as  $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ , and denote the vector of all  $\infty$ s as  $\infty$ . We use the usual conventions about addition and multiplication by  $\infty$  (see Jacod & Protter, 2004, p. 24):

$$\begin{aligned} \infty + \infty &= \infty, & \pm\infty \times 0 &= 0, \\ -\infty - \infty &= -\infty, & \infty \times a &= \infty \text{ for all } a \in (0, \infty], \\ a \pm \infty &= \pm\infty \text{ for all } a \in (-\infty, \infty), & \infty \times a &= -\infty \text{ for all } a \in [-\infty, 0). \end{aligned}$$

By convention, we define  $\max \emptyset := -\infty$  and  $\min \emptyset := \infty$  for the dimension appropriate for the context.

We say that  $x \in \mathbb{N}^p$  *lexicographically* precedes  $y \in \mathbb{N}^p$ , denoted  $x < y$ , whenever there exists an  $i \in [p]$  such that  $x_k = y_k$  for  $1 \leq k < i$  and  $x_i < y_i$  (For a discussion of order theory for integer vectors in terms of “monomial orders”, see Cox et al., 2015, pp. 2.2, 2.4). For example,  $(0, 2) < (1, 0)$  with  $i = 1$ . Lexicographic order is a *total order* on  $\mathbb{N}^p$ , meaning that every  $x, y \in \mathbb{N}^p$ , exactly one of  $x \leq y$  or  $x \geq y$  is true. Further, lexicographic order is compatible with addition, so that if we also have some  $z \in \mathbb{N}^p$ , then  $x < y$  implies that  $x + z < y + z$ . Finally, lexicographic order is a *well order*, meaning that every non-empty subset  $B$  of  $\mathbb{N}^p$  has a *minimum* or *least element*  $x \in B$ , so that for all  $y \in B$ ,  $x < y$ . The lexicographic order is also defined for real-valued vectors. In the real case,  $<$  ceases to be a well order.

While we stick with the usual entry-wise relation  $\leq$  for (extended) real vectors and matrices, we sometimes want to emphasize viewing vectors, especially packages (see subsection 4.1.2), as multisets. Suppose  $\mathcal{X}$  and  $\mathcal{Y}$  are multisets in a universe with  $p$  items, and they have corresponding characteristic vectors  $x, y \in \mathbb{N}^p$ . Then  $\mathcal{X} \subseteq \mathcal{Y}$  if and only if  $x \leq y$ , and  $\mathcal{X} \subset \mathcal{Y}$  if and only if  $x \not\leq y$ . The latter means that  $x \leq y$  and  $x \neq y$ , e.g.,  $(0, 2) \not\leq (1, 2)$ . The  $\leq$  relation on vectors is not a total order but rather a *partial order*: it may be the case that none of  $x \leq y$ ,  $x \geq y$ , or  $x = y$  hold. Such  $x$  and  $y$  are *not comparable*.

**Lemma 4.1.1.** *If  $x, y \in \overline{\mathbb{R}}^p$ , then  $x \leq y$  implies  $x \leq y$ . The converse is always true if and only if  $p = 1$ .*



*Proof.* Equality has the same meaning under both  $\leq$  and  $\preceq$ , so suppose that  $\mathbf{x} \preceq \mathbf{y}$ . Then  $x_k \leq y_k$  for all  $k \in [p]$ , and there is some least  $j$  for which  $x_j < y_j$ . Therefore  $x_k = y_k$  for  $k < j$  and  $x_j < y_j$ , i.e.,  $\mathbf{x} < \mathbf{y}$ .

If  $p = 1$ , then  $\leq$  and  $\preceq$  are both just the usual order on  $\mathbb{R}$ . A counterexample to prove that the converse fails if  $p > 1$  is  $\mathbf{x} := (0, 1, \dots)$  and  $\mathbf{y} := (1, 0, \dots)$ , where the ellipses stand for any values in the remaining  $p - 2$  coordinates. In this case  $\mathbf{x} < \mathbf{y}$ , but  $\mathbf{x}$  and  $\mathbf{y}$  are not comparable under  $\leq$ .  $\square$

**4.1.2 Formulation.** We are concerned with *multiunit* auctions in which some of the discrete, indivisible, and possibly heterogeneous *items* for sale are *perfect substitutes*, meaning that they're economically indistinguishable (D. Lehmann et al., 2006, p. 300). A *product* is an equivalence class of mutually substitutable items. For example, flights and train rides are products whereas specific seats are items. In this context *heterogeneity* means that there is more than one product. Let  $p \geq 1$  be the number of products for sale, each of which we assign a unique positional number in  $[p]$ . A *package* is a (column) vector  $\mathbf{q} \in \mathbb{N}^p$ , which represents a multiset of items by designating the number of items of each product. For some  $m \geq 1$ , we call (column) vectors  $\mathbf{v} \in \mathbb{R}^m$  *value* vectors. A pair  $(\mathbf{v}, \mathbf{q})$  of a value and a package vector constitutes a *package bid*, or just *bid*: an expression that a bidder is willing to pay  $v_1$ , the *bid price*, in exchange for the entire package  $\mathbf{q}$  but nothing at all for anything less.

Implicitly we are using a variant of the *exclusive-or* (XOR) *bidding language* for encoding bidders' valuations (Nisan, 2006, p. 220). The XOR bidding language has two distinct advantages. First it is fully expressive, meaning that it can express all valuations (Nisan, 2006, Prop. 9.2 on p. 220). Second it is the bidding language in actual use in several recent government auctions, including in the UK and Canada (Day & Cramton, 2012, p. 590; Ofcom, 2012, reg. 67.(4)(b); Industry Canada, 2013, para. 49; ISED Canada, 2018, para. 45).

Our formulation differs from the standard XOR bidding language in two ways. First we are using package vectors instead of *bundles*, or sets of items (Nisan, 2006, p. 219). We may think of packages as living in the quotient space of bundles modulo the product equivalence relation. More simply, we can break packages out into their individual items. After

this transformation, an XOR package bid becomes an XOR-of-OR bundle bid. For example, suppose there are two red marbles and two blue marbles for sale, so a package written in red-blue order might be  $q = (1, 2)$ . Once we break out the package  $q$  into an OR of bundles, it becomes  $\{\text{red}_1, \text{blue}_1, \text{blue}_2\}$  OR  $\{\text{red}_2, \text{blue}_1, \text{blue}_2\}$ .<sup>30</sup> The second difference is the possibility that the value vector has length  $m > 1$ . The first coordinate is always the monetary willingness to pay for the corresponding package. We call the subsequent entries *tie breakers*. In most real auctions the auctioneer, not the bidder, determines the values of the tie breakers. It makes no difference to our investigation of winner determination. One way or another, the auctioneer has in hand the entire value vector  $v$ .

We assume for the remainder of this report that  $b \geq 1$  bidders participated in a combinatorial auction. Each *bidder*  $j \in [b]$  submitted to the *auctioneer* bids  $(V_i, Q_i)$  indexed by positive integers  $i \in I^j \neq \emptyset$  such that the index sets  $I^1, \dots, I^b$  partition  $[n]$ , where  $n = \sum_j |I^j|$  is the total number of bids. The largest possible package, or *supply*, is  $s \in (\mathbb{N} \setminus \{0\})^p$ , so all other packages  $Q_i$  are such that  $Q_i \leq s$ . We place all the values as columns  $V_i, i \in [n]$ , of a *values matrix*  $V \in \mathbb{R}^{m \times n}$ , and all the packages as corresponding columns  $Q_i$  of a *packages matrix*  $Q \in \mathbb{N}^{p \times n}$ .

Table 4.2 on the following page illustrates a common representation of bids in a spreadsheet, called a *bid stack*, whose rows encode bids (transposing  $[V \ Q]$ ) labeled and sorted by bidder. Bid indices  $i$  are just the row numbers, which the sorting renders consecutive within each bidder's index set.<sup>31</sup> Bid stacks have been a popular data language among government auctioneers of radio spectrum (see, e.g., Ofcom, 2013b, Supplementary\_bids.csv; and ISED Canada, 2019a).

The classical *winner determination problem* for the XOR bidding language is to

---

<sup>30</sup>↑In this example the package is more succinct than ORs of bundles. The former requires four bits whereas the latter requires four bits per bundle times the number of bundles ORed together. However for any two bidding languages, there exists two valuations, one that is more succinct in the first bidding language and the other more succinct in the second (Nisan, 2006, p. 216).

<sup>31</sup>↑Formally, a bid stack is the matrix  $[j \ V^T \ Q^T]$ , where  $j \in [b]^n$  is a column vector labeling each row with the bidder who submitted that bid. Put differently,  $j_i$  is the unique solution to  $i \in I^{j_i}$ . Optionally we could stipulate that each of the bid index sets  $I^j$  contain only consecutive numbers.

Table 4.2. An illustrative bid stack

Bid $i$	Bidder $j$	Values $V^T$		Packages $Q^T$	
		\$	Rand.	Wine	Cheese
✓ 1	Alice	20	79	1	0
2	Alice	0	0	0	0
3	Bob	10	12	0	1
✓ 4	Bob	0	0	0	0
✓ 5	Carol	10	31	0	1
6	Carol	0	0	0	0
7	Dan	25	80	1	1
8	Dan	12	59	1	0
9	Dan	5	74	0	1
✓ 10	Dan	0	0	0	0

This is an example of a bid stack (orig. pub. as Schwartz, 2019) with  $n = 10$  bids from  $b = 4$  bidders, Alice, Bob, Carol, and Dan. They are bidding for  $p = 2$  products, a bottle of wine and a hunk of cheese. The supply of both products is one:  $s = [1, 1]^T$ . The values matrix  $V$  has  $m = 2$  rows, a dollar amount the bidder is willing to pay for each package and a random number to break ties. The sets of bid indices are  $\mathcal{I}^{\text{Alice}} = \{1, 2\}$ ,  $\mathcal{I}^{\text{Bob}} = \{3, 4\}$ ,  $\mathcal{I}^{\text{Carol}} = \{5, 6\}$ , and  $\mathcal{I}^{\text{Dan}} = \{7, 8, 9, 10\}$ . ✓ denotes the winning bids assuming no reserve price. The winning bids' total value is  $v^* = [30, 110]^T$ , which beats the highest single bid of \$25 by Dan. Carol's bid in row 5 beats Bob's in row 3, both of which are for the cheese without the wine, because of the tie breaker. Bob and Dan win their explicit *zero bids*  $(0, 0)$ , so they keep their cash, thirst, and hunger.

*allocate*<sup>32</sup> to each bidder exactly one of their bid packages in such a way that maximizes the *social-welfare function* while ensuring that the *winning allocation* does not exceed supply. The social-welfare function is the sum of bid prices corresponding to the bid packages that the winning allocation selects. Following Maldoom (2007, § 3) and D. Lehmann et al. (2006, p. 301), we write this problem as the zero-one integer linear program (ZOIP) in eq. (4.1) on the next page. The *decision variable*  $x_i$  for  $i \in [n]$  indicates whether the auctioneer selects bid  $i$  to win ( $x_i = 1$ ) or not ( $x_i = 0$ ); in particular,  $x$  represents an allocation as an indicator

<sup>32</sup>↑In contrast to an allocation's matching of packages of *products* to bidders, an *assignment* matches bundles of *items* to bidders (cf. Maldoom, 2007, § 1.1).

vector, and eq. (4.1)'s maximizers are winning allocations.

$$\begin{aligned}
 & \underset{\mathbf{x}}{\text{maximize}} && (\mathbf{V}\mathbf{x})_1 && (4.1) \\
 & \text{subject to} && \mathbf{Q}\mathbf{x} \leq \mathbf{s}, && \text{(supply)} \\
 & && \mathbf{x} \in \{0, 1\}^n, && \text{(nonnegativity and integrality)} \\
 & && \sum_{i \in I^j} x_i = 1, && j = 1, \dots, b. \quad \text{(exclusive or)}
 \end{aligned}$$

Since the first entry of each value vector is bid price,  $(\mathbf{V}\mathbf{x})_1$  is the sum of bid prices. The *decision version* of WDP returns whether a solution to eq. (4.1) exists with the the sum of bid prices equaling or exceeding some given number (D. Lehmann et al., 2006, p. 304). The *exclusive-or constraints* ensure that each bidder wins exactly one of its bids. The *supply constraints* ensure that the auctioneer has not sold more items than it has. The last constraint ensures each bid  $i$  either wins ( $x_i = 1$ ) or doesn't ( $x_i = 0$ ).

Many auctions have *reserve prices*: the least revenue an auctioneer accepts in exchange for a package (Cramton et al., 2006a, p. 622; Klemperer, 2004a, p. 109). We assume the auctioneer's valuation for all packages is linear, so we write *product  $t$ 's reserve price* as  $r_t \in \mathbb{R}$  and a *package  $q$ 's reserve price* as  $\mathbf{r}^\top \mathbf{q}$ , where  $\mathbf{r} := (r_1, \dots, r_p)$  is the vector of reserve prices. What *revenue* means depends on an how the auctioneer determines final payments from winners, which is the subject of section 4.5. For now, it suffices to say that if a bidder  $j$  is to win a package  $q$ , the auctioneer would charge the bidder  $\varphi^j(q)$  of currency. Reserve prices indicates the auctioneer's unwillingness to permit a bidder  $j$ 's winning any package  $q$  for which  $\varphi^j(q) < \mathbf{r}^\top \mathbf{q}$ . Recent auctions such as Canada's 700 MHz (2014) and 600 MHz (2019) spectrum auctions have implemented reserve prices by appending to the bid stack bids for each item of each product at the reserve price from separate dummy bidders for each quantity of demand (Industry Canada, 2013, para. 48; ISED Canada, 2018, para. 44; this idea has been known at least since Day & Cramton, 2012, p. 597). This ensures bidders must out bid the reserve price for every marginal unit of every product—at the computational cost of lengthening the bid stack by  $\sum_{t=1}^p (s_t + 1)$  bids. This is problematic because (the decision version of) WDP is NP-complete (D. Lehmann et al., 2006, Thm. 12.1 on p. 305).

Instead, we follow Day and Cramton (2012, pp. 597–599) in subtracting from the

social-welfare function the auctioneer's valuation of each package it's considering selling. (Incorporating reserves into the objective function loses no generality because we may always take  $r = 0$ .) Mathematically this means that definition 4.1.2 incorporates reserve prices directly into the objective function by adding  $r^\top(s - Qx)$ . Since that vector is fixed for any given  $x$ , we could follow Day and Cramton to the conclusion that we can replace the first row of  $V$  with those bid prices less  $r^\top Q$ , the length- $n$  row vector of reserve prices for each package. We could then solve WDP and add  $r^\top Q_i$  back into the final payments from each bidder winning a package  $Q_i$ . However, keeping reserve prices explicit in the formulation provides an analytical framework for understanding how reserve prices interact with subproblems and pricing. To this end, define the  $m \times p$  real *reserve-values matrix*

$$R := \begin{bmatrix} r^\top \\ \mathbf{0} \end{bmatrix} = e_1 r^\top, \quad (4.2)$$

where  $\mathbf{0}$  is the  $(m - 1) \times p$  matrix of all zeros and  $e_1$  is the first standard basis vector of  $\mathbb{R}^p$ . We assume that no bidder has submitted a bid below the reserve price:  $V_{1i} \geq r^\top Q_i$  for all  $i \in [n]$  (Clock auction rules generally prevent submitting bids below reserve prices. For example, see ISED Canada, 2018, para. 11; Day & Cramton, 2012, pp. 597–599, explained why this rule is economically desirable).

Most auctions have *tie-breaking rules* in case multiple combinations of bids achieve the same sum of bid prices. We model such tie-breaking rules using *multi-criteria* optimization, meaning that we combine multiple objective functions into one optimization according to some rule (Ehrgott, 2005). Tie-breaking rules in auctions typically require breaking ties among only the maximizers of eq. (4.1), and then maximizing another linear objective function within the optimal set of the previous optimization, and so on (For an example, see Power Auctions, 2019, para. 12). Performing this literal sequence of recursive optimizations is Algorithm 5.1 Ehrgott (2005, pp. 129–130). On behalf of the Federal Communications Commission (FCC), Hoffman et al. (2001, slides 13–16) first suggested breaking ties this way with a random linear objective function, arguing that randomly selecting among optimal winning combinations is fair and a ZOIP solution avoids enumerating the

possibly millions of tied combinations.<sup>33</sup> However, Maldoom (2007, § 4.4) did offer an algorithm for enumerating all ties. Pekeč and Rothkopf (2003, pp. 1498–1500) discussed why some auction mechanisms might experience more ties than others and some of the options for avoiding or breaking those ties. The authors point out that tie breaking is more important for government than private auctioneers because of the need for fairness and equal treatment as well as the risk of lawsuits. To generalize across multiple tie-breaking rules while avoiding the need to enumerate ties, we formulate a *lexicographic maximization* (Ehrgott, 2005, § 5.1 on pp. 129 sqq.), which we denote  $\max^{\prec}$ ,  $\arg \max^{\prec}$ , etc., in our case over the values vectors that constitute the columns of  $V$ . Auctioneers have explicitly used lexicographic optimization before for modified WDPs, such as the assignment phase of the FCC’s Auction 1002 (the Incentive Auction’s forward auction), which maximized three measures of spectrum contiguity (FCC, 2014, app. H, § 3). What we are offering here that is new is a unified way to view all the common auction rules in one formulation.

**Definition 4.1.2** (WDP<sup>⋄</sup>). The *lexicographic winner determination problem*, or WDP<sup>⋄</sup>, for the XOR bidding language is to find a maximizer  $x$  of the following lexicographic linear ZOLP, whose maximum we denote  $v^*$ .

$$\begin{aligned} & \underset{x}{\text{maximize}} \quad Vx + R(s - Qx) & (4.3) \\ & \text{subject to} \quad Qx \leq s, \\ & \quad \quad \quad x \in \{0, 1\}^n, \\ & \quad \quad \quad \sum_{i \in I^j} x_i = 1, \quad j = 1, \dots, b. \end{aligned}$$

Equation (4.3) (with  $R = 0$ ) is very similar to a  $p$ -dimensional version of the multi-criteria knapsack problem, about which there is a small literature; subsection 4.4.2 briefly discusses the multidimensional knapsack problem. Eben-Chaime (1996) discussed parametric solutions for the bi-criteria case ( $m = 2$ ) where, instead of lexicographic ordering,

---

<sup>33</sup>↑The FCC has continued using breaking ties with random numbers, including in its 2016–2017 Incentive Auction (see, e.g., FCC, 2017a, “Random Number” column). Canada’s 600 MHz (2019) spectrum auction used three tie breakers, the last of which was a random number (ISED Canada, 2018, paras. 46–47). The UK’s 4G (2013) spectrum auction used two tie breakers, the second of which was a random number (Ofcom, 2012, regs. 67.(5)–67.(6)).

the authors use the parameter to form a weighted sum of the two objective functions. Klamroth and Wiecek (2002) proposed several dynamic programming solutions to the multi-criteria knapsack problem. They note that most of the literature on the problem focuses on weighted sums of the objective functions. However, they point out, there is some literature on algorithms for finding Pareto optimal solutions or non-dominated points. The authors' dynamic programming solutions find the non-dominated solutions. Ehrgott and Gandibleux (2000) surveyed the state of the art—in 2000—of multi-criteria combinatorial optimization problems. It found that all the literature on the multi-criteria knapsack problem focused on finding Pareto optimal solutions or those optimizing certain parameterized dot products of the objective functions (Ehrgott & Gandibleux, 2000, § 6.8, p. 444).

Let us return for a moment to our comment that the decision version of WDP is NP-complete. A *decision problem* is one whose instance's solutions are either *true* or *false* (Cormen et al., 2009, p. 1054). We define the *decision version* of  $\text{WDP}^{\times}$  analogously to that of WDP: whether a solution to eq. (4.3) exists whose objective value  $v^* \geq a$  for some given  $a \in \overline{\mathbb{R}}^m$ . A decision problem is in **NP** if there is a polynomial-time algorithm that can verify a problem instance's solution given a certificate of that solution (Cormen et al., 2009, pp. 1049, 1063–1064). A *reduction* from a first decision problem to a second is an algorithm that converts instances of the first into instances of the second such that the solution to the first is *true* if and only if the solution to the second is *true* (Cormen et al., 2009, p. 1067). A decision problem is **NP-hard** if there is a polynomial time reduction from some NP-complete problem to the problem in question; a problem is **NP-complete** if it is in NP and is NP-hard (Cormen et al., 2009, pp. 1069, 1078–1079).

**Proposition 4.1.3.** *The decision version of  $\text{WDP}^{\times}$  is NP-complete.*

*Proof.* Given some certificate  $x \in \{0, 1\}^n$  and some target value  $a \in \overline{\mathbb{R}}^m$ , we can determine in polynomial time whether  $x$  is feasible for eq. (4.3) and whether  $Vx + R(s - Qx) \geq a$ . Therefore the decision version of  $\text{WDP}^{\times}$  is in NP.

As mentioned above, the decision version of WDP is NP-complete. Every instance of WDP is also an instance of  $\text{WDP}^{\times}$  with  $m = 1$ . That is, the identity is a reduction from the

NP-complete problem  $\text{WDP}$  to the NP problem  $\text{WDP}^\leq$ , so  $\text{WDP}^\leq$  is NP-complete.  $\square$

We return to computational complexity in subsection 4.4.2's theorem 4.4.7 on page 168 once theorem 4.4.4 gives us a recursive formula for  $\text{WDP}^\leq$ .

**4.1.3 Side Constraints.** Beyond the XOR and supply constraints, the combinatorial auction literature has paid some attention to *side constraints* (D. Lehmann et al., 2006, p. 307), which many real auctions have. For instance, the Canadian auction for the 600 MHz block had a “set-aside” constraint preventing the “set-aside ineligible” bidders from winning too much of any one product (Bono et al., 2019; ISED Canada, 2018, para. 45). In this subsection we describe how to implement a wide variety of side-constraints without modifying the form of eq. (4.3), and thus without writing new code. The key idea comes from linear optimization theory: *canonicalization*, or rewriting linear constraints in a standard form.

We do not intend to handle all possible side constraints. Real auctions use only linear constraints and are the ones we can handle easily. Such constraints can be written as a system of constraints  $\mathbf{a}^\top \mathbf{x} \leq d$ ,  $\mathbf{a}^\top \mathbf{x} = d$ , or  $\mathbf{a}^\top \mathbf{x} \geq d$ , where the  $\mathbf{a}$  is an  $n$ -vector and  $d$  is a scalar. We further restrict ourselves to considering only  $\mathbf{a}$ s and  $d$ s with (hopefully small) integer entries. This will keep our formulation compatible with the implementation ideas discussed in section 4.4. If the constraint comprises rational coefficients (all floating-point numbers are rational), we can multiply both sides by the least common multiple of all the denominators of the fractions written in lowest terms. Minoux (1983/1986, p. 249) explains how to reduce any linear integer program to a linear ZOIP.

The main idea is to rewrite side constraints as *virtual supply constraints* on *virtual products*. This is essentially the same idea as *dummy items* in Fujishima et al. (1999, August 6–/1999), and is in direct analogy with canonicalization for LPS as discussed in Bertsimas and Tsitsiklis (1997, pp. 5–6). We will construct a new packages matrix  $\tilde{Q}$  and new supply vector  $\tilde{s}$  to replace  $Q$  and  $s$  in eq. (4.3). Virtual products correspond to new rows appended to  $Q$  and new entries appended to  $s$  to form  $\tilde{Q}$  and  $\tilde{s}$ , respectively. Our main tool is the following lemma.

**Lemma 4.1.4.** *Let  $w, d \in \mathbb{R}$  and  $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$ . Suppose further that  $\mathbf{x}$  is a feasible solution to eq. (4.3).*



Then  $\mathbf{a}^\top \mathbf{x} \leq d$  if and only if  $(\mathbf{a} + w\mathbf{1})^\top \mathbf{x} \leq d + wb$ .

*Proof.* Since  $\mathbf{x}$  is a feasible solution to eq. (4.3), the XOR constraints say that, for each  $j \in [b]$ , we have  $1 = \sum_{i \in I^j} x_i$ , so

$$b = \sum_{j=1}^b 1 = \sum_{j=1}^b \sum_{i \in I^j} x_i = \sum_{i=1}^n x_i = \mathbf{1}^\top \mathbf{x}.$$

The second to last step works because the  $I^j$ 's partition  $[n]$ . The result then follows from reading the following inequality either forward or backward and then canceling.

$$(\mathbf{a} + w\mathbf{1})^\top \mathbf{x} = \mathbf{a}^\top \mathbf{x} + w\mathbf{1}^\top \mathbf{x} = \mathbf{a}^\top \mathbf{x} + wb \leq d + wb. \quad \square$$

If we let

$$w := -\min\{0, d, a_1, \dots, a_n\}$$

$$\hat{w} := \max\{0, d, a_1, \dots, a_n\}$$

in lemma 4.1.4, then  $\mathbf{0} \leq \mathbf{a} + w\mathbf{1}$ ,  $0 \leq d + wb$ ,  $\mathbf{0} \leq -\mathbf{a} + \hat{w}\mathbf{1}$ , and  $0 \leq -d + \hat{w}b$ . This allows us to construct  $\tilde{Q}$  and  $\tilde{s}$  while ensuring they contain only nonnegative integers (as long as  $\mathbf{a}$  and  $d$  have been normalized to be integers). Finally, the transformation algorithm works according to table 4.3.

Table 4.3. How to add side constraints to a canonical WDP<sup>Ⓢ</sup>

Form of side constraint	New columns to append to $\tilde{Q}^\top$	New entries to append to $\tilde{s}$
$\mathbf{a}^\top \mathbf{x} \leq d$	$\mathbf{a} + w\mathbf{1}$	$d + wb$
$\mathbf{a}^\top \mathbf{x} \geq d$	$-\mathbf{a} + \hat{w}\mathbf{1}$	$-d + \hat{w}b$
$\mathbf{a}^\top \mathbf{x} = d$	$\mathbf{a} + w\mathbf{1}$ and $-\mathbf{a} + \hat{w}\mathbf{1}$	$d + wb$ and $-d + \hat{w}b$

A solution  $\mathbf{x}$  is feasible for eq. (4.3) plus the side constraints if and only if it is feasible for eq. (4.3) with  $\tilde{Q}$  and  $\tilde{s}$  replacing  $Q$  and  $s$ . To prove this, observe that  $\mathbf{a}^\top \mathbf{x} = d$  is the same as  $\mathbf{a}^\top \mathbf{x} \leq d$  and  $\mathbf{a}^\top \mathbf{x} \geq d$  together; and that  $\mathbf{a}^\top \mathbf{x} \geq d$  is the same as  $-\mathbf{a}^\top \mathbf{x} \leq -d$ . Then apply lemma 4.1.4.

**4.1.4 Subproblems.** Let's break one big WDP<sup>Ⓢ</sup> into a bunch of smaller ones that are hopefully easier to solve. Such smaller problems, or *subproblems*, arise in algorithms based on both

*branch and bound* (B&B) and DP methodologies. In B&B the decision variables are partitioned, a part is chosen, the remaining decision variables are fixed at some prospective solution, the original problem is solved or estimated for the chosen part's decision variables, and either that combination of part/fixed-decisions is discarded as infeasible or the part is further partitioned, solved, and so on (Minoux, 1983/1986, pp. 248–253). These combinations of part/fixed-decisions form nodes in a *search tree* or *decision tree* (with edges describing subset relations between a parent's part and its children's partition). Solving the original problem after fixing decisions according to the node gives rise to one subproblem at each node. The scheme by which we partition the decision variables determines which *branch* of the decision tree to traverse next at each node in the tree. In DP recursion drives the search for an optimal solution, thereby giving rise to a *recursion tree* (Cormen et al., 2009, p. 37); the nodes of the recursion tree correspond directly to subproblems. Hence we can think about defining subproblems in terms of partitioning the decision variables, branching strategies, or recursion.

Sandholm (2006, § 14.2.1 on pp. 338–343) presented or mentioned formulations of subproblems for classical WDP that *branch on items* (“What bid should this item be assigned to?” (Sandholm, 2006, p. 338)); *branch on bids* (“Should this bid be accepted or rejected?” (Sandholm, 2006, p. 340)); and *branch on multivariate combinations* (“Of these eleven bids, are at least three winners?” (Sandholm, 2006, p. 342)). While the choice of branching strategy (which Sandholm called the *search formulation*) does not affect the size of the search tree (Sandholm, 2006, Prop. 14.2 on p. 342), combining a particular search formulation with a bounding strategy can substantially change the amount of the decision tree that the algorithm needs to search. Section 4.3 presents a bounding strategy for WDP<sup>∞</sup>, and subsection 4.4.2 briefly compares branching strategies.

Inspired by the recursive WDP formulae in Müller (2006, Thm. 13.7 on p. 333) and Maldoom (2007, Eq. (2) in § 4.1), the branching strategy we will consider is one that we call *branching on bids by bidder*: “Which one of this bidder's bids should win?” Each search-tree node has some *current bidder*  $j \leq b$  and some subset  $\mathcal{I}$  of bidder  $j$ 's bid indices  $\mathcal{I}^j$  indicating which of bidder  $j$ 's bids ( $V_i, Q_i$ ) we are considering accepting ( $x_i = 1$ )

or rejecting ( $x_i = 0$ ) from among the  $i \in \mathcal{I}$ . Sub-subsection 4.1.4.1 defines *subproblem* in terms of branching on bids by bidder, but further analysis of the motivation for it awaits subsection 4.4.2's recursive formula for subproblems, whereupon we will revisit branching.

We imagine fixing a prospective allocation of the supply  $s$  among some bidders not including the current bidder  $j$ , but now need to allocate the *residual supply*  $q \leq s$  ( $q \in \mathbb{N}^p$ ) among the remaining bidders including bidder  $j$ . Choosing for simplicity to keep bidders in the same order throughout the search tree—and choosing arbitrarily to work in descending order—let's say that the prospective allocation is for bidders  $j + 1, \dots, b$ . The auctioneer divvies up the residual supply  $q$  among the remaining bidders  $1, \dots, j$ .

But not just among the remaining bidders. As subsection 4.1.2 pointed out, the reserve-price rule is equivalent to the auctioneer's bidding for all packages at the reserve price. The auctioneer gets whatever the real bidders don't—that is the  $s - Qx$  term in eq. (4.3). Allocating to the auctioneer last will turn out to be convenient. Just as we have labeled the bidders  $1, \dots, b$ , we now label the auctioneer *zero*. However, we set  $\mathcal{I}^0 := \emptyset$  because we are expressing the auctioneer's reserve bids in the objective function rather than in the constraints of eq. (4.3). This way all formulas in the sequel work for a base case of the current bidder's being the auctioneer ( $j = 0$ ).

Maybe by the time we have gotten to a node in the search tree, we have already seen some feasible allocation. Its objective value gives us a lower bound  $a$  on the optimal value  $v^*$  of eq. (4.3) less the value of the prospective allocation to bidders  $j + 1, \dots, b$  (cf. the *MIN* variable in Sandholm et al., 2005; other lower bounds are discussed in Sandholm, 2006, pp. 349–350). We can give up in the middle of solving the subproblem as soon as we notice that the optimal value of the subproblem plus the value of the prospective allocation cannot exceed  $a$ . Summing the value of two independent allocations encourages us to think of subproblems as  $\text{WDP}^\times$  instances for “residual auctions” of the residual supply  $q$  to bidders  $0, 1, \dots, j$ . By analogy, we call  $a$  the *aggregate reserve* (AR) because the residual auctions cannot conclude unless the winning allocation's value exceeds it (see, e.g., FCC, 2021a, paras. 118–119). We consider any  $a \in \overline{\mathbb{R}}^m$  to accommodate completely relaxing the

constraint at the beginning of a search by setting  $\mathbf{a} = -\infty$ .<sup>34</sup>

**4.1.4.1 Notation.** For the sake of concision in definition 4.1.5 on the next page of *subproblem* as we use the term going forward, we introduce some notation to hide the details of the XOR constraints. Denote the standard basis of  $\mathbb{R}^n$  as  $e_1, \dots, e_n$ . For any  $\mathcal{J} \subseteq [n]$ , denote  $\mathcal{J}$ 's indicator vector as  $e_{\mathcal{J}} := \sum_{i \in \mathcal{J}} e_i$ , so  $e_{[n]} = \mathbf{1}$  and  $e_{\emptyset} = \mathbf{0}$ . Supposing  $\mathbf{x}$  is the vector of decision variables, then the dot product  $e_{\mathcal{I}}^{\top} \mathbf{x} = \sum_{i \in \mathcal{I}} x_i$  is the number of bids that  $\mathbf{x}$  accepts among bidder  $j$ 's bids indexed by  $\mathcal{I}$ . As we are considering only those of bidder  $j$ 's bid indices in  $\mathcal{I}$ , definition 4.1.5 ensures that subproblems are instances of WDP<sup>\*</sup> by requiring that the dot product equal one. This is the XOR constraint for bidder  $j$ , just without all of  $j$ 's bids. However if the current bidder is the auctioneer,  $j = 0$ , then  $\mathcal{I} \subseteq \mathcal{I}^j = \emptyset$  by definition, so the dot product is zero. To handle both cases, definition 4.1.5 requires that the dot product equal  $\min\{j, 1\}$ . We call this the *subset- $\mathcal{I}$  constraint*, or just the *subset constraint* if we don't know which subset. When  $\mathcal{I} = \mathcal{I}^j$ , bidder  $j$ 's XOR constraint  $\sum_{i \in \mathcal{I}^j} x_i = 1$  implies the subset- $\mathcal{I}$  constraint.

Define the *matrix of exclusive-or constraints* as the  $b \times n$  matrix of zeros and ones,

$$\mathbf{E} := \begin{bmatrix} e_{\mathcal{I}^1}^{\top} \\ \vdots \\ e_{\mathcal{I}^b}^{\top} \end{bmatrix}.$$

This way  $\mathbf{E}\mathbf{x} = \mathbf{1}$  if and only if  $\sum_{i \in \mathcal{I}^j} x_i = 1$  for all  $j = 1, \dots, b$ , which is all the XOR constraints. However, in a subproblem, bidders  $j+1, \dots, b$  are not bidding. Definition 4.1.5 imposes the  *$j$ th exclusive-or constraint* that  $\mathbf{E}\mathbf{x} = e_{[j]}$ , the  $b$ -vector with ones at coordinates  $1, \dots, j$  and zeros at  $j+1, \dots, b$ . When the current bidder is the auctioneer,  $j = 0$ , the zeroth XOR constraint that  $\mathbf{E}\mathbf{x} = e_{[0]} = \mathbf{0}$  together with the nonnegativity constraint imply  $\mathbf{x} = \mathbf{0}$  because  $\{\mathcal{I}^{\ell}\}_{\ell=1}^b$  partitions  $[n]$ .

Recall from the definition of  $\mathbf{R}$  in eq. (4.2) that the quantity  $\mathbf{R}\mathbf{q} = \begin{bmatrix} r^{\top} \mathbf{q} \\ \mathbf{0} \end{bmatrix}$  is a column vector whose first entry is the reserve price  $r^{\top} \mathbf{q}$  of the residual supply package  $\mathbf{q}$  and whose

<sup>34</sup>↑ An auctioneer with a real AR rule could initialize a solver with  $\mathbf{a}$  set appropriately. Even without a real rule, there may be economic or policy reasons that a practitioner might know *a priori* a finite lower bound. Starting search with  $\mathbf{a} > -\infty$  could speed up a B&B algorithm by pruning the search tree even before a feasible solution is observed.

subsequent  $m - 1$  entries are all zeros.

**Definition 4.1.5.** The  $\text{wDP}^x$  *subproblem* for current bidder  $j \in \{0, \dots, b\}$ , residual supply  $\mathbf{q} \in \mathbb{N}^p$  such that  $\mathbf{q} \leq \mathbf{s}$ , aggregate reserve  $\mathbf{a} \in \overline{\mathbb{R}}^m$ , and bid indices  $\mathcal{I} \subseteq \mathcal{I}^j$  is

$$\begin{aligned}
 & \underset{\mathbf{x}}{\text{maximize}} \quad \mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) & (4.4) \\
 & \text{subject to} \quad \mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) \geq \mathbf{a}, & \text{(aggregate reserve)} \\
 & \quad \mathbf{Q}\mathbf{x} \leq \mathbf{q}, & \text{(supply)} \\
 & \quad \mathbf{E}\mathbf{x} = \mathbf{e}_{[j]}, & \text{(}j\text{th exclusive or)} \\
 & \quad \mathbf{e}_{\mathcal{I}}^\top \mathbf{x} = \min\{1, j\}, & \text{(subset } \mathcal{I}\text{)} \\
 & \quad \mathbf{x} \in \{0, 1\}^n. & \text{(nonnegativity and integrality)}
 \end{aligned}$$

Ignoring decision variables  $x_i$  for  $i \in \bigcup_{\ell=j}^b \mathcal{I}^\ell \setminus \mathcal{I}$ , eq. (4.4) is just the  $\text{wDP}^x$  from eq. (4.3) with  $j$  replacing the maximum bidder index  $b$ ,  $\mathcal{I}$  replacing bidder  $j$ 's bid-index set  $\mathcal{I}^j$ , the residual supply  $\mathbf{q}$  replacing the supply  $\mathbf{s}$ , and, the aggregate reserve  $\mathbf{a}$  replacing, effectively,  $-\infty$  in an AR constraint that eq. (4.3) implicitly simplifies out. Thus we distinguish among subproblems of eq. (4.3) by the values of  $j$ ,  $\mathbf{q}$ ,  $\mathbf{a}$ , and  $\mathcal{I}$ . For each value of  $j$ , the triple  $(\mathbf{q}, \mathbf{a}, \mathcal{I})$  is in the domain set  $\mathcal{D}^j := \mathbb{N}^p \times \overline{\mathbb{R}}^m \times 2^{\mathcal{I}^j}$ . Throughout the remainder of this report, we will need repeatedly to refer to various aspects of eq. (4.4) or relaxations of it. Its nonnegativity, integrality, and  $j$ th XOR constraints depend at most on the current bidder  $j$ . The feasible set, set of maximizers, optimal value, and other notions depend on both  $j$  and  $(\mathbf{q}, \mathbf{a}, \mathcal{I})$ . In the notation below, the optimal value  $v^*$  of the full  $\text{wDP}^x$  in eq. (4.3) is  $v^b(\mathbf{s}, -\infty, \mathcal{I}^b)$ .

**Definition 4.1.6.** Denote eq. (4.4)'s set of feasible allocations as  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ , its subset of eq. (4.4)'s winning allocations (maximizers) as  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ , and eq. (4.4)'s optimal value as  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  (Fixing  $j$ ,  $\mathbf{a}$ , and  $\mathcal{I}$ , the *value function* of eq. (4.3) is  $\mathbf{q} \mapsto v^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ ). Nemhauser & Wolsey, 1999, § II.3.2 at p. 300). Define eq. (4.4)'s  *$j$ th-XOR-constrained polytope* and its integrality-constrained subset, as, respectively,  $\mathcal{P}^j := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{E}\mathbf{x} = \mathbf{e}_{[j]} \text{ and } \mathbf{x} \geq \mathbf{0}\}$  and  $\mathcal{E}^j := \mathcal{P}^j \cap \{0, 1\}^n$ .

We think of these functions as having default arguments:  $\mathcal{I}^j$  for  $\mathcal{I}$  and  $-\infty$  for  $\mathbf{a}$ . For each  $f \in \{\mathcal{X}, \mathcal{W}, \mathbf{x}, \mathbf{v}\}$ , we may write

$$f^j(\mathbf{q}, \mathbf{a}) := f^j(\mathbf{q}, \mathbf{a}, \mathcal{I}^j) \quad \text{and} \quad f^j(\mathbf{q}) := f^j(\mathbf{q}, -\infty) \quad (4.5)$$

**4.1.4.2 Basic Properties.** First let's recapitulate definition 4.1.6's notation with alternative expressions in case they're clearer. The definition covers eq. (4.4)'s feasible set, set of winning allocations, and the value of a winning allocation. We may think of them in a formal sense as functions, respectively,  $\mathcal{X}^j: \mathcal{D}^j \rightarrow 2^{\mathcal{E}^j}$ ,  $\mathcal{W}^j: \mathcal{D}^j \rightarrow 2^{\mathcal{E}^j}$ , and  $\mathbf{v}^j: \mathcal{D}^j \rightarrow \mathbb{R}^m \cup \{-\infty\}$  such that

$$\mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \left\{ x \in \mathcal{E}^j \left| \begin{array}{l} \mathbf{Q}x \leq \mathbf{q}, \\ \mathbf{V}x + \mathbf{R}(\mathbf{q} - \mathbf{Q}x) \geq \mathbf{a}, \text{ and} \\ \mathbf{e}_{\mathcal{I}}^\top x = \min\{1, j\} \end{array} \right. \right\},$$

$$\begin{aligned} \mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) &= \arg \max_x \mathbf{V}x + \mathbf{R}(\mathbf{q} - \mathbf{Q}x) & \mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) &= \max_x \mathbf{V}x + \mathbf{R}(\mathbf{q} - \mathbf{Q}x) \\ &\text{subject to } x \in \mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}), & &\text{subject to } x \in \mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}). \end{aligned}$$

Equation (4.4) is infeasible if and only if  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \emptyset$  if<sup>35</sup> and only if  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \emptyset$  if and only if  $\mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = -\infty$  by our convention that  $\max \emptyset = -\infty$ .  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  is exactly the set of feasible allocations  $x \in \mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  for which

$$\mathbf{V}x + \mathbf{R}(\mathbf{q} - \mathbf{Q}x) = \mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}). \quad (4.6)$$

When eq. (4.4) is feasible, eq. (4.6) is finite, and indeed  $\mathbf{V}x + \mathbf{R}(\mathbf{q} - \mathbf{Q}x)$  is finite for all  $x \in \mathcal{P}^j$ . This is because  $\mathbf{V}$ ,  $\mathbf{q}$ ,  $\mathbf{R}$ ,  $\mathbf{Q}$ , and  $x$  are finite.

We will use the following monotonicity result frequently enough to profit by packaging it into a lemma even though its proof is nothing more than “relax the AR constraint.”

**Lemma 4.1.7.** *The AR constraint in eq. (4.4) for the aggregate reserve  $\mathbf{a}$  is binding or infeasible if and only if  $\mathbf{a}$  lexicographically equals or exceeds the AR-unconstrained optimal value  $\mathbf{v}^j(\mathbf{q}, -\infty, \mathcal{I})$ . In*

<sup>35</sup>↑The *if* direction follows from the fact that  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  is the set of maximizers over the *finite* set  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ . Put differently, eq. (4.4) is either infeasible or has a finite optimal value; it cannot have an unbounded solution.

detail, the subproblem is infeasible if  $\mathbf{a} > \mathbf{v}^j(\mathbf{q}, -\infty, \mathcal{I})$ . If  $\mathbf{a} \leq \mathbf{v}^j(\mathbf{q}, -\infty, \mathcal{I})$ , then  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \mathcal{W}^j(\mathbf{q}, -\infty, \mathcal{I})$ , so  $\mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \mathbf{v}^j(\mathbf{q}, -\infty, \mathcal{I})$ .

An immediate corollary to lemma 4.1.7 is that eq. (4.4) is feasible only if  $\mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) \geq \mathbf{a}$ , but we can replace “only if” with “if and only if” when  $\mathbf{a} \neq -\infty$ . This holds even for the auctioneer ( $j = 0$ ), who implicitly bids for all packages at their reserve prices. That is, specializing lemma 4.1.7 for  $j = 0$  yields the following.

$$\mathcal{X}^0(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \mathcal{W}^0(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \begin{cases} \{\mathbf{0}\} & \mathbf{a} \leq R\mathbf{q} \\ \emptyset & \mathbf{a} > R\mathbf{q}, \end{cases} \quad \mathbf{v}^0(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \begin{cases} R\mathbf{q} & \mathbf{a} \leq R\mathbf{q} \\ -\infty & \mathbf{a} > R\mathbf{q}. \end{cases} \quad (4.7)$$

In subsection 4.4.2 the recursive formulation of  $\text{WDP}^\times$  subproblems in theorem 4.4.4 will use eq. (4.7) as a base case. Indeed, the simplicity of the formulas confirms the intuition we laid out in this subsection’s introduction that labeling the auctioneer as zero would be convenient.

We may use the aggregate reserve  $\mathbf{a}$  as a global “incumbent” lower bound because of the following proposition, which says that we can update  $\mathbf{a}$  to the best, feasible objective value seen so far at future search-tree nodes that examine more of the current bidder’s bids. This allows us to ratchet the AR constraint tighter without excluding any feasible, optimal solutions.

**Proposition 4.1.8.** *If we have eq. (4.4)’s maximum  $\mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  and we re-optimize with more of the current bidder’s bids, replacing the subset of bid indices  $\mathcal{I}$  with  $\mathcal{I}'$  for some  $\mathcal{I}' \subseteq \mathcal{I}^j$  such that  $\mathcal{I} \subseteq \mathcal{I}'$ , then we can replace the aggregate reserve  $\mathbf{a}$  with  $\mathbf{a}' := \max\{\mathbf{a}, \mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})\}$  without modifying the maximizers:*

$$\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}') = \mathcal{W}^j(\mathbf{q}, \mathbf{a}', \mathcal{I}'). \quad (4.8)$$

*Proof.* Since  $-\infty \leq \mathbf{a}$  and  $\mathcal{I} \subseteq \mathcal{I}'$ , substituting  $-\infty$  in for  $\mathbf{a}$  and  $\mathcal{I}'$  in for  $\mathcal{I}$  in eq. (4.4) relaxes the AR and subset- $\mathcal{I}$  constraints, so  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) \subseteq \mathcal{X}^j(\mathbf{q}, -\infty, \mathcal{I}')$ ;  $\mathbf{v}^j$  is maximization over the corresponding  $\mathcal{X}^j$ , so this implies  $\mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) \leq \mathbf{v}^j(\mathbf{q}, -\infty, \mathcal{I}')$ . Hence if  $\mathbf{a}' = \mathbf{v}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) \geq \mathbf{a}$ , then  $\mathbf{v}^j(\mathbf{q}, -\infty, \mathcal{I}')$  lexicographically equals or exceeds both  $\mathbf{a}$  and  $\mathbf{a}'$ . Lemma 4.1.7 therefore gives us both the left and right equalities,  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}') = \mathcal{W}^j(\mathbf{q}, -\infty, \mathcal{I}') = \mathcal{W}^j(\mathbf{q}, \mathbf{a}', \mathcal{I}')$ .

Otherwise  $a' = a > v^j(q, a, I)$ . If  $a' \leq v^j(q, -\infty, I')$ , then previous paragraph's conclusion still holds. Otherwise both  $a'$  and  $a$  lexicographically exceed  $v^j(q, -\infty, I')$ . Lemma 4.1.7 therefore gives us both the left and right equalities,  $\mathcal{W}^j(q, a, I') = \emptyset = \mathcal{W}^j(q, a', I')$ .  $\square$

## 4.2 Pre-Solving

This section considers operations on the bids of one bidder at a time that can speed up solving eq. (4.3). Since we only consider one bidder's bids here, we can simplify notation somewhat.  $V$  and  $Q$  are the matrices of just one bidder's bids, of which there are  $n$ . Bid  $i$  of the bidder's bids is  $(V_i, Q_i)$ . We also use  $j$  and  $k$  to denote bid indices.

**4.2.1 Dominated Bids.** Nothing in the auction rules that eq. (4.3) embodies prevents a bidder from outbidding himself. The resulting "dominated" bids can never win—we will prove this later on in proposition 4.4.8.

To set up the definition of *dominance*, consider a bidder's marginal value for adding one unit of some product: take the difference in values between two bids whose difference in packages is one unit of that product. More generally, if  $a, b \in [n]$  we can compute the  $a, b$  *margin* of  $(V_b, Q_b)$  over  $(V_a, Q_a)$  as  $V_b - V_a$  if  $Q_b - Q_a$  is a package, i.e., has all nonnegative entries, i.e.,  $Q_a \leq Q_b$ . To compute quickly or visualize tidily, we want to avoid dealing with the bid  $(V_a, Q_a)$  if there's another bid  $(V_c, Q_c)$  such that  $Q_a \leq Q_c \leq Q_b$  because  $V_b - V_a = (V_b - V_c) + (V_c - V_a)$ . That is, we can just compute the  $a, c$  and  $c, b$  margins, skipping computation of the  $a, b$  margin.

**Definition 4.2.1.** We say that a bid  $(V_a, Q_a)$  *dominates* another bid  $(V_b, Q_b)$  if  $Q_a \leq Q_b$  but  $V_a \geq V_b$ .

It is instructive to compare the definition of *dominance* with the following related concepts. Let  $Q \subseteq \mathbb{N}^p$ , and  $v: Q \rightarrow \mathbb{R}^m$ .  $q^* \in Q$  is *Pareto optimal* if no  $q \in Q$  exists such that  $v(q) \gneq v(q^*)$ , or, equivalently, if  $v(q) \geq v(q^*)$  implies  $v(q) = v(q^*)$ . Further,  $v(q^*)$  is a *non-dominated* point and  $q^*$  *dominates* any  $q$  for which  $v(q) \leq v(q^*)$  (Terminology varies across the literature, but we have presented that of Ehrgott, 2005, pp. 12, 23–24, after swapping Ehrgott's minimization viewpoint for WDP's maximization viewpoint). This meaning of



*dominates* is connected to definition 4.2.1's by Proposition 4.4.8 on page 168.

**Lemma 4.2.2** (Ehr Gott, 2005, Lem. 5.2 on p. 129). *If  $q^* \in Q$  and  $v(q) \leq v(q^*)$  for all  $q \in Q$ , then  $q^*$  is Pareto optimal.*

**4.2.2 Transitive Reduction.** We now develop an algorithm, culminating in proposition 4.2.5, to identify dominated bids.

If  $x$  and  $y$  are columns of  $Q$ , we say that  $y$  *covers* or is a *cover* of  $x$  in  $Q$  if  $x \preceq y$  and  $Q$  contains no  $z$  such that  $x \preceq z \preceq y$  (Hall, 1967, p. 15). The directed graph  $G$  whose nodes are the columns of  $Q$  (ignoring duplicates) and whose edges are its covering relations is called the *transitive reduction* of  $\leq$  on  $Q$  (cf. Aho et al., 1972, p. 135). That is  $x \rightsquigarrow_G y$  if and only if  $y$  covers  $x$  in  $Q$ .

**Lemma 4.2.3** (Aho et al., 1972). *The time complexity of an algorithm for computing the transitive reduction of a directed graph is at least a constant multiple of the time complexity of an algorithm for Boolean matrix multiplication.*

At the time of this writing, the best algorithm for matrix multiplication has time complexity worse than quadratic in the number of rows or columns.

We always orient directed edges upward from subset to superset to make it easy to draw *Hasse diagrams* of the packages with bigger packages higher up (Hasse diagrams are always drawn with edges implicitly pointing up. Gross & Yellen, 1999, p. 373).  $G$  is a directed, acyclic graph because  $\leq$  is a transitive relation.

**Lemma 4.2.4.** *The nodes  $x$  on a directed path in  $G$  ending at a node  $y$  are exactly the vectors satisfying  $x \leq y$ .*

*Proof.* A node in a graph is always on every path ending at itself, so we assume from now on that  $x \neq y$ .

( $\leq$ ). Suppose  $x$  lies on a directed path in  $G$  ending at  $y$ . Let  $\ell - 1$  be the length of the path, set  $x_1 := x$  and  $x_{\ell+1} := y$ , and denote the nodes in the interior of the path as  $x_i$  for  $i$  from 2 to  $\ell - 1$ , so that the path is

$$x = x_1 \rightsquigarrow x_2 \rightsquigarrow \cdots \rightsquigarrow x_\ell = y.$$

From the definition of  $G$ , we have  $x_i \preceq x_{i+1}$  for  $i \in [\ell]$ . Thus  $x \preceq y$ .

( $\geq$ ). Suppose  $x \preceq y$ . If  $y$  covers  $x$ , then  $x$  lies on the length-one path  $x \rightsquigarrow_G y$ . This establishes our base case. Otherwise, there's another column  $z$  of  $Q$  such that  $x \preceq z \preceq y$ . We assume recursively that there is a path from  $z$  to  $y$ . This recursion must terminate in the base case because  $Q$  has a finite number ( $n$ ) of columns.  $\square$

For bids we extend the notion of transitive reduction beyond the bids' packages. The *transitive reduction of bids*  $(V, Q)$  is the directed graph  $H$  constructed from the transitive reduction  $G$  of  $\leq$  on  $Q$  as follows. The main work here is being careful about *duplicate* bids: bids for the same package. For each column  $x$  of  $Q$ , create a node in  $H$  labeled

$$\min \left( \arg \max_{i \in [n]: Q_i = x} V_i \right)$$

That is, we take the indices of the lexicographically maximally valued bids for package  $x$ , and then break ties by picking the one with the lowest index. Then, for each pair of nodes  $i$  and  $j$  in  $H$ , we let  $H$  have the edge  $i \rightsquigarrow_H j$  if and only if  $G$  has the edge  $Q_i \rightsquigarrow_G Q_j$ .

The transitive reduction of bids  $H$  can help us identify dominated bids as follows. Define a vector  $M \in (\{-1\} \cup [n])^n$  by, for all nodes  $j \in H$ ,

$$M_j := \begin{cases} -1 & \text{if no edge of } H \text{ points at } j \\ \arg \max_{i \in [n]: Q_i \preceq Q_j} V_i & \text{else} \end{cases}$$

If there are ties in the maximization, we can break them arbitrarily.

**Proposition 4.2.5.** *Define  $V_{-1} := (-\infty, \dots, -\infty)^T \in \overline{\mathbb{R}}^m$ . Then a bid indexed by  $j$  is dominated if and only if  $V_{M_j} > V_j$ .*

*Proof.* ( $\implies$ ). Suppose  $j$ 's bid is dominated. Then there is some  $i \in [n]$  with  $Q_i \preceq Q_j$  and  $V_i > V_j$ . Thus the set over which  $M_j$  maximizes is non-empty. Then  $V_{M_j} > V_j$  from the definition of  $M$ .

( $\impliedby$ ). Suppose  $V_{M_j} > V_j$ . Since every vector in  $\mathbb{R}^m$  lexicographically succeeds  $V_{-1}$ , we know that  $M_j \neq -1$ , and hence that some edge of  $H$  points at  $j$ , say  $i \rightsquigarrow_H j$ . This implies that  $Q_i \preceq Q_j$ , so the set over which  $M_j$  maximizes is non-empty. Then  $Q_{M_j} \preceq Q_j$  from the definition of  $M$ , hence  $M_j$ 's bid dominates  $j$ 's.  $\square$

We can compute  $M$  using depth-first search in  $H$ 's *transpose*  $H^T$ . This is the directed graph with the same vertex set as  $H$  in which, for all nodes  $i, j$ , we have  $i \rightsquigarrow_{H^T} j$  if and only if  $j \rightsquigarrow_H i$ . In practice we compute  $H^T$ , not  $H$ .

The following result is useful for constructing  $H^T$ . Recall that in a directed graph, the *predecessors* of a node  $j$  are the nodes from which an edge points to  $j$ , and the *successors* of  $j$  are the nodes to which an edge points from  $j$ .

**Proposition 4.2.6.** *Suppose the bids are ordered so that  $Q_1 \geq Q_2 \geq \dots \geq Q_n$ . For  $i \in [n]$ , let  $H_i$  be the transitive reduction of bids indexed by  $[i]$  (i.e.,  $H$  but ignoring any bids after the first  $i$  bids). Then the successors of  $i$  in  $H^T$  are exactly the same as the successors of  $i$  in  $H_i^T$ .*

*Proof.* The successors of  $i$  in  $H^T$  are exactly those  $k \in [n]$  for which  $Q_i \leq Q_k$  and no  $j \in [n]$  exists for which  $Q_i \leq Q_j \leq Q_k$ . The successors of  $i$  in  $H_i^T$  have the same definition with  $n$  replaced by  $i$ . Thus, to prove the result, it suffices to show that all such  $k \in [n]$  satisfy  $k \leq i$ .

Suppose  $k \in [n]$  for which  $Q_i \leq Q_k$ . By lemma 4.1.1,  $Q_i \leq Q_k$ , which implies that  $k < i$  because of our ordering of the bids. □

### 4.3 Upper Bounds

Solving  $v^*$  in eq. (4.3) via B&B requires having upper bounds on subproblems' solutions. Subsection 4.1.4 defined and briefly analyzed the *branching* scheme for generating subproblems. This section develops various *bounds* for those subproblems (Minoux, 1983/1986, p. 252). In subsection 4.3.1 we define the LP relaxation and Lagrangian dual of WDP<sup>⋈</sup> subproblems, concluding that they are equal in theorem 4.3.5, the subsection's main result. In subsection 4.3.2 we develop a fathoming algorithm from the LP relaxation in the subsection's main result, theorem 4.3.7. Any B&B algorithm for WDP<sup>⋈</sup> can employ theorem 4.3.7. We have in mind to use it in a combination with the DP algorithm in section 4.4.

**4.3.1 Dual Relaxations.** Linear programming relaxations have been experimentally successful upper bounds in WDP, e.g., in Sandholm et al. (2005). The main result of this subsection explains why: theorem 4.3.5 says that the optimal value of the LP relaxation equals that of the Lagrangian dual, which in general is the tighter bound for a ZOIP. Techniques other than LP exist for finding upper bounds, but Sandholm (2006, p. 348) found in a review of

the literature at the time that for general WDP, LP is superior.

The main difficulty in formulating the standard relaxations of  $\text{WDP}^\times$  is the lexicographic maximization, which bundles a sequence of  $<$  programs into a single  $<$  program. Ehrgott (2005, Alg. 5.1 on pp. 129–130) says that  $\text{WDP}^\times$ 's bundling of the sequence of tie-breaker optimizations is reversible: When optimizing the objective function from row  $k$  of the values matrix  $\mathbf{V}$ , add a constraint to the program ensuring that the  $k$ th program meets the  $(k - 1)$ 's optimal value. This forms a sequence of  $k - 1$  AR constraints using  $\geq$  in direct analogy to the single AR constraint using  $\geq$  that appears in the subproblem in eq. (4.4).

So instead of relaxing the  $\text{WDP}^\times$  subproblem in eq. (4.4), we relax the sequence programs in eq. (4.9) below for  $k \in [m]$ . Let  $e_1, \dots, e_m$  denote the standard basis of  $\mathbb{R}^m$ .

**Definition 4.3.1.** For current bidder  $j \in \{0, \dots, b\}$ , tie breaker  $k \in [m]$ , residual supply  $\mathbf{q}$ , and aggregate reserve  $\mathbf{a}$ , we define the  $k$ th *tie-breaker subproblem* to be the following linear ZOIP:

$$\begin{aligned}
 & \underset{\mathbf{x}}{\text{maximize}} && e_k^\top [\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x})] && (4.9) \\
 & \text{subject to} && e_1^\top \mathbf{V}\mathbf{x} + \mathbf{r}^\top (\mathbf{q} - \mathbf{Q}\mathbf{x}) \geq a_1, \\
 & && \vdots \\
 & && e_{k-1}^\top \mathbf{V}\mathbf{x} \geq a_{k-1}, \\
 & && \mathbf{Q}\mathbf{x} \leq \mathbf{q}, \\
 & && \mathbf{x} \in \mathcal{E}^j.
 \end{aligned}$$

Equation (4.9)'s AR constraints (those involving  $\mathbf{a}$ ) look different from its objective function only because eq. (4.2)'s definition of the reserve-values matrix  $\mathbf{R} = e_1 \mathbf{r}^\top$  allows us to simplify  $e_k^\top \mathbf{R}$  depending on  $k$ .

With a small enough aggregate reserve  $\mathbf{a}$ , the tie-breaker subproblem in eq. (4.9) and the  $\text{WDP}^\times$  subproblem in eq. (4.4) have the same optimal values  $v_k^j(\mathbf{q}, \mathbf{a})$  at the first few indices  $k$ . We won't bother explicating *small enough* or *few* because we won't solve the tie-breaker subproblem directly in practice. Instead, definition 4.3.2 on the following page

defines the LP relaxation of the tie-breaker subproblem, and theorem 4.3.7 on page 152 details the connection between the LP relaxation and  $v^j(\mathbf{q}, \mathbf{a})$ .

**Definition 4.3.2.** The *linear programming relaxation* of the tie-breaker subproblem in eq. (4.9) is the following LP, whose optimal value we denote  $z_k^j(\mathbf{q}, \mathbf{a})$ :<sup>36</sup>

$$\begin{aligned}
 & \underset{\mathbf{x}}{\text{maximize}} && e_k^\top [\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x})] && (4.10) \\
 & \text{subject to} && e_1^\top \mathbf{V}\mathbf{x} + \mathbf{r}^\top(\mathbf{q} - \mathbf{Q}\mathbf{x}) \geq a_1, \\
 & && \vdots \\
 & && e_{k-1}^\top \mathbf{V}\mathbf{x} \geq a_{k-1}, \\
 & && \mathbf{Q}\mathbf{x} \leq \mathbf{q}, \\
 & && \mathbf{x} \in \mathcal{P}^j.
 \end{aligned}$$

The only difference between eqs. (4.9) and (4.10) is that the latter relaxes the integrality constraint by maximizing over the  $j$ th-XOR-constrained polytope  $\mathcal{P}^j$  in place of its integrality-constrained subset  $\mathcal{E}^j = \mathcal{P}^j \cap \{0, 1\}^n$ . LPS' optimal values occur at the extreme points, or basic feasible solutions (BFS) (Bertsimas & Tsitsiklis, 1997, Thm. 2.3 in § 2.2), of their feasible sets (Bertsimas & Tsitsiklis, 1997, § 2.6 on pp. 65–67).  $\mathcal{P}^j$  is not the feasible set of eq. (4.10), but the following lemma does hint at trying to find a related mathematical program whose feasible set is either  $\mathcal{P}^j$  or  $\mathcal{E}^j$ . After the lemma's proof, we will see just such a program, the Lagrangian relaxation, whose feasible set is  $\mathcal{E}^j$ .

**Lemma 4.3.3.** *The set of  $\mathcal{P}^j$ 's extreme points is  $\mathcal{E}^j$ , and the convex hull of  $\mathcal{E}^j$  is  $\mathcal{P}^j$ .*

*Proof.* Write the  $i$ th column of  $\mathbf{E}$  as  $\mathbf{E}_i$ . Because  $\mathbf{E}$  has orthogonal—and thus linearly independent—rows,  $\mathbf{x} \in \mathbb{R}^n$  is a BFS for  $\mathcal{P}^j$  if and only if  $\mathbf{x} \geq \mathbf{0}$ ,  $\mathbf{E}\mathbf{x} = \mathbf{e}_{[j]}$ , there exists  $\mathcal{B} = \{i_1, \dots, i_b\} \subseteq [n]$  such that  $\mathbf{E}_{i_1}, \dots, \mathbf{E}_{i_b}$  is linearly independent in  $\mathbb{R}^b$ , and  $x_i = 0$  if  $i \notin \mathcal{B}$  (Bertsimas & Tsitsiklis, 1997, Thm. 2.4, § 2.3). Note that if  $i \in I^j$ , then  $\mathbf{E}_i = \mathbf{e}_j$  by the definition of  $\mathbf{E}$ . Thus linear independence of  $\mathbf{E}_{i_1}, \dots, \mathbf{E}_{i_b}$  requires  $|\mathcal{B} \cap I^j| = 1$  for each  $j \in [b]$ . Without loss of generality, we may write  $\mathcal{B}$  so that  $i_j \in I^j$  for each  $j \in [b]$ . Thus, if

<sup>36</sup>↑Formally,  $z^j: \mathbb{R}^p \times \overline{\mathbb{R}}^m \rightarrow (\mathbb{R} \cup \{-\infty\})^m$  as eq. (4.10) is either infeasible or has a finite optimum.

$\mathbf{x} \in \mathcal{P}^j$ , then  $\mathbf{x}$  is a BFS if and only if such a set  $\mathcal{B}$  exists and  $i \notin \mathcal{B}$  implies  $x_i = 0$ . This can happen if and only if we have

$$\mathbf{e}_{\mathcal{I}^k}^\top \mathbf{x} = x_{i_k} + \sum_{\substack{i \in \mathcal{I}^k \\ i \neq i_k}} x_i = x_{i_k} = \begin{cases} 1 & k \in [j], \\ 0 & k \in \{j+1, \dots, b\}. \end{cases} \quad (4.11)$$

$\mathbf{x} \in \mathcal{P}^j$  satisfies eq. (4.11) if and only if  $\mathbf{x} \in \mathcal{E}^j$ , and hence  $\mathbf{x} \in \mathbb{R}^n$  is a BFS if and only if it's in  $\mathcal{E}^j$ .

Finally, the convex hull of  $\mathcal{E}^j$ , or the set of all convex combinations of elements of  $\mathcal{E}^j$ , is  $\mathcal{P}^j$ . This is because  $\mathcal{P}^j$  is a bounded polyhedron, and such sets are the convex hulls of their extreme points (Bertsimas & Tsitsiklis, 1997, Thm. 2.9, § 2.7). To see that  $\mathcal{P}^j$  is bounded, observe that its elements are nonnegative, so, for all  $\mathbf{x} \in \mathcal{P}^j$ ,  $\|\mathbf{x}\|_1 = \sum_{k=1}^b \sum_{i \in \mathcal{I}^k} x_i = \sum_{k=1}^j 1 = j$ .  $\square$

As usual, the linear ZOIP in eq. (4.9) is bound from above by its LP relaxation in definition 4.3.2, but ZOIPs generally get tighter bounds from their Lagrangian duals (Bertsimas & Tsitsiklis, 1997, p. 499). We develop eq. (4.9)'s by loosely adapting Bertsimas and Tsitsiklis (1997, § 11.4). The trick is to keep the nonnegativity, integrality, and  $j$ th XOR constraints of  $\mathcal{E}^j$  but to *dualize* the AR and supply constraints, replacing them with new terms in the objective function. In the method of *Lagrange multipliers* each new term is the product of a *dual variable* times the removed constraints' *violations*  $\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) - \mathbf{a}$  and  $\mathbf{q} - \mathbf{Q}\mathbf{x}$ , respectively. For each  $k \in [m]$ , we write the dual variables as  $\Delta_k \in \mathbb{R}^p$  for the supply constraints, and  $\Gamma_{k\ell} \in \mathbb{R}$  for each AR constraint  $\ell \in [k-1]$ . For convenience, let  $\Delta$  be the  $m \times p$  matrix whose rows are the  $\Delta_k^\top$ s. Form the lower triangle of the  $m \times m$  matrix  $\Gamma$  from the  $\Gamma_{k\ell}$ s, filling the diagonal and upper triangle with anything (we won't use them). We now define the *Lagrangian relaxation* (Nemhauser & Wolsey, 1999, p. 324) of the tie-breaker subproblem in eq. (4.9) for fixed dual variables  $\Delta$  and  $\Gamma$ , current bidder  $j$ , residual supply  $\mathbf{q}$ , and aggregate reserve  $\mathbf{a}$ , as the following linear ZOIP, whose optimal

value we denote  $\zeta_k^j(\Delta, \Gamma)$ :<sup>37</sup>

$$\begin{aligned} & \underset{\mathbf{x}}{\text{maximize}} \quad e_k^\top (\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x})) + \Delta_k^\top (\mathbf{q} - \mathbf{Q}\mathbf{x}) + \sum_{\ell=1}^{k-1} \Gamma_{k\ell} e_\ell^\top (\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) - \mathbf{a}) \quad (4.12) \\ & \text{subject to} \quad \mathbf{x} \in \mathcal{E}^j. \end{aligned}$$

The goal is to use  $\zeta^j$  as an upper bound for the optimal value  $v^j(\mathbf{q}, \mathbf{a})$  of the  $\text{wDP}^\times$  subproblem in eq. (4.4). However, we must be careful with how the tie-breaker subproblem in eq. (4.9) converts the  $\text{wDP}^\times$  subproblem's AR constraint from a lexicographic inequality to a sequence of regular inequalities. As lemma 4.1.7 points out, eq. (4.4)'s lexicographic inequality is feasible only if  $v^j(\mathbf{q}, \mathbf{a}) \geq \mathbf{a}$ . In that case, the definition of  $\geq$  means that, for some  $k \in [m]$ ,

$$v_\ell^j(\mathbf{q}, \mathbf{a}) = a_\ell \text{ for each } \ell \in [k-1] \quad (4.13)$$

The next lemma says  $\zeta^j$  indeed bounds  $v^j(\mathbf{q}, \mathbf{a})$  in those first  $k$  entries if dual variables are nonnegative (*cf.* Bertsimas & Tsitsiklis, 1997, Lem. 11.1 on p. 494). We can always cheat our way into applying the lemma: eq. (4.13) is trivially true for  $k = 1$ .

**Lemma 4.3.4.** *If  $\Delta$  and  $\Gamma$  have all nonnegative entries and eq. (4.13) holds for some  $k \in [m]$ , then  $\zeta_\ell^j(\Delta, \Gamma) \geq v_\ell^j(\mathbf{q}, \mathbf{a})$  for each  $\ell \in [k]$ .*

*Proof.* Assume that the  $\text{wDP}^\times$  subproblem in eq. (4.4) is feasible—if it weren't, then our convention says that  $v^j(\mathbf{q}, \mathbf{a}) = -\infty \leq \zeta(\Delta, \Gamma)$ . Choose any winning allocation  $\mathbf{x} \in \mathcal{W}^j(\mathbf{q}, \mathbf{a})$ . Further, assume that  $\Delta$  has all nonnegative entries. Then  $\mathbf{Q}\mathbf{x} \leq \mathbf{q}$ , so  $\Delta(\mathbf{q} - \mathbf{Q}\mathbf{x}) \geq \mathbf{0}$ .

Additionally suppose that eq. (4.13) holds for some  $k \in [m]$  and that  $\Gamma$  has all nonnegative entries. By eq. (4.6),  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}^j) = \mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x})$ , which is finite. Thus the first  $k-1$  entries of  $\mathbf{a}$  are not  $\infty$  and of  $\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) - \mathbf{a}$  are nonnegative, as is their product with the corresponding entries of  $\Gamma$ . For any  $\ell \in [k]$ , the first inequality below follows from the fact that  $\mathbf{x} \in \mathcal{E}^j(\mathbf{q}, \mathbf{a})$ , and the second from the foregoing non-negativity

---

<sup>37</sup>↑Formally,  $\zeta^j: \mathbb{R}^{m \times p} \times \mathbb{R}^{m \times m} \rightarrow \mathbb{R}^m \cup \{-\infty\}$  as eq. (4.12) is either infeasible or has a finite optimum for all  $k \in [m]$ .

facts.

$$\begin{aligned}
\zeta_\ell^j(\Delta, \Gamma) &\geq e_\ell^\top(\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x})) + \Delta_\ell^\top(\mathbf{q} - \mathbf{Q}\mathbf{x}) + \sum_{h=1}^{\ell-1} \Gamma_{\ell h} e_h^\top(\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) - \mathbf{a}) \\
&\geq e_\ell^\top(\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x})) \\
&= v_\ell^j(\mathbf{q}, \mathbf{a}).
\end{aligned}$$

□

The problem of finding the best possible bound on the  $\text{wDP}^\times$  subproblem in eq. (4.4) from the Lagrangian relaxation in eq. (4.12) is the *Lagrangian dual* (Nemhauser & Wolsey, 1999, p. 324) for  $k \in [m]$ , denoted

$$\hat{\zeta}_k^j := \min_{\Delta, \Gamma \geq 0} \zeta_k^j(\Delta, \Gamma). \quad (4.14)$$

This is well defined for a single choice each of  $\Delta$  and of  $\Gamma$  across all  $m$  coordinates. That is, there are minimizers  $\hat{\Delta}$  and  $\hat{\Gamma}$  of eq. (4.14) such that  $\hat{\zeta}_\ell^j = \zeta_\ell^j(\hat{\Delta}, \hat{\Gamma})$  for all  $\ell \in [m]$ . In eq. (4.12), the  $k$ th entry of  $\zeta^j(\Delta, \Gamma)$  uses only the  $k$ th rows of  $\Delta$  and  $\Gamma$ , so the minimization in eq. (4.14) is independent for different values of  $k$ .

The importance of the next theorem, which says that the LP relaxation and the Lagrangian dual have equal optima, is that the LP relaxation is an upper bound for the  $\text{wDP}^\times$  subproblem (under the right conditions), and a relatively tight one at that. By lemma 4.3.4,  $\hat{\zeta}_\ell^j \geq v_\ell^j(\mathbf{q}, \mathbf{a})$  for  $\ell \in [k]$  whenever  $\mathbf{a}$  is small enough in its first  $k - 1$  entries. The same is true of the LP relaxation's optimum  $z^j(\mathbf{q}, \mathbf{a})$  in place of  $\hat{\zeta}^j$  since they're equal. This is reassuring because definition 4.3.2 defines eq. (4.10) to be the LP relaxation of the tie-breaker subproblem in eq. (4.9), not of the  $\text{wDP}^\times$  subproblem in eq. (4.4) for which  $v^j$  gives the optimal value. An LP relaxation is not always as tight a bound as a Lagrangian dual (Bertsimas & Tsitsiklis, 1997, p. 499), but in this case they're the same. This result extends from subproblems to  $\text{wDP}^\times$  by taking  $\mathbf{a} = -\infty$ ,  $j = b$  and  $\mathbf{q} = \mathbf{s}$ ; and the result extends from  $\text{wDP}^\times$  to classical WDP by taking  $m = 1$ .

**Theorem 4.3.5.** *The tie-breaker subproblem's LP relaxation and Lagrangian dual have the same optimal value (even if the LP relaxation is infeasible).*

*Proof.* Fixing an arbitrary  $k \in [m]$ , the goal is to prove that the Lagrangian dual's optimal value  $\hat{\zeta}_k^j = z_k^j(\mathbf{q}, \mathbf{a})$ , the LP relaxation's optimal value. Lemma 4.3.3 on page 147 says that the



convex hull of  $\mathcal{E}^j$  has integer extreme points, so Bertsimas and Tsitsiklis (1997, p. 500) say that the optimal value of eq. (4.9)'s LP relaxation equals the optimal value of its Lagrangian dual problem (see also Nemhauser & Wolsey, 1999, Cor. 6.6 in § II.3.6 on p. 329). Equations (4.10) and (4.14) match Bertsimas and Tsitsiklis's definition of LP relaxation and Lagrangian dual for the tie-breaker subproblem in eq. (4.9) when  $a_1, \dots, a_{k-1}$  are all finite, so  $\hat{\zeta}_k^j = z_k^j(\mathbf{q}, \mathbf{a})$  in that case. However, we need to check what happens when one of the AR entries is infinite.

First suppose  $a_\ell = \infty$  for some entry  $\ell \in [k-1]$  but that  $a_{\ell'} > -\infty$  for all  $\ell' \in [k-1]$ . Then eq. (4.10) is infeasible due to its  $\ell$ th AR constraint, so  $z_k^j(\mathbf{q}, \mathbf{a}) = -\infty$ . For any  $\hat{\mathbf{x}} \in \mathcal{E}^j$ , the  $\ell$ th entry of  $\mathbf{V}\hat{\mathbf{x}} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\hat{\mathbf{x}}) - \mathbf{a}$  is  $-\infty$ . To minimize  $\zeta_k^j(\Delta, \Gamma)$  for any finite  $\Delta \geq \mathbf{0}$  over finite  $\Gamma \geq \mathbf{0}$ , eq. (4.14) may achieve  $\hat{\zeta}_k^j \leq \zeta_k^j(\Delta, \Gamma) = -\infty$  by choosing  $\Gamma_{k\ell} > 0$  because our convention is that  $\gamma \times -\infty = -\infty$  for any finite, positive  $\gamma$ . Therefore  $\hat{\zeta}_k^j = -\infty = z_k^j(\mathbf{q}, \mathbf{a})$ .

Second suppose  $a_\ell = -\infty$  for some entry  $\ell \in [k-1]$  but that  $a_{\ell'} < \infty$  for all  $\ell' \in [k-1]$ . For any  $\hat{\mathbf{x}} \in \mathcal{E}^j$ , the  $\ell$ th entry of  $\mathbf{V}\hat{\mathbf{x}} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\hat{\mathbf{x}}) - \mathbf{a}$  is  $\infty$ . For any dual variables  $\Delta$  and  $\Gamma$  for which  $\Gamma_{k\ell} \neq 0$ ,  $\zeta_k^j(\Delta, \Gamma) = \infty$ . To minimize  $\zeta_k^j(\Delta, \Gamma)$  for any finite  $\Delta \geq \mathbf{0}$  over finite  $\Gamma \geq \mathbf{0}$ , eq. (4.14) may achieve  $\hat{\zeta}_k^j \leq \zeta_k^j(\Delta, \Gamma) < \infty$  only by choosing  $\Gamma_{k\ell} = 0$  because our convention is that  $0 \times \infty = 0$ . Moreover, by choosing  $\Gamma_{k\ell'} = 0$  for all  $\ell' \in [k-1]$  for which  $a_{\ell'} = -\infty$ , we get a finite  $\hat{\zeta}_k^j$ , as though the  $\ell'$ th AR constraint were not involved in eq. (4.12) in the first place. Indeed, that is exactly the case for eqs. (4.9) and (4.10): AR constraints for which the AR is  $-\infty$  can never be binding and thus act as though they are not part of the definitions of those problems. Therefore the conclusion of equality we came to in the case when  $a_1, \dots, a_{k-1}$  are all finite also applies when some or all of them are  $-\infty$ .

Finally, if the first  $k-1$  entries of  $\mathbf{a}$  contain both  $\infty$  and  $-\infty$ , then eq. (4.14) minimizes the Lagrangian dual by choosing the  $\Gamma_{k\ell}$ s so that it's positive in the columns corresponding to  $\mathbf{a}$ 's  $\infty$ s and zero in the columns corresponding to  $\mathbf{a}$ 's  $-\infty$ s. Then  $\hat{\zeta}_k^j = -\infty = z_k^j(\mathbf{q}, \mathbf{a})$ , same as above.  $\square$

**4.3.2 Fathoming.** Theorem 4.3.7 below suggests an algorithm to *fathom*—preemptively solve or skip—the subproblem  $v^j(\mathbf{q}, \mathbf{a})$ .

*Remark 4.3.6.* The algorithm works as follows. Using an LP software package such as **GLPK**,

Gurobi, `lp_solve`, and CPLEX, solve the LP relaxation in eq. (4.10) for each  $k$  until  $z_k^j(\mathbf{q}, \mathbf{a})$  differs from  $a_k$ . If  $z_1^j(\mathbf{q}, \mathbf{a})$  is infeasible (item 1) or  $z_k^j(\mathbf{q}, \mathbf{a}) < a_k$  (item 2c), then *prune*—declare infeasible—the subproblem. If the LP’s decision variables are all zeros and ones, then either tighten the aggregate reserve  $\mathbf{a}$  (item 4(b)i) or use the decision variables as a solution to the subproblem (item 4(b)ii).

When reading theorem 4.3.7, think of  $k^*$  as the “current” optimization that the algorithm is running, and note that  $\mathbf{a}$  is already tight for entries one through  $k^* - 1$  (item 2b).

To abate some of this and the next subsections’ index chasing, we define the notation  $\mathbf{d}_{[k]} := (d_1, \dots, d_k) \in \overline{\mathbb{R}}^k$ , the first  $k \in [m]$  entries of any  $m$ -vector  $\mathbf{d} \in \overline{\mathbb{R}}^m$ . We define propositions of the form  $\mathbf{d}_{[0]} \leq \mathbf{d}'_{[0]}$  to be (trivially) true for any  $\mathbf{d}, \mathbf{d}' \in \overline{\mathbb{R}}^m$ .

**Theorem 4.3.7.** *Let  $k^* \in [m]$  such that  $z_{[k^*-1]}^j(\mathbf{q}, \mathbf{a}) = \mathbf{a}_{[k^*-1]}$  (which is trivially true if  $k^* = 1$ ).*

1.  $v_1^j(\mathbf{q}) \leq z_1^j(\mathbf{q}, \mathbf{a})$ .
2.  $z^j(\mathbf{q}, \mathbf{a})$  bounds  $v^j(\mathbf{q}, \mathbf{a})$  as follows.
  - (a)  $v_{[k^*]}^j(\mathbf{q}, \mathbf{a}) \leq z_{[k^*]}^j(\mathbf{q}, \mathbf{a})$ , and  $v_{k^*+1}^j(\mathbf{q}, \mathbf{a}) \leq z_{k^*+1}^j(\mathbf{q}, \mathbf{a})$  if  $k^* < m$ .
  - (b) If  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}) \neq \emptyset$ , then  $v_{[k^*-1]}^j(\mathbf{q}, \mathbf{a}) = z_{[k^*-1]}^j(\mathbf{q}, \mathbf{a})$  and  $a_{k^*} \leq v_{k^*}^j(\mathbf{q}, \mathbf{a}) \leq z_{k^*}^j(\mathbf{q}, \mathbf{a})$ .
  - (c) If  $z_{k^*}^j(\mathbf{q}, \mathbf{a}) < a_{k^*}$  then  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}) = \emptyset$  and eq. (4.10) is infeasible for  $k \in \{k^* + 1, \dots, m\}$ .
3. If  $k^* > 1$  and  $a_1 > -\infty$ , then eq. (4.10) is feasible for  $k = k^*$  and its maximizers are optimal for  $k < k^*$ .
4. Suppose that  $\mathbf{x}$  is a feasible solution to eq. (4.10) for  $k = k^*$  and that  $\xi_{k^*} \geq a_{k^*}$ , where we abbreviate the objective value  $\xi := \mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x})$ .
  - (a) If  $k^* < m$ , then  $\xi_{k^*+1} \leq z_{k^*+1}^j(\mathbf{q}, \mathbf{a})$ .
  - (b) Suppose  $\mathbf{x}$  is integral.
    - i. If  $\xi_{k^*} > a_{k^*}$  or  $k^* = m$ , then  $\mathbf{x} \in \mathcal{X}^j(\mathbf{q}, \mathbf{a})$  and  $\mathcal{W}^j(\mathbf{q}, \xi) = \mathcal{W}^j(\mathbf{q}, \mathbf{a})$ .
    - ii. If  $k^* = m$  and  $\mathbf{x}$  is optimal, then  $\mathbf{x} \in \mathcal{W}^j(\mathbf{q}, \mathbf{a})$ .

Note that item 2a implies that, if  $k^* \geq m - 1$ , then  $v^j(\mathbf{q}, \mathbf{a}) \leq z^j(\mathbf{q}, \mathbf{a})$ .

Items 3 and 4 assume that  $x$  is a feasible—not necessarily (except item 4(b)ii) optimal—solution. When  $m = 1$ , Item 4(b)ii is classical WDP’s “INTEGER case”, as Sandholm (2006, pp. 355–356) described it, in which the LP relaxation yields an optimal, integer solution  $x$ . However, for WDP<sup>s</sup> with  $k^* < m$  (or when  $x$  is not optimal), item 4(b)i does not permit fathoming the subproblem with an optimal solution.

What item 4(b)i does permit is that, if we find an integer solution  $x$  such that  $\xi_{k^*} > a_{k^*}$ , then we can *replace*  $a$  with  $\xi = Vx + R(q - Qx)$ , analogously to proposition 4.1.8. In particular,  $v^j(q, a) = v^j(q, \xi)$ , and, by item 3,  $a_{[k^*-1]} = \xi_{[k^*-1]}$ , so  $z_{[k^*]}^j(q, a) = z_{[k^*]}^j(q, \xi)$ . We are also assured that  $z^j(q, \xi) \geq \xi$  because  $x$  is a feasible solution for  $z_k^j(q, \xi)$  for all  $k$ . If additionally  $x$  is optimal, then we would have  $z_{[k^*]}^j(q, \xi) = \xi_{[k^*]}$ . This allows us to carry on computing  $z_{k^*+1}^j(q, \xi)$  when  $k^* < m$ , acting acting as though  $a$  had been  $\xi$  the whole time. We now have tighter bounds in the sense that

$$z_{k^*+1}^j(q, a) \geq z_{k^*+1}^j(q, \xi) \geq v_{k^*+1}^j(q, \xi) = v_{k^*+1}^j(q, a) \geq \xi_{k^*+1}.$$

Theorem 4.3.7 also permits for some shortcuts. Item 4a tells us that, when  $k^* < m$ , even if  $\xi_{k^*} = a_{k^*}$ , we need not optimize eq. (4.10) for  $k = k^* + 1$  as long as  $\xi_{k^*+1} > a_{k^*+1}$ . That check only takes  $O(n)$  time to compute, whereas just one step of the revised simplex method takes  $O(n(p + b + k^* - 1))$  in the worst case (Bertsimas & Tsitsiklis, 1997, § 3.3). However, we might still want to optimize eq. (4.10) for  $k = k^* + 1$  so we can take advantage of item 4b. One heuristic for deciding whether to take a chance on optimizing eq. (4.10) for  $k = k^* + 1$  is being sure to do so if there is an integer feasible solution at hand to eq. (4.10) for  $k = k^*$ . Even without treating item 4a as a shortcut, it allows for checking solver software for numerical inaccuracies in computing the value of the objective function.

Another shortcut is available for moving from one subproblem to another. When a bid’s package  $Q_i$  is big enough and its value  $V_i$  is small enough, we can avoid calculating  $z^{j-1}(\phi, u)$  for  $\phi = q - Q_i$  and  $u = a - V_i$  if we have already computed  $z^j(q, a)$ . Given theorem 4.3.7, we would only be fathoming  $v^{j-1}(\phi, u)$  if  $z^j(q, a) \geq a$ .

*Proof of theorem 4.3.7.* A recurring but implicit theme in the sub-proofs below is the application to the AR constraints of lemma 4.1.1, which says that  $\geq$  implies  $\geq$ .

*Proof of item 1.* If  $\mathcal{W}^j(\mathbf{q}) = \emptyset$ , then  $v_1^j(\mathbf{q}) = -\infty \leq z_1^j(\mathbf{q}, \mathbf{a})$ . If  $x \in \mathcal{W}^j(\mathbf{q})$ , then  $x \in \mathcal{E}^j \subseteq \mathcal{P}^j$  and  $\mathbf{Q}x \leq \mathbf{q}$ , so  $x$  is feasible for eq. (4.10) when  $k = 1$ , so  $z_1^j(\mathbf{q}, \mathbf{a}) \geq \mathbf{e}_1^\top \mathbf{V}x + \mathbf{r}^\top(\mathbf{q} - \mathbf{Q}x) = v_1^j(\mathbf{q})$ .  $\lrcorner$

*Proof of item 2.* If  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}) = \emptyset$ , then  $v^j(\mathbf{q}, \mathbf{a}) = -\infty$ , but  $-\infty \leq z^j(\mathbf{q}, \mathbf{a})$ .

Suppose  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}) \neq \emptyset$ . Then  $v^j(\mathbf{q}, \mathbf{a}) \geq \mathbf{a}$  by lemma 4.1.7. In particular,  $v_1^j(\mathbf{q}, \mathbf{a}) \geq a_1$ . Since  $\hat{\zeta}^j$  minimizes over the closed set of matrices  $\Delta, \Gamma$  with nonnegative real entries, there exist such matrices for which  $\zeta^j(\Delta, \Gamma) = \hat{\zeta}^j = z^j(\mathbf{q}, \mathbf{a})$  by theorem 4.3.5. By lemma 4.3.4,  $v_1^j(\mathbf{q}, \mathbf{a}) \leq \zeta_1^j(\Delta, \Gamma) = z_1^j(\mathbf{q}, \mathbf{a})$ . If  $k^* > 1$ , then we have equality across:  $a_1 = v_1^j(\mathbf{q}, \mathbf{a}) = z_1^j(\mathbf{q}, \mathbf{a})$ . Indeed, suppose  $k^* > 1$ , and, by way of induction on  $k^*$ , that  $\mathbf{a}_{[k^*-1]} = v_{[k^*-1]}^j(\mathbf{q}, \mathbf{a}) = z_{[k^*-1]}^j(\mathbf{q}, \mathbf{a})$ . Then the same reasoning as for  $k^* = 1$  dictates for  $k^* > 1$  that  $a_{k^*} \leq v_{k^*}^j(\mathbf{q}, \mathbf{a}) \leq z_{k^*}^j(\mathbf{q}, \mathbf{a})$ . If  $a_{k^*} = z_{k^*}^j(\mathbf{q}, \mathbf{a})$ , then, again, there's equality across, thereby completing the induction. Regardless of equality at  $k^*$ , if  $k^* < m$ , then, again by lemma 4.3.4 and theorem 4.3.5,  $v_{k^*+1}^j(\mathbf{q}, \mathbf{a}) \leq z_{k^*+1}^j(\mathbf{q}, \mathbf{a})$ .

However, if  $z_{k^*}^j(\mathbf{q}, \mathbf{a}) < a_{k^*}$ , then the inequality above is a contradiction, so  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}) = \emptyset$ . Moreover, no  $x \in \mathcal{P}^j$  such that  $\mathbf{Q}x \leq \mathbf{q}$  can satisfy  $(\mathbf{V}x + \mathbf{R}(\mathbf{q} - \mathbf{Q}x))_{[k-1]} \geq \mathbf{a}_{[k-1]}$  for  $k > k^*$ , so eq. (4.10) is infeasible for  $k > k^*$ .  $\lrcorner$

*Proof of item 3.* If  $k^* > 1$  and  $a_1 > -\infty$ , then  $z_1^j(\mathbf{q}, \mathbf{a}) = a_1 > -\infty$ . Supposing inductively that  $z_k^j(\mathbf{q}, \mathbf{a}) > -\infty$  for  $k \in [k^* - 1]$ , there exists a feasible solution  $x$  satisfying the constraints of eq. (4.10) for  $k = k^*$ . Hence  $z_{k^*}^j(\mathbf{q}, \mathbf{a}) > -\infty$ . Abbreviate the objective function's value  $\xi := \mathbf{V}x + \mathbf{R}(\mathbf{q} - \mathbf{Q}x)$ .

Further,  $x$  is feasible for eq. (4.10) for  $k \in [k^*]$ , so  $\xi_{[k^*]} \leq z_{[k^*]}^j(\mathbf{q}, \mathbf{a})$  because  $\mathbf{a}_{[k^*-1]} \leq \xi_{[k^*-1]}$ . Hence  $\xi_{[k^*-1]} = z_{[k^*-1]}^j(\mathbf{q}, \mathbf{a})$ .  $\lrcorner$

*Proof of item 4.* Suppose that  $x \in \mathcal{P}^j$  such that  $\mathbf{Q}x \leq \mathbf{q}$  and  $\xi_{[k^*]} \geq \mathbf{a}_{[k^*]}$ .

If  $k^* < m$ , then  $x$  is a feasible solution to eq. (4.10) for  $k = k^* + 1$ . Hence  $\xi_{k^*+1} \leq z_{k^*+1}^j(\mathbf{q}, \mathbf{a})$ .

Now suppose  $x$  is integral, which implies that  $x \in \mathcal{E}^j$ .

Combining  $\xi_{[k^*]} \geq \mathbf{a}_{[k^*]}$  with either  $\xi_{k^*} > a_{k^*}$  or  $k^* = m$  yields  $\xi \geq \mathbf{a}$ . Therefore  $x \in \mathcal{X}^j(\mathbf{q}, \mathbf{a})$ . Hence  $\xi \leq v^j(\mathbf{q}, \mathbf{a})$ , so, by lemma 4.1.7,  $\mathcal{W}^j(\mathbf{q}, \xi) = \mathcal{W}^j(\mathbf{q}, \mathbf{a})$ .

Further, suppose  $k^* = m$  and  $x$  is optimal. From item 2 we have that  $v^j(q, a) \leq z^j(q, a)$ , and hence  $v^j(q, a) \leq z^j(q, a)$ . From item 3 and the optimality of  $x$ , we have that  $\xi = z^j(q, a)$ . Putting this together with the facts that  $\xi \leq v^j(q, a)$  and  $x \in \mathcal{X}^j(q, a)$ , we conclude that  $x \in \mathcal{W}^j(q, a)$ .  $\square$

**4.3.3 Reusing Relaxations.** Under certain circumstances, we can reuse solutions to LP relaxations we computed for one subproblem when we descend the search tree to another subproblem. The previous subsection suggests an algorithm for fathoming the subproblem with current bidder  $j$ , residual supply  $q$ , and aggregate reserve  $a$ . The algorithm involves computing the LP relaxation in eq. (4.10) for  $k = 1$  until the LP relaxation's optimal value no longer equals the  $k$ th entry of  $a$ . Theorem 4.3.7 denotes by  $k^*$  that last  $k$ . Proposition 4.3.8 below suggests hanging onto certain functions of the  $k^*$  maximizers  $x$  of those  $k^*$  LPs. They serve as pruning bounds if we descend the search tree from bidder  $j$  to bidder  $j - 1$  at a subproblem with too large an aggregate reserve  $u$  and too small a residual supply  $\phi$ . The proposition may allow us to prune the new search-tree node before we even compute its LP relaxation for the cost of a few vector-matrix multiplications and storing the resulting  $k^*$  floating-point numbers and  $p$  integers.

Since the goal of proposition 4.3.8 is to tell us something about a subproblem when the current bidder is  $j - 1$  based on data gathered about a subproblem when the current bidder was  $j$ , we need notation for removing bidder  $j$  from a solution. Let  $P \in \mathbb{R}^{n \times n}$  be the linear projection that zeros out all entries except those indexed by  $I^j$ :

$$P := \sum_{i \in I^j} e_i e_i^\top.$$

If  $I$  is the  $n \times n$  identity matrix, then  $I - P$  zeros out just the entries indexed by  $I^j$ , so that if  $x \in \mathcal{P}^j$ , then  $(I - P)x \in \mathcal{P}^{j-1}$ .

**Proposition 4.3.8.** *Suppose  $j \in [b]$ . Let  $k^*$  be the same as given in theorem 4.3.7;  $x$  be a maximizer of eq. (4.10) for  $k = k^*$ ;  $u \in \overline{\mathbb{R}}^m$  such that  $u_{[k^*]} > (V(I - P)x + R(q - Qx))_{[k^*]}$ ; and  $\phi \in \mathbb{N}^p$  such that  $\phi \leq q - QPx$ . Then  $\mathcal{X}^{j-1}(\phi, u) = \emptyset$ .*

*Proof.* Our goal is to show that

$$\mathbf{u} > \mathbf{z}^{j-1}(\phi, \mathbf{u}), \quad (4.15)$$

which, by item 2c of theorem 4.3.7, proves that  $\mathcal{X}^{j-1}(\phi, \mathbf{u}) = \emptyset$ .

We will limit ourselves to the first few entries of these vectors, as follows. From the hypothesis that  $\mathbf{u}_{[k^*]} > (\mathbf{V}(\mathbf{I} - \mathbf{P})\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}))_{[k^*]}$ , we know there is a  $\tilde{k} \in [k^*]$  such that

$$\mathbf{u}_{[\tilde{k}-1]} = (\mathbf{V}(\mathbf{I} - \mathbf{P})\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}))_{[\tilde{k}-1]} \quad (4.16)$$

and  $u_{\tilde{k}} > e_{\tilde{k}}^\top (\mathbf{V}(\mathbf{I} - \mathbf{P})\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}))$ . Put differently,  $\tilde{k}$  is the least for which

$$\mathbf{u}_{[\tilde{k}]} > (\mathbf{V}(\mathbf{I} - \mathbf{P})\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}))_{[\tilde{k}]}. \quad (4.17)$$

Theorem 4.3.7's item 3 and the fact that  $\mathbf{x}$  is a maximizer of eq. (4.10) for  $k = k^* \geq \tilde{k}$  combine to show the second equality below:

$$\begin{aligned} (\mathbf{V}(\mathbf{I} - \mathbf{P})\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}))_{[\tilde{k}]} &= (\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}))_{[\tilde{k}]} - (\mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}]} \\ &= \mathbf{z}_{[\tilde{k}]}^j(\mathbf{q}, \mathbf{a}) - (\mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}]} \\ &= (\mathbf{z}^j(\mathbf{q}, \mathbf{a}) - \mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}]}. \end{aligned} \quad (4.18)$$

We claim, and will prove at the end, that

$$(\mathbf{z}^j(\mathbf{q}, \mathbf{a}) - \mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}]} \geq \mathbf{z}_{[\tilde{k}]}^{j-1}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}, \mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x}). \quad (4.19)$$

Moreover, by the definition of  $k^*$  in theorem 4.3.7 and the fact that  $\tilde{k} \leq k^*$ , we also have

$$\begin{aligned} \mathbf{z}_{[\tilde{k}-1]}^j(\mathbf{q}, \mathbf{a}) - (\mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}-1]} &= \mathbf{a}_{[\tilde{k}-1]} - (\mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}-1]} \\ &= (\mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}-1]}, \end{aligned}$$

which, in combination with eqs. (4.16) and (4.18), implies

$$(\mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x})_{[\tilde{k}-1]} = \mathbf{u}_{[\tilde{k}-1]}.$$

From that equation and our assumptions that  $\phi \leq \mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}$ , we see that relaxing the supply constraints in eq. (4.10) yields

$$\mathbf{z}_{[\tilde{k}]}^{j-1}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}, \mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x}) \geq \mathbf{z}_{[\tilde{k}]}^{j-1}(\phi, \mathbf{u}). \quad (4.20)$$

Combining eqs. (4.17) to (4.20) achieves eq. (4.15), our goal. (Don't forget that  $\leq$  implies  $\leq$  by lemma 4.1.1).

*Proof of eq. (4.19).* Suppose there is a  $\mathbf{y} \in \mathcal{P}^{j-1}$  such that  $\mathbf{Q}\mathbf{y} \leq \mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}$ . If there isn't, the right-hand side of eq. (4.19) is  $-\infty$  and we're done. Let  $\hat{k} \in [\tilde{k}]$  be the largest for which we can furthermore make such a  $\mathbf{y}$  satisfy

$$(\mathbf{V}\mathbf{y} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x} - \mathbf{Q}\mathbf{y}))_{[\hat{k}-1]} \geq (\mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x})_{[\hat{k}-1]}. \quad (4.21)$$

This way  $\mathbf{y}$  is a feasible decision variable for  $z_k^{j-1}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}, \mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x})$  for each  $k \in [\hat{k}]$ , so

$$z_{[\hat{k}]}^{j-1}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}, \mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x}) \geq (\mathbf{V}\mathbf{y} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x} - \mathbf{Q}\mathbf{y}))_{[\hat{k}]}. \quad (4.22)$$

Because  $z_k^{j-1}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}, \mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x}) = -\infty$  for each  $k \in \{\hat{k} + 1, \dots, \tilde{k}\}$ , it remains only to prove eq. (4.19) with  $\hat{k}$  replacing  $\tilde{k}$ .

We will now show that  $\hat{\mathbf{y}} := \mathbf{y} + \mathbf{P}\mathbf{x}$  is a feasible decision variable in eq. (4.10) for each  $k \in [\hat{k}]$ . Since  $j > 0$  we have  $\mathbf{e}_{[j]}^\top \mathbf{x} = 1$ , and since  $\hat{\mathbf{y}} \geq \mathbf{0}$ , we have

$$\mathbf{E}\hat{\mathbf{y}} = \mathbf{e}_{[j-1]} + \mathbf{E}\mathbf{P}\mathbf{x} = \mathbf{e}_{[j-1]} + \mathbf{e}_j = \mathbf{e}_{[j]},$$

so  $\hat{\mathbf{y}} \in \mathcal{P}^j$ .  $\hat{\mathbf{y}}$  satisfies the rest of the constraints:

$$\begin{aligned} \mathbf{Q}\hat{\mathbf{y}} &= \mathbf{Q}\mathbf{y} + \mathbf{Q}\mathbf{P}\mathbf{x} & (\mathbf{V}\hat{\mathbf{y}} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\hat{\mathbf{y}}))_{[\hat{k}-1]} \\ &\leq \mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x} + \mathbf{Q}\mathbf{P}\mathbf{x} & = (\mathbf{V}\mathbf{y} + \mathbf{V}\mathbf{P}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{y} - \mathbf{Q}\mathbf{P}\mathbf{x}))_{[\hat{k}-1]} \\ &= \mathbf{q}; & \geq (\mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x} + \mathbf{V}\mathbf{P}\mathbf{x})_{[\hat{k}-1]} \\ & & = \mathbf{a}_{[\hat{k}-1]}, \end{aligned}$$

where the inequality on the right comes from eq. (4.21).  $\hat{\mathbf{y}}$ 's consequent feasibility implies

$$\begin{aligned} z_{[\hat{k}]}^j(\mathbf{q}, \mathbf{a}) &\geq (\mathbf{V}\hat{\mathbf{y}} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\hat{\mathbf{y}}))_{[\hat{k}]} \\ &= (\mathbf{V}\mathbf{y} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x} - \mathbf{Q}\mathbf{y}))_{[\hat{k}]} + (\mathbf{V}\mathbf{P}\mathbf{x})_{[\hat{k}]}. \end{aligned}$$

By choosing the best possible  $\mathbf{y}$  (possibly a different  $\mathbf{y}$  for each value of  $k$ ), we can conclude that, for each  $k \in [\hat{k}]$ ,

$$z^j(\mathbf{q}, \mathbf{a})_{[\hat{k}]} \geq z_{[\hat{k}]}^{j-1}(\mathbf{q} - \mathbf{Q}\mathbf{P}\mathbf{x}, \mathbf{a} - \mathbf{V}\mathbf{P}\mathbf{x}) + (\mathbf{V}\mathbf{P}\mathbf{x})_{[\hat{k}]}. \quad \square$$

In practice, after computing  $z_{[k^*]}^j(\mathbf{q}, \mathbf{a})$ , we need to store  $k^*$ ,  $(V(I - P)\mathbf{x})_{[k^*]}$ , and  $\mathbf{q} - QP\mathbf{x}$ . For convenience, let  $k_b := k^*$ ,  $\mathbf{b} := V(I - P)\mathbf{x}$ , and  $\mathbf{q}_b := \mathbf{q} - QP\mathbf{x}$ . When can we replace the stored data  $k_b, \mathbf{b}, \mathbf{q}_b$  with new data  $k_d \in [m]$ ,  $\mathbf{d} \in \mathbb{R}^m$ , and  $\mathbf{q}_d \in \mathbb{R}^p$ ? It is always safe when  $u_{[k_b]} > \mathbf{b}_{[k_b]}$  implies  $u_{[k_b]} > \mathbf{b}_{[k_b]}$  and  $\phi \leq \mathbf{q}_b$  implies  $\phi \leq \mathbf{q}_b$  for all  $\mathbf{u} \in \overline{\mathbb{R}}^m$  and  $\phi \in \mathbb{N}^p$ . This occurs exactly when  $k_b \leq k_d$ ,  $\mathbf{b}_{[k_b]} \geq \mathbf{d}_{[k_b]}$ , and  $\mathbf{q}_b \leq \mathbf{q}_d$ .

#### 4.4 Dynamic Programming

This section develops the elements of a *dynamic programming* algorithm for  $\text{WDP}^\times$ . The development comprises subsections 4.4.1 and 4.4.2's recursive description of  $\text{WDP}^\times$  in theorem 4.4.4 as well as subsection 4.4.3's memoization scheme to avoid computing subproblems more than once. While Sniedovich (1992), a monograph focused on the mathematical theory of DP (see also Minoux, 1983/1986, § 9 on pp. 375 sqq.), averred that there was no consensus on just what constitutes DP, we use the term here in the operational sense of Cormen et al. (2009, chap. 15 on pp. 359–413): DP starts with an optimization problem that has *optimal substructure* in that the optimal values of the independent, overlapping subproblems appear in a recursive description of the problem's optimum, and the subproblems admit a scheme to avoid repeating computations of them. *Independent* means that choosing a solution to one subproblem does not affect the feasible sets for other subproblems. *Overlapping* means that solving two subproblems requires computing the same, third subproblem.

Dynamic programming fits into the B&B structure we have been developing. Subsection 4.1.4 defined branching on bids by bidder, and section 4.3 extended bounding with LP relaxations to  $\text{WDP}^\times$ 's lexicographic maximization. Minoux (1983/1986, § 9.3.2 on pp. 395–399) discussed combining DP and B&B via the example of the longest-path problem. The principal is the same as other B&B techniques. Before the full optimization at a node of the search tree, quickly compute an upper bound. In DP, the recursion tree determines the search tree, and the full optimization involves the enumeration of the node's children. Theorem 4.4.4's eq. (4.28) explicates how that enumeration works for  $\text{WDP}^\times$  with branching on bids by bidder. (In particular, DP maximization algorithms are always equivalent to the longest-path problem in a certain supergraph of the recursion tree (Minoux, 1983/1986,



p. 391).)

*Remark 4.4.1.* Thus the full algorithm for computing  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}^j)$  is, more or less, as follows.

1. Check the memo per lemma 4.4.12.
2. If necessary, check the LP relaxation per theorem 4.3.7.
3. If necessary, perform a full optimization per theorem 4.4.4.
4. Record the result into the memo per remark 4.4.11.

While other techniques, such as cutting planes (for a discussion of branch-and-cut algorithms for WDP, see Sandholm, 2006, pp. 348–349), fit into the B&B structure as well, we focus on DP for a few reasons. First it streamlines algebraic reasoning about  $\text{WDP}^\times$ . Theorem 4.4.7 and proposition 4.4.8 offer results generic to  $\text{WDP}^\times$  (not just a particular algorithm) with simple proofs using theorem 4.4.4’s recursive formula. Second the DP memos can speed up certain popular pricing computations, the topic of section 4.5, in particular corollary 4.5.2.

Third DP has fewer opportunities for numerical problems than LP-based algorithms such as the B&B algorithm using LP relaxations in Nemhauser and Wolsey (1999, § II.4.2 on pp. 355–366). The simplex algorithm for solving LPS can have bad numerical properties (Ogryczak, 1988). Governments auctioning public assets and bidders spending tens of billions of dollars have little tolerance for rounding errors. For instance, in the UK’s 2013 auction of 4G spectrum, the auctioneer Ofcom contracted with outside auditors to compare the results of three independent implementations of WDP (Smith Institute, 2013, February 14/2013). (This is why we are not discussing probabilistic or approximation algorithms (Bertsimas & Tsitsiklis, 1997, § 11.5 on pp. 507 sqq.).) In a DP algorithm by contrast testing prospective solutions’ feasibility can occur in exact, integer rather than floating-point (FP) arithmetic. The one place a DP algorithm for  $\text{WDP}^\times$  needs to use FP arithmetic is summing up values vectors  $V_i$  and comparing them lexicographically. Assuming the FP implementation is compatible with the ubiquitous IEEE-754 standard, rounding from a real number to an FP number is monotonic, and addition and subtraction are exactly rounded (Higham, 1996/

2002, pp. 38, 41), so there is no opportunity for numerical problems in maximizing on the first values-matrix row. However, more research is needed on lexicographic comparisons of FP vectors. Recall from subsection 4.1.1 that for two real  $m$ -vectors,  $\alpha < \beta$  when  $\alpha_k = \beta_k$  for each  $k < k^*$  for some  $k^* \in [m]$  and  $\alpha_{k^*} < \beta_{k^*}$ . Monotonicity of FP arithmetic does not guarantee that the first  $k^* - 1$  entries of these vectors continue to equal each other exactly after rounding to FP numbers. Indeed standard textbooks and references on numerical analysis and FP systems, such as Goldberg (1991), Higham (1996/2002), Muller et al. (2010), and Trefethen and Bau (1997), virtually ignore the topic of FP comparison.

We briefly mention some literature on DP for WDP. Maldoom (2007) presented a full DP algorithm for multiunit WDP. Müller (2006) reviewed recursive formulations for a variety of classes of WDP instances chosen for computationally tractability, except for Thm. 13.7 on p. 333, which regards generic, multiunit WDP. That theorem generalizes results from two other studies of multiunit WDP. Tennenholtz (2000, p. 102) looked at a version of multiunit WDP through the lens of keeping the number of products  $p$  fixed and analyzed the problem's computational tractability using a certain, recursively defined graph. Van Hoesel and Müller (2001, p. 29) also kept  $p$  fixed, showing specifically for  $p = 1$  that an optimal DP algorithm can have a running time in  $O(bs)$ , where  $s$  is the supply of the one product; cf. theorem 4.4.7 on page 168.<sup>38</sup>

**4.4.1 From Indicators to Indices.** So far we have represented allocations of products to bidders as indicator vectors  $x$  of winning bids. In subsection 4.4.2 eq. (4.28)'s recursive formulation of eq. (4.4)'s WDP<sup>⊆</sup> subproblem is in terms of winning bids' indices  $i$  in the columns of  $V$  and  $Q$ . This subsection maps between the two representations. We assume  $j \neq 0$  because the auctioneer submits no bids. Lemma 4.4.2 below provides for the existence of the indices given an indicator vector. We will use it like "given a feasible allocation  $x \in \mathcal{X}^j(q, a, I)$ , select bidder  $j$ 's winning bid index  $i$  according lemma 4.4.2".

**Lemma 4.4.2.** *Suppose  $j \neq 0$  and  $x$  obeys the nonnegativity, integrality, and  $j$ th XOR constraints:  $x \in \mathcal{E}^j$ . For each bidder  $\beta \in [j]$ , there is a unique **winning bid index**  $i_\beta \in I^\beta$  for which  $x_{i_\beta} = 1$ ,*

---

<sup>38</sup>↑By theorem 4.4.7, assuming away the asymptotics of  $p$  assumes away the exponential upper bound on DP's running time for WDP.

and that  $i_\beta$  is the unique element of the set  $\arg \max_{i \in \mathcal{I}^\beta} x_i$ . Moreover, if  $\mathbf{x}$  obeys the subset- $\mathcal{I}$  constraint, then  $i_j \in \mathcal{I}$ .

*Proof.* The result follows directly from the definitions of the respective constraints.  $\square$

Next, lemma 4.4.3 provides for the optimality of an index for the current bidder  $j$  given an indicator vector representation of an allocation for bidder  $j - 1$ 's subproblem. This inductive result sets up the proof of theorem 4.4.4's recursive formula for the optimal value  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  of the  $\text{wDP}^\times$  subproblem in eq. (4.4). In particular, the lemma is one way of describing  $\text{wDP}^\times$ 's optimal substructure via subsection 4.1.4's branch-on-bids-by-bidder scheme. Let  $\mathbf{e}_1, \dots, \mathbf{e}_n$  denote the standard basis of  $\mathbb{R}^n$ .

**Lemma 4.4.3.** *Suppose that  $j \neq 0$  and that  $\mathbf{x}$  is a winning allocation for the subproblem in eq. (4.4), i.e.,  $\mathbf{x} \in \mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ . Per lemma 4.4.2 select bidder  $j$ 's unique winning bid index  $i \in \mathcal{I}$ , for which  $x_i = 1$ . Then  $\mathbf{Q}_i \leq \mathbf{q}$  and there is an  $\mathbf{x}' \in \mathcal{W}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - \mathbf{V}_i, \mathcal{I}^{j-1})$  such that*

$$\mathbf{x} - \mathbf{e}_i \in \mathcal{W}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - \mathbf{V}_i, \mathcal{I}^{j-1}), \quad (4.23)$$

$$\mathbf{x}' + \mathbf{e}_i \in \mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}). \quad (4.24)$$

*Proof.* The proofs of eqs. (4.23) and (4.24) are very similar, but eq. (4.23) itself is the reason we can know that  $\mathbf{x}' \in \mathcal{W}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - \mathbf{V}_i, \mathcal{I}^{j-1})$  is even possible. To avoid repeating several lengthy equations, we cheat a little by assuming that  $\mathbf{Q}_i \leq \mathbf{q}$  and that  $\mathbf{x}'$  exists so that we can prove both equations in parallel. The proof of eq. (4.23) does not depend on  $\mathbf{x}'$  at all, so it will fully justify  $\mathbf{x}'$ 's existence. Along the way it will also prove the technical condition that  $\mathbf{Q}_i \leq \mathbf{q}$ .

First we show feasibility,  $\mathbf{x} - \mathbf{e}_i \in \mathcal{X}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - \mathbf{V}_i)$  and  $\mathbf{x}' + \mathbf{e}_i \in \mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ , starting with the respective AR, supply, and XOR constraints on  $\mathbf{x}$  and  $\mathbf{x}'$ . In the following list of implications, the left column presents facts arising from  $\mathbf{x} \in \mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  and  $\mathbf{x}' \in \mathcal{X}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - \mathbf{V}_i)$ . Equalities trailing on from there follow by adding the zero vector in the form of  $-\mathbf{V}_i + \mathbf{V}_i$ ,  $-\mathbf{Q}_i + \mathbf{Q}_i$ , or  $-\mathbf{e}_j + \mathbf{e}_j$  (where  $\mathbf{e}_j$ 's length is  $b$  and  $\mathbf{e}_i$ 's is  $n$ ) and then rearranging using  $\mathbf{V}\mathbf{e}_i = \mathbf{V}_i$ ,  $\mathbf{Q}\mathbf{e}_i = \mathbf{Q}_i$ , and, since  $i \in \mathcal{I}^j$ ,  $\mathbf{E}\mathbf{e}_i = \mathbf{e}_j$ . The right column justifies the set memberships that we want to show.

Aggregate-reserve constraint

$$\begin{aligned} \mathbf{a} &\leq \mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) \\ &= \mathbf{V}(\mathbf{x} - \mathbf{e}_i) + \mathbf{V}_i &\implies & \mathbf{a} - \mathbf{V}_i \leq \mathbf{V}(\mathbf{x} - \mathbf{e}_i) \\ &\quad + \mathbf{R}((\mathbf{q} - \mathbf{Q}_i) - \mathbf{Q}(\mathbf{x} - \mathbf{e}_i)) && \quad + \mathbf{R}((\mathbf{q} - \mathbf{Q}_i) - \mathbf{Q}(\mathbf{x} - \mathbf{e}_i)) \end{aligned}$$

$$\begin{aligned} \mathbf{a} - \mathbf{V}_i &\leq \mathbf{V}\mathbf{x}' + \mathbf{R}((\mathbf{q} - \mathbf{Q}_i) - \mathbf{Q}\mathbf{x}') \\ &= \mathbf{V}\mathbf{x}' + \mathbf{R}(\mathbf{q} - \mathbf{Q}(\mathbf{x}' + \mathbf{e}_i)) &\implies & \mathbf{a} \leq \mathbf{V}(\mathbf{x}' + \mathbf{e}_i) + \mathbf{R}(\mathbf{q} - \mathbf{Q}(\mathbf{x}' + \mathbf{e}_i)) \end{aligned}$$

Supply constraints

$$\mathbf{q} \geq \mathbf{Q}\mathbf{x} = \mathbf{Q}(\mathbf{x} - \mathbf{e}_i) + \mathbf{Q}_i \implies \mathbf{q} - \mathbf{Q}_i \geq \mathbf{Q}(\mathbf{x} - \mathbf{e}_i) \quad (4.25)$$

$$\mathbf{q} - \mathbf{Q}_i \geq \mathbf{Q}\mathbf{x}' \implies \mathbf{q} \geq \mathbf{Q}(\mathbf{x}' + \mathbf{e}_i)$$

Exclusive-or constraints

$$\mathbf{e}_{[j]} = \mathbf{E}\mathbf{x} = \mathbf{E}(\mathbf{x} - \mathbf{e}_i) + \mathbf{e}_j \implies \mathbf{E}(\mathbf{x} - \mathbf{e}_i) = \mathbf{e}_{[j]} - \mathbf{e}_j = \mathbf{e}_{[j-1]}$$

$$\mathbf{E}\mathbf{x}' = \mathbf{e}_{[j-1]} = \mathbf{e}_{[j]} - \mathbf{e}_j \implies \mathbf{e}_{[j]} = \mathbf{E}\mathbf{x}' + \mathbf{e}_j = \mathbf{E}(\mathbf{x}' + \mathbf{e}_i)$$

Next we consider the nonnegativity and integrality constraints.  $\mathbf{x}$ ,  $\mathbf{x}'$ , and  $\mathbf{e}_i$  are each in  $\{0, 1\}^n$ , and  $\mathbf{e}_i$  is zero everywhere but the  $i$ th coordinate. We specifically chose  $i$  so that  $x_i = 1$ , hence  $\mathbf{x} - \mathbf{e}_i \in \{0, 1\}^n$ . So is  $\mathbf{x}' + \mathbf{e}_i$  as long as  $x'_i = 0$ , which it is because  $i \in \mathcal{I} \subseteq \mathcal{I}^j$  but  $\mathbf{E}\mathbf{x}' = \mathbf{e}_{[j-1]}$  implies that  $\mathbf{e}_{\mathcal{I}^j}^\top \mathbf{x}' = \sum_{\ell \in \mathcal{I}^j} x'_\ell = 0$  even though  $\mathbf{x}' \geq \mathbf{0}$ . We now have enough to conclude that  $\mathbf{q} \geq \mathbf{Q}_i$ : The right side of eq. (4.25) and the facts that  $\mathbf{Q}$  and  $\mathbf{x} - \mathbf{e}_i$  are nonnegative imply  $\mathbf{q} - \mathbf{Q}_i \geq \mathbf{0}$ .

The last elements of feasibility are the respective subset constraints. We saw above that  $\mathbf{x} - \mathbf{e}_i$  satisfies the  $(j-1)$ th XOR constraint, and that in turn implies that it satisfies the subset- $\mathcal{I}^{j-1}$  constraint. Consequently  $\mathbf{x} - \mathbf{e}_i \in \mathcal{X}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - \mathbf{V}_i, \mathcal{I}^{j-1})$ , which justifies the existence of  $\mathbf{x}'$ . Per the previous paragraph  $0 = \mathbf{e}_{\mathcal{I}^j}^\top \mathbf{x}' \geq \mathbf{e}_{\mathcal{I}}^\top \mathbf{x}' \geq 0$  because  $\mathcal{I} \subseteq \mathcal{I}^j$ , and  $\mathbf{e}_{\mathcal{I}}^\top \mathbf{e}_i = 1$  because  $i \in \mathcal{I}$ , so  $\mathbf{e}_{\mathcal{I}}^\top (\mathbf{x}' + \mathbf{e}_i) = 1 = \min\{1, j\}$ . Consequently  $\mathbf{x}' + \mathbf{e}_i \in \mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ .

From that feasibility consequence and the fact that  $\mathbf{x} \in \mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  we have that  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  equals the left side of

$$\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) \geq \mathbf{V}(\mathbf{x}' + \mathbf{e}_i) + \mathbf{R}(\mathbf{q} - \mathbf{Q}(\mathbf{x}' + \mathbf{e}_i)). \quad (4.26)$$

Rearrange this using  $\mathbf{q} - \mathbf{Q}\mathbf{x} = (\mathbf{q} - \mathbf{Q}_i) - \mathbf{Q}(\mathbf{x} - \mathbf{e}_i)$  and  $\mathbf{q} - \mathbf{Q}(\mathbf{x}' + \mathbf{e}_i) = (\mathbf{q} - \mathbf{Q}_i) - \mathbf{Q}\mathbf{x}'$  to obtain

$$V(\mathbf{x} - \mathbf{e}_i) + \mathbf{R}((\mathbf{q} - \mathbf{Q}_i) - \mathbf{Q}(\mathbf{x} - \mathbf{e}_i)) \geq V\mathbf{x}' + \mathbf{R}((\mathbf{q} - \mathbf{Q}_i) - \mathbf{Q}\mathbf{x}'), \quad (4.27)$$

the right side of which equals  $v^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - V_i)$  because  $\mathbf{x}' \in \mathcal{W}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - V_i)$ . Consequently  $\mathbf{x} - \mathbf{e}_i \in \mathcal{W}^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - V_i)$ .

From that optimality consequence, the inequality in eq. (4.27) and hence also the one in eq. (4.26) are both equalities. The latter equality then implies that  $\mathbf{x}' + \mathbf{e}_i \in \mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ .  $\square$

**4.4.2 Recursion.** The main result of this subsection is theorem 4.4.4, which rewrites  $\text{WDP}^\times$  from a lexicographic, linear ZOIP as in eq. (4.3) to a recursive maximization problem. Our focus here remains on definition 4.1.5's subproblems, and the theorem's eq. (4.28) is in terms of the optimal-value function  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ . Let  $\mathbf{e}_1, \dots, \mathbf{e}_n$  denote the standard basis of  $\mathbb{R}^n$ .

**Theorem 4.4.4.** *The following recursive  $\text{WDP}^\times$  Bellman equation holds for the optimal value of the  $\text{WDP}^\times$  subproblem in eq. (4.4) (cf. Bellman's equation in the graph-shortest-path problem defined in Bertsimas & Tsitsiklis, 1997, § 7.9):*

$$\begin{aligned} v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = & \max_i v^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - V_i) + V_i & (4.28) \\ \text{subject to } & v^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - V_i) + V_i \geq \mathbf{a}, & (\text{aggregate reserve}) \\ & v^{j-1}(\mathbf{q} - \mathbf{Q}_i, \mathbf{a} - V_i) \neq -\infty, & (\text{recursive feasibility}) \\ & \mathbf{Q}_i \leq \mathbf{q}, & (\text{supply}) \\ & i \in \mathcal{I}, & (\text{subset } \mathcal{I}) \end{aligned}$$

for all residual supplies  $\mathbf{q} \leq \mathbf{s}$ , all aggregate reserves  $\mathbf{a} \in \overline{\mathbb{R}}^m$ , and all bid-index subsets  $\mathcal{I} \subseteq \mathcal{I}^j$  for the current bidder  $j \in [b]$ . When  $j = 0$ , eq. (4.7) on page 141 serves as the recursion's base case. Further, let  $\mathcal{B}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  denote the set of maximizers or **winning bid indices**  $i$  of eq. (4.28)'s right-hand side when  $j \neq 0$ .<sup>39</sup> Then  $\mathcal{B}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  equals the set of  $i \in \mathcal{I}$  for which  $x_i = 1$  for some  $\mathbf{x} \in \mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  via lemma 4.4.2.

<sup>39</sup>↑Formally,  $\mathcal{B}^j: \mathcal{D}^j \rightarrow 2^{\mathcal{I}^j}$  for each bidder  $j \in [b]$ .

*Remark 4.4.5.* Another way to read the last part of theorem 4.4.4 is that  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  equals the set of vectors  $e_{i_1} + \dots + e_{i_j}$  for any sequence of indices<sup>40</sup>

$$\begin{aligned} i_j &\in \mathcal{B}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}), \\ i_{j-1} &\in \mathcal{B}^{j-1}(\mathbf{q} - \mathbf{Q}_{i_j}, \mathbf{a} - \mathbf{V}_{i_j}, \mathcal{I}^{j-1}), \\ &\vdots \\ i_1 &\in \mathcal{B}^1(\mathbf{q} - \mathbf{Q}_{i_j} - \dots - \mathbf{Q}_{i_2}, \mathbf{a} - \mathbf{V}_{i_j} - \dots - \mathbf{V}_{i_2}, \mathcal{I}^1). \end{aligned} \quad (4.29)$$

*Proof of theorem 4.4.4.* Let  $\beta$  denote the maximum objective value on right side of eq. (4.28).

First we show that  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) \geq \beta$ . If  $\mathcal{B}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \emptyset$ , then our convention says that  $\beta = -\infty$ , so we're done. Suppose instead that  $i' \in \mathcal{B}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ . Equation (4.28)'s recursive-feasibility constraint implies that there is a winning allocation  $\mathbf{x}' \in \mathcal{W}^{j-1}(\mathbf{q} - \mathbf{Q}_{i'}, \mathbf{a} - \mathbf{V}_{i'})$ . Applying eq. (4.6) to get eq. (4.30) below, the wDP<sup>κ</sup> objective value for  $\mathbf{x} := \mathbf{x}' + e_{i'}$  is

$$\begin{aligned} \mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) &= \mathbf{V}(\mathbf{x}' + e_{i'}) + \mathbf{R}[\mathbf{q} - \mathbf{Q}(\mathbf{x}' + e_{i'})] = \mathbf{V}\mathbf{x}' + \mathbf{V}_{i'} + \mathbf{R}[(\mathbf{q} - \mathbf{Q}_{i'}) - \mathbf{Q}\mathbf{x}'] \\ &= v^{j-1}(\mathbf{q} - \mathbf{Q}_{i'}, \mathbf{a} - \mathbf{V}_{i'}) + \mathbf{V}_{i'} \quad (4.30) \\ &= \beta. \end{aligned}$$

Therefore we have to prove that  $\mathbf{x}$  is a feasible allocation for eq. (4.4) so that its wDP<sup>κ</sup> objective value is at most  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ , and thus so is  $\beta$ . Since  $i' \in \mathcal{B}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) \subseteq \mathcal{I} \subseteq \mathcal{I}^j$  but the current bidder for  $\mathbf{x}'$  is  $j-1$ , all the entries of  $\mathbf{x}' + e_{i'}$  are either zero or one. By the same token,  $\mathbf{x}$  satisfies the  $j$ th XOR and subset- $\mathcal{I}$  constraints. Because of eq. (4.30), eq. (4.4)'s and eq. (4.28)'s AR constraints are the same. Finally,  $\mathbf{x}$  satisfies the supply constraint because  $\mathbf{Q}\mathbf{x} = \mathbf{Q}\mathbf{x}' + \mathbf{Q}_{i'}$ , and our choice of  $\mathbf{x}'$  implies  $\mathbf{Q}\mathbf{x}' \leq \mathbf{q} - \mathbf{Q}_{i'}$ . We may now conclude that  $\mathbf{x}$  is a feasible allocation, so

$$\beta = \mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) \leq v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}). \quad (4.31)$$

Second, we show that  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) \leq \beta$ . If  $\mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = \emptyset$ , then our convention says that  $v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}) = -\infty$ , so we're done. Suppose instead that  $\hat{\mathbf{x}} \in \mathcal{W}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ . Lemma 4.4.2

---

<sup>40</sup>↑Equation (4.29)'s term  $\mathbf{V}_{i_j} + \dots + \mathbf{V}_{i_\ell}$ , for  $\ell \in \{2, \dots, j\}$ , is the  $m$ -dimensional analog of  $g$ , the "sum of prices of the bids accepted on the [search-tree] path", from Sandholm, 2006, Eq. 14.2 on p. 339.

implies that there is a unique winning bid index  $i \in \mathcal{I}$ , for which  $\hat{x}_i = 1$ . Applying eq. (4.6) and replacing  $x$  with  $\hat{x}$  in lemma 4.4.3's eq. (4.23) to get eq. (4.32) below, the Bellman-equation objective value for  $i$  is

$$\begin{aligned} v^{j-1}(q - Q_i, a - V_i) + V_i &= V(\hat{x} - e_i) + R[(q - Q_i) - Q(\hat{x} - e_i)] + V_i & (4.32) \\ &= V\hat{x} + R[q - Q\hat{x}] \\ &= v^j(q, a, \mathcal{I}). \end{aligned}$$

Therefore we have to prove that  $i$  is a feasible bid index for eq. (4.28) so that its Bellman-equation objective value is at most  $\beta$ , and thus so is  $v^j(q, a, \mathcal{I})$ . We already know that  $i \in \mathcal{I}$ . Lemma 4.4.3 tells us that  $i$  satisfies eq. (4.28)'s supply and recursive-feasibility constraints. Because of the latter, lemma 4.1.7 says that  $v^{j-1}(q - Q_i, a - V_i) \geq a - V_i$ , so eq. (4.28)'s AR constraint holds. We may now conclude that  $i$  is a feasible solution to the right-hand side of eq. (4.28), so

$$v^j(q, a, \mathcal{I}) = v^{j-1}(q - Q_i, a - V_i) + V_i \leq \beta. \quad (4.33)$$

Putting together eqs. (4.31) and (4.33) gives equality across: the wDP<sup>κ</sup> and Bellman-equation objective values of  $x$  and  $i$  both equal  $\beta = v^j(q, a, \mathcal{I})$ , which is the same thing as eq. (4.28). This means that  $x$  and  $i$  are not just feasible but optimal solutions for eqs. (4.4) and (4.28), i.e.,  $x \in \mathcal{W}^j(q, a, \mathcal{I})$  and  $i \in \mathcal{B}^j(q, a, \mathcal{I})$ .  $\square$

From eq. (4.28), we can derive a Bellman equation for  $v^j(q)$  with zero reserve prices ( $r = 0$ ) and no AR constraint ( $a = -\infty$ ) that looks almost exactly like the one in Maldoom (2007, Eq. (2) in § 4.1) for wDP (see also Müller, 2006, Thm. 13.7 on p. 333). For  $j \in [b]$ , we have

$$\begin{aligned} v^j(q) &= \max_i^{\check{}} v^{j-1}(q - Q_i) + V_i & (4.34) \\ &\text{subject to } Q_i \leq q \\ & i \in \mathcal{I}^j \end{aligned}$$

Equation (4.34) uses lexicographic maximization “ $\max^{\check{}}$ ” whereas Maldoom's Bellman equation used regular maximization “ $\max$ ” over just bid prices. Lexicographic maximization allows us to solve the wDP<sup>κ</sup> in eq. (4.3) not as a sequence of instances of wDP but

directly, taking advantage of the information between tie breakers  $k \in [m]$  available from a fathoming algorithm based on theorem 4.3.7. Maldoom (2007, § 4.4) provided an algorithm for extracting all tied outcomes. With eq. (4.34), there is no need: any maximal solution satisfying eq. (4.34) has already fulfilled all the tie-breaking rules.

Equation (4.28) finally gives full clarity to what *branching on bids by bidder* meant in subsection 4.1.4, so we turn back for a moment to that subsection's discussion of branching schemes. Branching on bids by bidder produces a search tree whose nodes (for subproblems whose solution sets we call)  $\mathcal{B}^j(\mathbf{q}, \mathbf{a}, \mathcal{I}^j)$  at the same distance from the root  $\mathcal{B}^b(\mathbf{s}, -\infty, \mathcal{I}^b)$  have the same current bidder  $j$ , and the edges coming out of a node correspond to each of bidder  $j$ 's packages  $Q_i \leq \mathbf{q}$ ,  $i \in \mathcal{I}^j$ . In contrast, branching on bids, which Sandholm (2006, p. 340) found to yield faster algorithms than branching on items, produces essentially the same search tree as the naive DP algorithm, e.g., Müller (2006, Thm. 13.7 on p. 333), for the *multidimensional knapsack problem* (MKP) in eq. (4.35) below (Minoux, 1983/1986, p. 379). Holte (2001) was among the first to notice that MKP can be reinterpreted as WDP.

$$\begin{aligned} & \underset{x}{\text{maximize}} \quad (\mathbf{V}x)_1 & (4.35) \\ & \text{subject to} \quad Qx \leq \mathbf{s}, \\ & \quad \quad \quad x \in \{0, 1\}^n. \end{aligned}$$

The MKP knows nothing of the *bidder* structure of WDP. At one extreme where there is only one bidder ( $b = 1$ ), the WDP in eq. (4.1) is the MKP. In any case the virtual-products canonicalization from subsection 4.1.3 can convert a  $b$ -bidder WDP into a single-bidder WDP (A large class of integer linear programs can be written as a MKP. See subsection 4.1.3 and Minoux, 1983/1986, p. 379). At the other extreme where there is one bidder per bid ( $b = n$ ), eq. (4.1) is either infeasible or has the trivial solution  $x^* = \mathbf{1}$ . Between those two extremes, bidders submit approximately the same number of bids, which is the same as saying that the maximum number of bids any one bidder has,  $\max_j |\mathcal{I}^j|$ , is approximately the average  $n/b$ ; and the number of bids per bidder approximately equals the number bidders, which is the same as saying that  $n/b \approx b$ .<sup>41</sup> In subsection 4.1.4 we mentioned that branching

<sup>41</sup>↑The number of bids per bidder's approximating the number of bids means  $n/b \approx b$ , which implies



strategies affect running time of algorithms not through the size of the search tree but through the *bounding* half of *branch and bound*. Under this intermediate regime, branching on bids by bidders in a B&B algorithm allows for computing the LP relaxation for multiple bids at once while still breaking up the state space. This leads us to the following conjecture.

**Conjecture 4.4.6.** *When restricted to  $WDP^\leq$  instances in which  $\max_j |I^j| \approx n/b \approx \sqrt{n}$ , B&B algorithms that branch on bids by bidder are faster asymptotically as  $n \rightarrow \infty$  than those that branch on bids.*

Another asymptotic acceleration might be available by being careful about how we enumerate  $\{i \in I^j \mid Q_i \leq q\}$  in eq. (4.28). By lemma 4.1.1  $Q_i \leq q$  implies  $Q_i \leq q$ , so storing the packages in lexicographic order at least facilitates avoiding the inspection of packages lexicographically larger than  $q$  because we can stop iterating as soon as we see such a package. One data structure that stores strings—here, the positive (Matrices of packages tend in practice to be very sparse, so avoiding the storage of zeros may be profitable. A variety of methods for compressing sparse matrices exist, including the compressed-column representation described in Golub & Van Loan, 1983/2013, § 11.1.1 on pp. 598–599) entries of a package vector—in lexicographic order is the *radix tree* (Cormen et al., 2009, pp. 304–305). Suppose we index each bidder’s packages  $\{Q_i \mid i \in I^j\}$  with distinct radix trees whose edge labels correspond to the entries of  $Q_i$  and whose node labels correspond to  $i$ . Concatenating the edge labels of a path from the tree’s root ending at a node labeled  $i$  allows us to read off the entries of  $Q_i$ . Ordering the out-edges of each node ascending by edge label results in a lexicographic depth-first search order for the tree. Careful control of the tree’s traversal might yield an enumeration algorithm that minimizes how many columns of  $Q$  it examines beyond those that obey  $q$ ’s supply constraint. If such an algorithm inspects  $O(1)$  extra packages, then lemma 4.1.1 would enjoy an asymptotic speed up as  $n \rightarrow \infty$  compared to naively enumerating  $\{i \in I^j \mid Q_i \leq q\}$  and checking each package for the supply constraint.

Assuming some optimal enumeration scheme, we generalize to  $WDP^\leq$  from Müller

---

$n \approx b^2$ . Hence  $b \approx \sqrt{n} \approx n/b$ .

(2006, Thm. 13.7 on p. 333), which says that an optimal algorithm for classical WDP has an asymptotic running time of  $\mathcal{O}(bp(\|s\|_\infty + 1)^p \max_j |\mathcal{I}^j|)$ , where  $\|s\|_\infty := \max_t s_t$ , the maximum supply of any of the  $p$  different products. The proof follows Müller’s closely.

**Theorem 4.4.7.** *An optimal algorithm to solve the optimal value  $v^*$  of the WDP<sup>s</sup> in eq. (4.3) has an asymptotic running time in  $\mathcal{O}(b(p + m)(\|s\|_\infty + 1)^p \max_j |\mathcal{I}^j|)$ .*

*Proof.* We don’t need the AR constraint for now, so write  $v^j(q)$  in place of  $v^j(q, -\infty)$ . By theorem 4.4.4,  $v^* = v^b(s)$ , which we can compute by comparing and adding at most  $\max_j |\mathcal{I}^j|$  numbers of the form  $v_k^j(q_1, \dots, q_p)$  and subtracting  $p$  numbers (the  $q - Q_i$  term in eq. (4.28)). In that expression,  $k \in [m]$ ,  $j \in [b]$ ,  $q_1 \in \{0, \dots, s_1\}, \dots$ , and  $q_p \in \{0, \dots, s_p\}$ . However, the  $p$  subtractions need happen only once per expression  $v^j(q_1, \dots, q_p)$  for all  $k$  entries. That’s a total of  $b(p + m)(\max_j |\mathcal{I}^j|) \prod_{t=1}^p (s_t + 1) \leq b(p + m)(\|s\|_\infty + 1)^p (\max_j |\mathcal{I}^j|)$  numbers to compare and add in the worst case.  $\square$

Theorem 4.4.4 gives us the tools finally to prove that dominated bids (definition 4.2.1) are never necessary for a bidder to submit. Recall that  $(V_\alpha, Q_\alpha)$  dominates  $(V_\beta, Q_\beta)$  if and only if  $Q_\alpha \leq Q_\beta$  and  $V_\beta \leq V_\alpha$ , i.e., bid  $\alpha$  is for less stuff at a higher price than bid  $\beta$ .

**Proposition 4.4.8.** *Suppose  $j > 0$ . If  $(V_\alpha, Q_\alpha)$  dominates  $(V_\beta, Q_\beta)$  for some  $\alpha, \beta \in \mathcal{I}$  for some  $\mathcal{I} \subseteq \mathcal{I}^j$ , then  $\beta \in \mathcal{B}^j(q, a, \mathcal{I})$  only if  $\alpha \in \mathcal{B}^j(q, a, \mathcal{I})$ . Moreover, if  $V_\beta < V_\alpha$ , then  $\beta \notin \mathcal{B}^j(q, a, \mathcal{I})$ .*

*Remark 4.4.9.* If we modify eq. (4.3) to embody auction rules that disallow the winning allocation to leave any number of items unsold in each product, that is by replacing the “ $\leq$ ” in the supply constraints with an “ $=$ ”, proposition 4.4.8 could fail. The auctioneer might select a bid for a larger package at a lower value—preferring  $(V_\beta, Q_\beta)$  over  $(V_\alpha, Q_\alpha)$  in proposition 4.4.8—if doing so allowed the auction to clear with no unsold items.

*Proof of proposition 4.4.8.* Suppose  $\alpha, \beta \in \mathcal{I} \subseteq \mathcal{I}^j$ ,  $Q_\alpha \leq Q_\beta$  and  $V_\beta \leq V_\alpha$ . Suppose further that  $\beta$  satisfies the constraints of eq. (4.28):  $v^{j-1}(q - Q_\beta, a - V_\beta) + V_\beta \geq a$ ,  $v^{j-1}(q - Q_\beta, a - V_\beta) \neq -\infty$ , and  $Q_\beta \leq q$ . (Without this assumption, we would be done already with  $\beta \notin \mathcal{B}^j(q, a, \mathcal{I})$ .) Thus  $Q_\alpha \leq q$ .

Further,  $q - Q_\alpha \geq q - Q_\beta$  and  $a - V_\alpha \leq a - V_\beta$ . Thus subtracting the  $\alpha$  bid from the residual supply and aggregate reserve is a relaxation of the subproblem's supply and AR constraints compared with doing the same thing with  $\beta$ :

$$v^{j-1}(q - Q_\alpha, a - V_\alpha) \geq v^{j-1}(q - Q_\beta, a - V_\beta),$$

and thus

$$v^{j-1}(q - Q_\alpha, a - V_\alpha) + V_\alpha \geq v^{j-1}(q - Q_\beta, a - V_\beta) + V_\beta.$$

Therefore  $\alpha$  satisfies the constraints of eq. (4.28) and with at least as great an objective value as  $\beta$ 's bid. Hence bid  $\beta$  wins only if  $\alpha$  does too.

Now suppose further that  $V_\beta < V_\alpha$ . Then the inequality above becomes strict, and the objective value for bid  $\alpha$  strictly exceeds the objective value for bid  $\beta$ . Hence bid  $\beta$  cannot win.  $\square$

**4.4.3 Memoization.** To reduce the number of times we actually solve a subproblem recursively using eq. (4.28), we save the results of prior computations, including in the case of infeasibility. Maldoom (2007, § 4.1) proposed a DP algorithm based on the WDP Bellman eq. (4.34). That algorithm avoided repeating computations using the *bottom-up method* (Cormen et al., 2009, p. 365): by computing  $v^1(q)$  for all  $q \leq s$ , then  $v^2(q)$  for all  $q \leq s$ , etc., until reaching bidder  $b$  and computing  $v^b(s)$ . The other choice of method—ours here—in DP algorithm design is *top-down with memoization*: we check if we already have saved the result of the current subproblem and return it if so; otherwise we compute WDP's Bellman eq. (4.28) recursively as written, but save or *memoize* (or *cache* (Sandholm, 2006, p. 357)) the result before returning. While bottom-up DP algorithms tend to have better constant factors in their asymptotic running times, memoized top-down algorithms can be asymptotically faster if solving the problem's Bellman equation recursively typically does not require solving all subproblems (Cormen et al., 2009, pp. 365, 389).

Indeed, real WDP (and thus WDP<sup>⊆</sup>) instances tend to be quite sparse in this sense. To make this concrete, consider the set  $\{q - Q_i \mid i \in \mathcal{I}^j \text{ and } Q_i \leq q\}$  of residual supplies in the recursive calls to  $v^{j-1}(q - Q_i, a - V_i)$  in eq. (4.28) when  $j \in [b]$ . That set contains at most  $|\mathcal{I}^j|$  package vectors, the number of bids that bidder  $j$  submitted. It is typically much

smaller than eq. (4.28)'s *state space*, the set of all residual supplies  $\mathbf{q}'$  such that  $\mathbf{q}' \leq \mathbf{q}$ . The cardinality of the state space generally determines the practical computational difficulty of a dynamic programming problem (Minoux, 1983/1986, p. 378). In our case, the state space contains exactly  $\prod_{t=1}^p (q_t + 1)$  package vectors, which is on the order of  $2^p$ .<sup>42</sup> Indeed, real auctions usually limit by rule the number of bids a bidder may submit whereas  $2^p$  may be quite large. For example, the UK's 2013 auction of 4G spectrum had  $2^{39}$  possible packages for sale (Day & Cramton, 2012, p. 589), but the auctioneer permitted at most 4,000 bids per bidder (Ofcom, 2012, reg. 43 at para. 10). Canada's 2019 auction of the 600 MHz band offered bidders either  $5^{16}$  or  $8^{16}$  distinct packages (depending on the bidder) (Bono et al., 2019; ISED Canada, 2018, paras. 4, 10, 32), but no bidder submitted more than 517 bids (TELUS Communications Inc. submitted 517 bids in the supplementary round. ISED Canada, 2019a). This leads us to posit the following Conjecture.

**Conjecture 4.4.10.** *When restricted to  $WDP^\times$  instances in which the maximum number  $\max_j |\mathcal{I}^j|$  of bids that any bidder submitted is bounded by a polynomial function of  $p$  and  $\max_t s_t$ , a top-down DP algorithm with memoization for the Bellman eq. (4.28) is faster than a bottom-up DP algorithm asymptotically in  $p$  and in  $n$ .*

The purpose of memoization is to record the results of each call to  $v^j(\mathbf{q}, \mathbf{a})$  implemented in terms of the recursion in eq. (4.28) with  $\mathcal{I} = \mathcal{I}^j$ . We do not need to save results in the middle of iterating over  $\mathcal{I}^j$ , so we drop the notation for subsets of bid indices, relying on our convention that  $v^j(\mathbf{q}, \mathbf{a}) \equiv v^j(\mathbf{q}, \mathbf{a}, \mathcal{I}^j)$ , etc., for the remainder of this subsection. We also do not need to save results in the base case because eq. (4.7) on page 141 provides a complete solution when  $j = 0$ , so we assume  $j \in [b]$  for the remainder of this subsection. While some memoization schemes for some problems record just the optimal cost of a subproblem, we can save significant time reconstructing the winning allocation from the optimal objective values if we store the winning bid index at each call (Cormen et al., 2009, p. 387). To abstract away the order of iteration over  $\mathcal{I}^j$  (Bid ordering when branching on

<sup>42</sup>↑A proof of conjecture 4.4.10 would need to take into account the number of entries of  $\mathbf{q}$  that are zero. When  $\mathbf{q} = \mathbf{s}$ , the full supply, every entry of  $\mathbf{q}$  is non-zero, so  $\prod_{t=1}^p (q_t + 1) \geq \prod_{t=1}^p 2 = 2^p$ . If, for example, a bidder  $j' > j$  has a bid for a package  $\mathbf{Q}_{i'}$  that demands the entire supply  $s_t = Q_{ti'}$  of some product  $t$ , then the residual supply  $q_t$  of product  $t$  is zero.

bids in WDP is discussed in Sandholm, 2006, pp. 353–354), we suppose only that a *winning-bid-index function*  $v^j$  returns either  $\infty$  for an infeasible subproblem or some winning bid index  $v^j(\mathbf{q}, \mathbf{a}) \in \mathcal{B}^j(\mathbf{q}, \mathbf{a})$  for a feasible subproblem. Sandholm (2006, pp. 357–358) reviewed techniques for memoization in WDP subproblems that branch on bids. What remains in this subsection is to derive a technique appropriate for WDP<sup>κ</sup> when branching on bids by bidder. We then define a *memo entry* as<sup>43</sup>

$$\mu^j(\mathbf{q}, \mathbf{a}) := \begin{cases} (v^j(\mathbf{q}, \mathbf{a}), v^j(\mathbf{q}, \mathbf{a})) & \text{if } v^j(\mathbf{q}, \mathbf{a}) \neq -\infty \\ (\infty, \mathbf{a}) & \text{if } v^j(\mathbf{q}, \mathbf{a}) = -\infty. \end{cases}$$

After each time that  $v^j(\mathbf{q}, \mathbf{a})$  returns an optimal value (or  $-\infty$  for infeasibility) in the recursion of the Bellman eq. (4.28), we record the residual supply  $\mathbf{q}$  and the aggregate reserve  $\mathbf{a}$  by appending them as new columns to the matrices  $\Psi^j \in \mathbb{N}^{p \times *}$  and  $A^j \in \overline{\mathbb{R}}^{m \times *}$ , respectively. Here  $p \times *$  and  $m \times *$  mean that  $\Psi^j$  has  $p$  rows,  $A^j$  has  $m$  rows, and they both have an equal number of columns depending on how many times  $v^j$  has returned a result. We also store a memo entry in bidder  $j$ 's *memo*, which we define as<sup>44</sup>

$$\begin{aligned} c^j(\Psi^j, A^j, \mathbf{q}) &:= \mu^j(\mathbf{q}, \eta) \\ &\text{where } \eta := \min\{A_\ell^j \mid \Psi_\ell^j = \mathbf{q}\}. \end{aligned} \tag{4.36}$$

*Remark 4.4.11.* The matrices  $\Psi^j$  and  $A^j$  represent the state of the algorithm specific to bidder  $j$ , and, in particular, the order of calls to eq. (4.28). Recall our convention that  $\min \emptyset = \infty$ . Thus, if  $\mathbf{q}$  is not a column of  $\Psi^j$ , then  $c^j(\Psi^j, A^j, \mathbf{q}) = (\infty, \infty)$ . This indicates that  $v^j(\mathbf{q}, \mathbf{a})$  has not been computed yet for any aggregate reserve  $\mathbf{a}$  except possibly  $\mathbf{a} = \infty$ , for which we already know the subproblem is infeasible. After each call to eq. (4.28), we replace  $\Psi^j$  and  $A^j$  with  $[\Psi^j \ \mathbf{q}]$  and  $[A^j \ \mathbf{a}]$  respectively. As an optimization, if  $\mathbf{q}$  is already column  $\ell$  of  $\Psi^j$  and  $\mathbf{a} < \eta$ , we may instead skip updating  $\Psi^j$  and update  $A^j$  in place by replacing  $\eta$  with  $\mathbf{a}$  in column  $\ell$ . Further, if the call to eq. (4.28) found that  $\{i \in I^j \mid Q_i \leq \mathbf{q}\} = \emptyset$ , then we know both  $v^j(\mathbf{q}, \mathbf{a}) = -\infty$  and  $v^j(\mathbf{q}, -\infty) = -\infty$ , so we may update  $A^j$  by inserting  $-\infty$  rather

<sup>43</sup>↑Formally,  $v^j: \mathbb{N}^p \times \overline{\mathbb{R}}^m \rightarrow I^j \cup \{\infty\}$  and  $\mu^j: \mathbb{N}^p \times \overline{\mathbb{R}}^m \rightarrow (I^j \cup \{\infty\}) \times \overline{\mathbb{R}}^m$ .

<sup>44</sup>↑Formally,  $c^j: \bigcup_{\ell=0}^{\infty} (\mathbb{N}^{p \times \ell} \times \overline{\mathbb{R}}^{m \times \ell}) \times \mathbb{N}^p \rightarrow (I^j \cup \{\infty\}) \times \overline{\mathbb{R}}^m$ .

than  $\mathbf{a}$  in place of  $\boldsymbol{\eta}$ . Finally, we can accelerate the search for  $\mathbf{q}$  among the columns of  $\Psi^j$  by indexing the columns with a radix tree, discussed briefly in subsection 4.4.2, whose edge labels correspond to the entries of  $\Psi_\ell^j$  and whose node labels correspond to  $\ell$ .

**Lemma 4.4.12.** *Suppose  $j > 0$ . Let  $i \in \mathcal{I}^j \cup \{\infty\}$  and  $\mathbf{u} \in \overline{\mathbb{R}}^m$  such that  $(i, \mathbf{u}) := c^j(\Psi^j, \mathbf{A}^j, \mathbf{q}) = \mu^j(\mathbf{q}, \boldsymbol{\eta})$ , where  $\Psi^j, \mathbf{A}^j$ , and  $\boldsymbol{\eta}$  are the same as in eq. (4.36). Then  $\mathbf{u} \geq \mathbf{v}^j(\mathbf{q})$ . Equality holds if and only if  $\mathbf{u} = -\infty$  or  $i \neq \infty$ . If  $i$  is finite then so is  $\mathbf{u}$ . Moreover,  $\mathbf{v}^j(\mathbf{q}, \mathbf{a}) = -\infty$  if either  $\mathbf{u} < \mathbf{a}$ , or  $\mathbf{u} = \mathbf{a}$  and  $i = \infty$ ; and  $\mathbf{v}^j(\mathbf{q}, \mathbf{a}) = \mathbf{u}$  if  $\mathbf{u} \geq \mathbf{a}$  and  $i \neq \infty$ .*

We summarize lemma 4.4.12 in table 4.4 (cf. the bulleted list at Sandholm, 2006, p. 358). Notice that  $\mathbf{u} = -\infty$  implies  $u_1 = -\infty$ , and, since  $\mathbf{u} \geq \mathbf{v}^j(\mathbf{q})$ , if  $u_1 = -\infty$ , then  $\mathbf{v}^j(\mathbf{q})$  cannot be entirely finite, so it is  $-\infty$ . The lemma says that we need only compute  $\mathbf{v}^j(\mathbf{q}, \mathbf{a})$  from scratch if  $\mathbf{u} > \mathbf{a}$ ,  $i = \infty$ , and  $u_1 \neq -\infty$ ; this includes the scenario in which  $\mathbf{q}$  is not a column of  $\Psi^j$  because then  $\mathbf{u} = \infty$  and  $i = \infty$ .

Table 4.4. Summary of lemma 4.4.12: how to read the recursion memo

	$i = \infty$	$i \neq \infty$
$u_1 = -\infty$	$\times \mathbf{v}^j(\mathbf{q}) = -\infty$	(impossible)
$u_1 \neq -\infty, \mathbf{u} \notin \mathbb{R}^m$	$? \mathbf{v}^j(\mathbf{q}) < \mathbf{u}$	
$\mathbf{u} \in \mathbb{R}^m$		$\checkmark \mathbf{v}^j(\mathbf{q}) = \mathbf{u}$
$\mathbf{u} < \mathbf{a}$	$\times \mathbf{v}^j(\mathbf{q}, \mathbf{a}) = -\infty$	$\times \mathbf{v}^j(\mathbf{q}, \mathbf{a}) = -\infty$
$\mathbf{u} = \mathbf{a}$		$\checkmark \mathbf{v}^j(\mathbf{q}, \mathbf{a}) = \mathbf{u}$
$\mathbf{u} > \mathbf{a}$	$? \text{inconclusive if } u_1 \neq -\infty$	

What does the memo entry  $(i, \mathbf{u}) := c^j(\Psi^j, \mathbf{A}^j, \mathbf{q})$  tell us about subproblem  $\mathbf{v}^j(\mathbf{q}, \mathbf{a})$ ?

$\checkmark$  Memo contains solution.  $\times$  Subproblem is infeasible.  $?$  Must solve subproblem.

*Proof of lemma 4.4.12.* First suppose  $\mathbf{v}^j(\mathbf{q}) < \boldsymbol{\eta}$  or  $\mathbf{v}^j(\mathbf{q}) = \boldsymbol{\eta} = -\infty$ . According to lemma 4.1.7,  $\mathbf{v}^j(\mathbf{q}, \boldsymbol{\eta}) = -\infty$ , so  $i = \infty$  and  $\mathbf{u} = \boldsymbol{\eta}$ . Hence  $\mathbf{u} \geq \mathbf{v}^j(\mathbf{q})$ . Under our current assumptions,  $\boldsymbol{\eta} = \mathbf{v}^j(\mathbf{q})$  if and only if  $\mathbf{v}^j(\mathbf{q}) = \boldsymbol{\eta} = -\infty$ , which happens if and only if  $\mathbf{u} = -\infty$ .

Next suppose that  $\boldsymbol{\eta} \leq \mathbf{v}^j(\mathbf{q})$  and either  $-\infty \neq \boldsymbol{\eta}$  or  $\boldsymbol{\eta} \neq \mathbf{v}^j(\mathbf{q})$ . Then  $\mathbf{v}^j(\mathbf{q}) \neq -\infty$ . According to lemma 4.1.7,  $\mathbf{v}^j(\mathbf{q}, \boldsymbol{\eta}) = \mathbf{v}^j(\mathbf{q}) \neq -\infty$ , so  $i = \iota^j(\mathbf{q}, \boldsymbol{\eta}) \neq \infty$  and  $\mathbf{u} = \mathbf{v}^j(\mathbf{q}, \boldsymbol{\eta}) = \mathbf{v}^j(\mathbf{q})$ .

Now we can see that  $i$ 's being finite implies  $\mathbf{u}$ 's being finite. The suppositions leading the previous two paragraphs are mutually exclusive and cover all possibilities. In the first

one,  $i = \infty$ , whereas in the second one, both  $i$  and  $u$  were finite.

Finally, we deduce  $v^j(\mathbf{q}, \mathbf{a})$  from  $i$  and  $u$ . An aggregate reserve of  $-\infty$  completely relaxes the AR constraint, so  $v^j(\mathbf{q}, \mathbf{a}) \leq v^j(\mathbf{q}, -\infty)$ . By eq. (4.5),  $v^j(\mathbf{q}, -\infty) = v^j(\mathbf{q})$ . From the foregoing considerations, we have that  $v^j(\mathbf{q}) \leq u$ , so  $v^j(\mathbf{q}, \mathbf{a}) \leq u$ . According to lemma 4.1.7,  $u < \mathbf{a}$  implies  $v^j(\mathbf{q}, \mathbf{a}) = -\infty$ . If  $i = \infty$ , then  $v^j(\mathbf{q}, \boldsymbol{\eta}) = -\infty$  and  $u = \boldsymbol{\eta}$ . If additionally  $u = \mathbf{a}$ , then  $\mathbf{a} = \boldsymbol{\eta}$ , so  $v^j(\mathbf{q}, \mathbf{a}) = -\infty$ . However, if  $i \neq \infty$ , then  $v^j(\mathbf{q}, \boldsymbol{\eta}) \neq -\infty$  and  $u = v^j(\mathbf{q}, \boldsymbol{\eta}) = v^j(\mathbf{q})$ . If additionally  $u \geq \mathbf{a}$ , then  $v^j(\mathbf{q}) \geq \mathbf{a}$ , so  $v^j(\mathbf{q}, \mathbf{a}) = v^j(\mathbf{q}) = u$  by lemma 4.1.7.  $\square$

The memoization strategy for WDP that Sandholm (2006, pp. 357–358) presented includes two extensions that we have not contemplated here. The first is to store the result of the LP relaxation of a subproblem in the memo entry. Then in the “? inconclusive” case in table 4.4, if the cached upper bound is less than  $\mathbf{a}$ , we can prune the current subproblem as infeasible. In the context of WDP<sup>ε</sup> with  $m > 1$  using theorem 4.3.7, we would have to store both  $k^*$  and  $z_{[k^*]}^j(\mathbf{q}, \mathbf{a})$ . Such a technique requires further study to determine if the resulting pruning justifies the enlarged memo entries. The second extension is to eject large or rarely used entries from the cache, or to be more selective about which subproblems get saved at all. The need arises in large problems when the memos start to take up too much memory. Sandholm represented packages as sets of items whereas we represent them as vectors. Our representation may be amenable to carefully ordering the iteration over  $\mathcal{I}^j$  in the optimization in eq. (4.28), e.g., by sorting the columns of  $\mathbf{Q}$  lexicographically. It may be possible to prove some subproblems won’t be used again after a certain point as long as unexplored nodes of the search tree always iterate in the same order. We could then delete the corresponding columns of  $\boldsymbol{\Psi}^j$  and  $\mathbf{A}^j$  and corresponding memo entries.

#### 4.5 Price Determination

An auction’s *pricing rule*  $\wp$  is how the auctioneer determines how much to charge winning bidders in exchange for the packages they win. In detail, if we let BS stand for the entire bid stack  $\{\mathcal{I}^j\}_{j=1}^b$ ,  $\mathbf{V}$ , and  $\mathbf{Q}$ , then  $\wp^j(\text{BS}, \mathbf{q}) \in \mathbb{R}$  is the amount bidder  $j$  would have to pay to the auctioneer in exchange for winning package  $\mathbf{q}$ . This formulation allows us to discuss hypotheticals, such as the reserve-price rule that prevents a bidder from winning a

package  $\mathbf{q}$  if  $\mathbf{r}^\top \mathbf{q} < \varphi^j(\text{BS}, \mathbf{q})$ . The pricing rules of choice in combinatorial auctions are from the class of *opportunity-cost pricing rules* in which the auctioneer charges the winner of a package at least as much as any other bidder (or combination of bidders) would have been willing to pay for that package. Our focus in this section is on *forward auctions*, which, for our purposes, are auctions for which all bid prices—the first row of the values matrix  $\mathbf{V}$ —are all nonnegative. A *procurement* or *reverse auction*, in which the auctioneer is buying rather than selling (Cramton et al., 2006a, pp. 622–623), has all nonpositive bid prices.

Our pricing algorithms fairly closely follow Maldoom (2007)'s rendition of core-constraint generation (CCG) that Day and Raghavan (2007, § 4 on pp. 1396–1398) proposed. The main difference is that we tend to formulate the relevant mathematical programs in terms of prices rather than payoffs. Day and Cramton (2012) described how the algorithm works in practice in several real-world auctions.

**4.5.1 Counterfactuals.** A common tool we use is the *counterfactual WDP*, for which we need the following notation. The *positive part*  $\alpha^+$  of a number (or infinity)  $\alpha \in \overline{\mathbb{R}}$  is  $\max\{\alpha, 0\}$ . Define the *reduction*  $\beta_j \in \mathbb{R} \cup \{\infty\}$  for bidder  $j > 0$ . Then we may define the *counterfactual WDP reduced by  $\beta$*  to be

$$\rho_\beta^* := \max_{\mathbf{x}} \mathbf{r}^\top (\mathbf{s} - \mathbf{Q}\mathbf{x}) + \sum_{j=1}^b \sum_{i \in \mathcal{I}^j} (V_{1i} - \beta_j)^+ x_i \quad (4.37)$$

subject to  $\mathbf{x} \in \mathcal{X}^b(\mathbf{s})$ .

Equation (4.37) is just an instance of  $\text{WDP}^\times$  with  $m = 1$  tie breaker (hence no need for lexicographic maximization) and  $(V_{1i} - \beta_j)^+$  replacing  $V_i$  for each  $i \in \mathcal{I}^j$  and each  $j \in [b]$ . In the objective function of eq. (4.37), the summand  $(V_{1i} - \beta_j)^+$  is always finite because  $V_{1i}$  is real so  $(V_{1i} - \infty)^+ = 0$  for all  $j \in [b]$  and all  $i \in \mathcal{I}^j$ . Moreover, since we are currently assuming the first row of  $\mathbf{V}$  has all nonnegative entries, we have  $(V_{1i} - 0)^+ = V_{1i}$ .

In analogy to definitions 4.1.5 and 4.1.6, we define the *subproblem* of the counterfactual WDP reduced by  $\beta \in (\mathbb{R} \cup \{\infty\})^b$  for current bidder  $j \in \{0, \dots, b\}$ , residual supply  $\mathbf{q} \in \mathbb{N}^p$  such that  $\mathbf{q} \leq \mathbf{s}$ , aggregate reserve  $a_1 \in \overline{\mathbb{R}}$ , and bid indices  $\mathcal{I} \subseteq \mathcal{I}^j$  to be eq. (4.38),



and we denote its maximum value as  $\rho_\beta^j(\mathbf{q}, a_1, \mathcal{I})$ .<sup>45</sup>

$$\underset{\mathbf{x}}{\text{maximize}} \mathbf{r}^\top(\mathbf{q} - \mathbf{Q}\mathbf{x}) + \sum_{k=1}^b \sum_{i \in \mathcal{I}^k} (V_{1i} - \beta_k)^+ x_i \quad (4.38)$$

subject to  $\mathbf{x} \in \mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$ ,

where  $\mathbf{a} := [a_1, -\infty, \dots, -\infty]^\top$ .

As in eq. (4.5), we set  $\rho_\beta^j(\mathbf{q}, a_1) := \rho_\beta^j(\mathbf{q}, a_1, \mathcal{I}^j)$  and  $\rho_\beta^j(\mathbf{q}) := \rho_\beta^j(\mathbf{q}, -\infty)$ . It follows that  $\rho_\beta^b(s) = \rho_\beta^*$ .

**Lemma 4.5.1.** Fix  $a_1 \in \overline{\mathbb{R}}$  and let  $\mathbf{a} := [a_1, -\infty, \dots, -\infty]^\top$ . If  $\beta_k = 0$  for all  $k \leq j$ , then

$$\rho_\beta^j(\mathbf{q}, a_1, \mathcal{I}) = v_1^j(\mathbf{q}, \mathbf{a}, \mathcal{I}), \quad (4.39)$$

*Proof.* The feasible set  $\mathcal{X}^j(\mathbf{q}, \mathbf{a}, \mathcal{I})$  of eq. (4.4) is the same as that of eq. (4.38). Every  $\mathbf{x}$  in the feasible set has  $x_i = 0$  for all  $i \in \mathcal{I}^k$  for all  $k > j$ . Since  $V_{1i} \in [0, \infty)$  and  $k \leq j$  implies  $\beta_k = 0$ , we have  $(V_{1i} - \beta_k)^+ = V_{1i}$  for all  $i \in \mathcal{I}^k$  for all  $k \leq j$ . By eq. (4.2),  $\mathbf{e}_1^\top \mathbf{R} = \mathbf{r}$ . Hence the objective function in eq. (4.38) is

$$\mathbf{r}^\top(\mathbf{s} - \mathbf{Q}\mathbf{x}) + \sum_{k=1}^b \sum_{i \in \mathcal{I}^k} (V_{1i} - \beta_k)^+ x_i = \mathbf{r}^\top(\mathbf{s} - \mathbf{Q}\mathbf{x}) + \sum_{k=1}^j \sum_{i \in \mathcal{I}^k} V_{1i} x_i = \mathbf{e}_1^\top [\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{s} - \mathbf{Q}\mathbf{x})].$$

Lexicographic maximization in eqs. (4.3) and (4.4) maximizes the first entry  $\mathbf{e}_1^\top \mathbf{V}\mathbf{x}$  of  $\mathbf{V}\mathbf{x}$  first, so, for any  $\mathcal{Y} \subseteq \{0, 1\}^n$ ,

$$\max\{\mathbf{e}_1^\top \mathbf{V}\mathbf{x} + \mathbf{r}^\top(\mathbf{q} - \mathbf{Q}\mathbf{x}) \mid \mathbf{x} \in \mathcal{Y}\} = \mathbf{e}_1^\top \max\{\mathbf{V}\mathbf{x} + \mathbf{R}(\mathbf{q} - \mathbf{Q}\mathbf{x}) \mid \mathbf{x} \in \mathcal{Y}\}.$$

Hence the objective function of  $v_1^j$  is  $\mathbf{e}_1^\top \mathbf{V}\mathbf{x} + \mathbf{r}^\top(\mathbf{q} - \mathbf{Q}\mathbf{x})$  maximized over  $(\overline{\mathbb{R}}, <)$ . With the same feasible set, objective function, and order, the left and right sides of eq. (4.39) are equal.  $\square$

We say that  $\rho_\beta^*$  in eq. (4.37) is the *counterfactual WDP for bidders*  $\mathcal{A} \subseteq [b]$  if

$$\beta_j = \begin{cases} \infty & j \in \mathcal{A} \\ 0 & j \notin \mathcal{A} \end{cases}$$

<sup>45</sup>Formally,  $\rho_\beta^j: \mathbb{N}^p \times \overline{\mathbb{R}} \times 2^{\mathcal{I}^j} \rightarrow \overline{\mathbb{R}}$ .

We will have no need for sets  $\mathcal{A}$  for which any bidder  $j \in \mathcal{A}$  *lost* the auction for a winning allocation  $\mathbf{x}$  in eq. (4.3), which is to say that  $V_{1i} = 0$  for the  $i \in \mathcal{I}^j$  such that  $x_i = 1$ . (A loser may still have won a positive package; losing simply refers to the package's being associated with no bid value in the first values row.)

**4.5.2 Reusing Memos.** We may speed up computation of a counterfactual WDP reduced by  $\beta$  as follows. Suppose we have computed the WDP<sup>\*</sup> in eq. (4.3) using dynamic programming via theorem 4.4.4. Every time we evaluated  $v^j(\mathbf{q}, \mathbf{a})$  for  $j \in [b]$ ,  $\mathbf{q} \in \mathbb{N}^p$  such that  $\mathbf{q} \leq \mathbf{s}$ , and  $\mathbf{a} \in \overline{\mathbb{R}}^m$ , we stored  $\mathbf{q}$  and  $\mathbf{a}$  as columns of matrices  $\Psi^j$  and  $A^j$  respectively. From there we computed memos  $c^j$  as in eq. (4.36). Let  $\mathcal{A} := \{j \in [b] \mid \beta_j \neq 0\}$ . Putting together lemmas 4.4.12 and 4.5.1, we obtain the following corollary.

**Corollary 4.5.2.** *Consider the setup in the paragraph above. Suppose  $j \in [\min \mathcal{A} - 1]$ . Let  $i \in \mathbb{N} \cup \{\infty\}$  and  $\mathbf{u} \in \overline{\mathbb{R}}^m$  such that  $c^j(\Psi^j, A^j, \mathbf{q}) =: (i, \mathbf{u})$ . If  $\mathbf{u} = -\infty$  or  $i \neq \infty$ , then  $\rho_\beta^j(\mathbf{q}) = u_1$ . If  $u_1 < a$  or  $u_1 = a$  and  $i = \infty$ , then  $\rho_\beta^j(\mathbf{q}, a) = -\infty$ .*

When computing the counterfactual for a set of bidders the least of whose id is  $j$ , the recursion memo storing  $v^b(\mathbf{s})$  does not need to modify its stored value of  $v^k(\mathbf{q})$  for  $k < j$  and any package  $\mathbf{q}$ . This allows subsequent runs of the dynamic programming solver to avoid recomputing its entire recursion memo.

## BIBLIOGRAPHY

- Adke, S. R., & Swamy, R. J. R. (1979). Some problems of inference about markov sequences with exponential type transition densities. *Journal of the Indian Statistical Association*, 17, 173–188.
- Aho, A. V., Garey, M. R., & Ullman, J. D. (1972). The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1(2), 131–137. <https://doi.org/10.1137/0201008>
- Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels (T. Jaakkola, Ed.). *Journal of Machine Learning Research*, 9, 1981–2014. Retrieved April 2, 2019, from <http://www.jmlr.org/papers/v9/airoldio8a.html>
- Al-Eideh, B., Abu-Salih, M. S., & Capar, U. (1988). Consistency of maximum likelihood estimators for multiparameter Markov chains. *Journal of Information and Optimization Sciences*, 9(3), 391–404. <https://doi.org/10.1080/02522667.1988.10698938>
- Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331), 1248–1255. <https://doi.org/10.2307/2284291>
- Auction of Flexible-Use Service Licenses in the 3.45–3.55 GHz Band for Next-Generation Wireless Services; Notice and Filing Requirements, Minimum Opening Bids, Upfront Payments, and Other Procedures for Auction 110; Bidding to Begin October 5, 2021 AU Docket No. 21-62, US (2021, June 9). Retrieved October 1, 2021, from <https://www.fcc.gov/document/procedures-established-auction-110-345-355-ghz-band>
- Ausubel, L. M., & Milgrom, P. (2006). The lovely but lonely vickrey auction [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 1–40). MIT Press.
- Axler, S. (1997). *Linear algebra done right* (2nd ed.). Springer.
- Axler, S. (2015, September 26). *Showing that  $n$  exponential functions are linearly independent*. Mathematics Stack Exchange. Retrieved August 10, 2021, from <https://math.stackexchange.com/a/1451686>
- Bannister, M. J., Devanny, W. E., & Eppstein, D. (2014, December 4). *ERGMs are hard*.
- Barndorff-Nielsen, O. E. (1978). *Information and exponential families: In statistical theory*. John Wiley & Sons. <https://doi.org/10.1002/9781118857281>

- Berger, R. L., & Boos, D. D. (1994). *P* Values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427), 1012–1016. <https://doi.org/10.1080/01621459.1994.10476836>
- Bertsimas, D., & Tsitsiklis, J. N. (1997). *Introduction to linear optimization*. Athena Scientific; Dynamic Ideas.
- Bhat, B. R. (1988). On exponential and curved exponential families in stochastic processes. *Mathematical Scientist*, 13(2), 121–134. Retrieved February 9, 2018, from [http://www.appliedprobability.org/data/files/TMS%20articles/13\\_2\\_5.pdf](http://www.appliedprobability.org/data/files/TMS%20articles/13_2_5.pdf)
- Bhat, B. R., & Gani, J. M. (1960). A note on sufficiency in regular Markov chains. *Biometrika*, 47, 452–457. <https://doi.org/10.2307/2333317>
- Bikhchandani, S., & Ostroy, J. M. (2006). From the assignment model to combinatorial auctions [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 189–214). MIT Press.
- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice* (R. J. Light & F. Mosteller, Collaborators). Springer-Verlag. <https://doi.org/10.1007/978-0-387-72806-3>. Repr. of Light, R. J., & Mosteller, F. (Collaborators). (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press
- Blum, J. R., & Hanson, D. L. (1963). Further results on the representation problem for stationary stochastic processes with trivial tail field. *Journal of Mathematics and Mechanics*, 12(6), 935–943.
- Bofinger, E. (1965). Sufficiency for multinomial and transition probabilities. *Journal of Applied Probability*, 2(2), 470–474.
- Bono, J., Ingraham, A. T., Ravi, S., Schwartz, W. K., & Sojourner, C. (2019, September 20). *An analysis of allocation phase pricing and clock round price increases in the canadian 600 mhz auction*. <https://doi.org/10.2139/ssrn.3463933>
- Broadcast Incentive Auction Scheduled To Begin on March 29, 2016; Procedures for Competitive Bidding in Auction 1000, US (2015, October 14). Retrieved October 29, 2021, from <https://www.govinfo.gov/content/pkg/FR-2015-10-14/xml/FR-2015-10-14.xml#seqnum61917>
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.) [Twentieth Indian Reprint 2017]. Cengage Learning.

- Chatterjee, S., & Diaconis, P. (2013). Estimating and understanding exponential random graph models. *Annals of Statistics*, 41(5), 2428–2461. <https://doi.org/10.1214/13-AOS1155>. An earlier version of the preprint contains a § 2.2, apparently cut before publication, that gives a closed-form formula for the partition function of the Erdős-Rényi graph model of example 2.5.2: *Estimating and understanding exponential random graph models*. (2011, April 6)
- Chatterjee, S., Diaconis, P., & Sly, A. (2011). Random graphs with a given degree sequence. *Annals of Applied Probability*, 21(4), 1400–1435. <https://doi.org/10.1214/10-AAP728>
- Chung, K. L. (2000, October 9). *A course in probability theory* (3rd ed.). Academic Press. (Original work published 1968)
- Clark, A. (1984). *Elements of abstract algebra* [Corrected republication]. Dover. (Original work published 1971)
- Clarke, E. H. (1971). Multipart pricing of public goods. *Public Choice*, 11(1), 17–33. <https://doi.org/10.1007/BF01726210>
- Comment Sought on Competitive Bidding Procedures for Broadcast Incentive Auction 1000, Including Auctions 1001 and 1002 Docket no. AU 14-252, GN 12-268, US (2014, December 17). Retrieved October 29, 2021, from <https://www.fcc.gov/document/broadcast-incentive-auction-comment-pn>
- Commission for Communications Regulation. (2017, June 1). *3.6 GHz band spectrum award*. Ireland. Retrieved August 2, 2021, from <https://www.comreg.ie/industry/radio-spectrum/spectrum-awards/3-6ghz-band-spectrum-award/>
- Commission for Communications Regulation. (2021, May 31). *Multi band spectrum award 2021*. Ireland. Retrieved August 2, 2021, from <https://www.comreg.ie/industry/radio-spectrum/spectrum-awards/proposed-multi-band-spectrum-award/>
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3rd ed.). MIT Press.
- Cox, D., Little, J., & O’Shea, D. (2015). *Ideals, varieties, and algorithms: An introduction to computational algebraic geometry and commutative algebra* (4th ed.). Springer. <http://www.dm.unipi.it/~caboara/Misc/Cox,%20Little,%20O’Shea%20-%20Ideals,%20varieties%20and%20algorithms.pdf>
- Cramton, P., Shoham, Y., & Steinberg, R. *Combinatorial auction glossary* (P. Cramton, Y. Shoham, & R. Steinberg, Eds.) [First MIT Press paperback edition, 2010]. In: ed., with an introd., by Cramton, P., Shoham, Y., & Steinberg, R. With a forew. by Smith, V. L. First MIT Press paperback edition, 2010. Cambridge, Massachusetts: MIT Press, 2006, 613–625. ISBN: 978-0-262-51413-2.

- Cramton, P., Shoham, Y., & Steinberg, R. (Eds.). (2006b). *Combinatorial auctions* [First MIT Press paperback edition, 2010]. MIT Press.
- Cramton, P., Shoham, Y., & Steinberg, R. *Introduction to combinatorial auctions* (P. Cramton, Y. Shoham, & R. Steinberg, Eds.) [First MIT Press paperback edition, 2010]. In: ed., with an introd., by Cramton, P., Shoham, Y., & Steinberg, R. With a forew. by Smith, V. L. First MIT Press paperback edition, 2010. Cambridge, Massachusetts: MIT Press, 2006, 1–13. ISBN: 978-0-262-51413-2.
- Danish Energy Agency. (2021a, April 21). *Spectrum: Auctions*. Ministry of Climate, Energy & Utilities. Retrieved October 29, 2021, from <https://ens.dk/en/our-responsibilities/spectrum/auctions>
- Day, R. W., & Cramton, P. (2012–June). Quadratic core-selecting payment rules for combinatorial auctions. *Operations Research*, 60(3), 588–603. <https://doi.org/10.1287/opre.1110.1024>
- Day, R. W., & Raghavan, S. (2007). Fair payments for efficient allocations in public sector combinatorial auctions. *Management Science*, 53(9), 1389–1406. <https://doi.org/10.1287/mnsc.1060.0662>
- Denny, J. L. (1972). Sufficient statistics and discrete exponential families. *Annals of Mathematical Statistics*, 43(4), 1320–1322.
- Diaconis, P. (1988). *Group representations in probability and statistics* (S. S. Gupta, Ed.). Institute of Mathematical Statistics.
- Diaconis, P., & Freedman, D. (1981, December 16–19). Partial exchangeability and sufficiency. In J. K. Ghosh & J. Roy (Eds.), *Statistics: Applications and new directions* (pp. 205–236). Statistical Publication Society. Retrieved November 19, 2019, from [https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis\\_freedman\\_PES.pdf](https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis_freedman_PES.pdf)
- Diaconis, P., & Freedman, D. (1999). Iterated random functions. *SIAM Review*, 41(1), 45–76. Retrieved July 8, 2021, from <https://statweb.stanford.edu/~cgates/PERSI/papers/iterate.pdf>
- Diaconis, P., & Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26(1), 363–397. <https://doi.org/10.1214/aos/1030563990>
- Eben-Chaime, M. (1996). Parametric solution for linear bicriteria knapsack models. *Management Science*, 42(11), 1565–1575.
- Ehrgott, M. (2005). *Multicriteria optimization* (2nd ed.). Springer. <https://doi.org/10.1007/3-540-27659-9>

- Ehrgott, M., & Gandibleux, X. (2000). A survey and annotated bibliography of multiobjective combinatorial optimization. *OR-Spektrum*, 22(4), 425–460. <https://doi.org/10.1007/s002910000046>
- Federal Communications Commission. (2017a, April 13). *Public reporting system: Incentive Auction* [Assignment phase bids]. Retrieved October 11, 2021, from <https://auctiondata.fcc.gov/public/projects/1000/reports/assignment-bids>. Available from *Public reporting system: Incentive Auction dashboard*. (2017, April 13). Retrieved October 11, 2021, from <https://auctiondata.fcc.gov/public/projects/1000>. Data provenance described at paras. 7 and 21 in Incentive Auction Closing and Channel Reassignment Public Notice; Incentive Auction Closes; Reverse Auction and Forward Auction Results Announced; Final Television Band Channel Assignments Announced; Post-Auction Deadlines Announced Docket no. AU 14-252, GN 12-258, WT 12-269, and MB 16-306, US (2017, April 13). Retrieved October 11, 2021, from <https://www.fcc.gov/document/fcc-announces-results-worlds-first-broadcast-incentive-auction-0>
- Federal Communications Commission. (2021b, November 9). *Public reporting system: Auction 110 – 3.45 ghz*. Retrieved October 29, 2021, from <https://auctiondata.fcc.gov/public/projects/auction110>
- Feigin, P. D. (1981). Conditional exponential families and a representation theorem for asymptotic inference. *Annals of Statistics*, 9(3), 597–603. <https://doi.org/10.1214/aos/1176345463>
- Fienberg, S. E., Meyer, M. M., & Wasserman, S. S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389), 51–67. <https://doi.org/10.2307/2288040>
- Fienberg, S. E., & Rinaldo, A. (2012a). Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40(2), 996–1023. <https://doi.org/10.1214/12-AOS986>
- Fienberg, S. E., & Rinaldo, A. (2012b). Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40(2), 996–1023. <https://doi.org/10.1214/12-AOS986>
- 1500 MHz, 2100 MHz, 2300 MHz, 3.5 GHz and 26 GHz Auction 2021, Denmark (2021, February 12). Retrieved October 29, 2021, from [https://ens.dk/sites/ens.dk/files/Tele/information\\_memorandum\\_1.pdf](https://ens.dk/sites/ens.dk/files/Tele/information_memorandum_1.pdf)
- Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395), 832–842. <https://doi.org/10.2307/2289017>
- Fudenberg, D., & Tirole, J. (1991, August). *Game theory*. The MIT Press. Retrieved October 29, 2021, from <https://homepage.univie.ac.at/Mariya.Teteryatnikova/WS2011/FT.pdf>

- Fujishima, Y., Leyton-Brown, K., & Shoham, Y. (1999). Taming the computational complexity of combinatorial auctions: Optimal and approximate approaches. In T. Dean (Ed.), *Proceedings* (pp. 548–553). Morgan Kaufmann Publishers. Retrieved July 27, 2021, from <https://www.ijcai.org/Proceedings/99-1/Papers/079.pdf>
- Gani, J. M. (1955). Some theorems and sufficiency conditions for the maximum-likelihood estimator of an unknown parameter in a simple Markov chain. *Biometrika*, *42*, 342–359. <https://doi.org/10.1093/biomet/42.3-4.342>. For a correction to its § IV.1, see Corrigenda. (1956). *Biometrika*, *43*, 497–498. <https://doi.org/10.1093/biomet/43.3-4.497>
- Gani, J. M. (1956). Sufficiency conditions in regular Markov chains and certain random walks. *Biometrika*, *43*, 276–284. <https://doi.org/10.2307/2332906>
- Givens, G. H., & Hoeting, J. A. (2012, November). *Computational statistics* (2nd ed.) [1st printing]. John Wiley & Sons, Inc. Retrieved August 13, 2021, from [http://home.ustc.edu.cn/~liweiyu/documents/Geof%20H.%20Givens%20%20Jennifer%20A.%20Hoeting\(auth.\)%20-%20Comp.pdf](http://home.ustc.edu.cn/~liweiyu/documents/Geof%20H.%20Givens%20%20Jennifer%20A.%20Hoeting(auth.)%20-%20Comp.pdf). Errata available from *Computational statistics: Second edition*. (2014, March 25). Retrieved August 13, 2021, from <https://www.stat.colostate.edu/computationalstatistics/>
- Makhorin, A. (2020, December). *GNU linear programming kit* (comp. software; Version 5.0). Retrieved July 28, 2021, from <https://www.gnu.org/software/glpk/>
- Godsil, C., & Royle, G. (2001). *Algebraic graph theory*. Springer.
- Goldberg, D. (1991). What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, *23*(1), 5–48. <https://doi.org/10.1145/103162.103163>
- Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, *2*(2), 129–233. <https://doi.org/10.1561/2200000005>
- Golub, G. H., & Van Loan, C. (2013). *Matrix computations* (4th ed.). The Johns Hopkins University Press. Retrieved October 12, 2021, from <http://math.ecnu.edu.cn/~jypan/Teaching/books/2013%20Matrix%20Computations%204th.pdf> (Original work published 1983)
- Goodreau, S. M. (2007). Advances in exponential random graph ( $p^*$ ) models applied to a large social network. *Social Networks*, *29*(2), 231–248. <https://doi.org/10.1016/j.socnet.2006.08.001>
- Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, *85*(4), 1033–1063. <https://doi.org/10.3982/ECTA12679>



- Grindrod, P., & Higham, D. J. (2010). Evolving graphs: Dynamical models, inverse problems and propagation. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2115), 753–770. <https://doi.org/10.1098/rspa.2009.0456>
- Grindrod, P., & Higham, D. J. (2013). A matrix iteration for dynamic network summaries. *SIAM Review*, 55(1), 118–128. <https://doi.org/10.1137/110855715>
- Grindrod, P., & Parsons, M. C. (2011). Social networks: Evolving graphs with memory dependent edges. *Physica A: Statistical Mechanics and its Applications*, 390, 3970–3981. <https://doi.org/10.1016/j.physa.2011.06.015>
- Grindrod, P., Parsons, M. C., Higham, D. J., & Estrada, E. (2011). Communicability across evolving networks. *Physical Review E*, 83(4), Article 046120. <https://doi.org/10.1103/PhysRevE.83.046120>
- Gross, J., & Yellen, J. (1999). *Graph theory and its applications* (K. H. Rosen, Ed.). CRC Press.
- Groves, T. (1973). Incentive in teams. *Econometrica*, 41(4), 617–631. <https://doi.org/10.2307/1914085>
- Gurobi Optimization, LLC. (2021). *Gurobi Optimizer reference manual*. Comp. software. Version 9.1.2. Retrieved July 28, 2021, from <https://www.gurobi.com/documentation/9.1/refman/index.html>
- Haberman, S. J. (1981). Comment [Comment on “An exponential family of probability distributions for directed graphs”]. *Journal of the American Statistical Association*, 76(373), 60–61. <https://doi.org/10.1080/01621459.1981.10477602>
- Hall, M., Jr. (1967). *Combinatorial theory* (G. Birkhoff & A. W. Tucker, Eds.). Blaisdell Publishing Company.
- Handcock, M. S. (2003, December 31). *Assessing degeneracy in statistical models of social networks* (Working Paper No. 39). Center for Statistics and the Social Science, University of Washington. Seattle. Retrieved July 1, 2021, from <https://csss.uw.edu/Papers/wp39.pdf>
- Hanneke, S., Fu, W., & Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4, 585–605. <https://doi.org/10.1214/09-EJS548>
- Hanneke, S., & Xing, E. P. (2006). Discrete temporal models of social networks: Models, issues, and new directions (E. M. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, & A. X. Zheng, Eds.) [Revised selected papers], (4503), 115–125. <https://doi.org/10.1007/978-3-540-73133-7>
- Hanson, D. L. (1963). On the representation problem for stationary stochastic processes with trivial tail field. *Journal of Mathematics and Mechanics*, 12(2), 293–301.

- Hayashi, M., & Watanabe, S. (2016). Information geometry approach to parameter estimation in Markov chains. *Annals of Statistics*, 44(4), 1495–1535. <https://doi.org/10.1214/15-AOS1420>
- Heyde, C. C., & Feigin, P. D. (1974, July 29–August 10). On efficiency and exponential families in stochastic process estimation. In G. P. Patil, S. Kotz, & J. K. Ord (Eds.), *A modern course on statistical distributions in scientific work. Vol. 1. Models and structures* (pp. 227–240). D. Reidel Publishing Company. [https://doi.org/10.1007/978-94-010-1842-5\\_18](https://doi.org/10.1007/978-94-010-1842-5_18)
- Higham, N. J. (2002). *Accuracy and stability of numerical algorithms* (2nd ed.). Society for Industrial; Applied Mathematics. <https://doi.org/10.1137/1.9780898718027> (Original work published 1996)
- Hildebrand, M. (2005). A survey of results on random random walks on finite groups. *Probability Surveys*, 2, 33–63. <https://doi.org/10.1214/154957805100000087>
- Hirayama, T., & Yano, K. (2013). Strong solutions of Tsirel'son's equation in discrete time taking values in compact spaces with semigroup action. *Statistics & Probability Letters*, 83(3), 824–828. <https://doi.org/10.1016/j.spl.2012.11.033>
- Hoel, P. G., Port, S. C., & Stone, C. J. (1972). *Introduction to stochastic processes*. Waveland Press.
- Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *Annals of Applied Statistics*, 9(3), 1169–1193. <https://doi.org/10.1214/15-AOAS839>
- Hoffman, K., Dunford, M., Durbin, M., Menon, D., Sultana, R., & Wilson, T. (2001, October 27). *Issues in scaling up the 700 MHz auction design* (presentation) [K. Hoffman presented on behalf of the FCC] [available from <https://www.fcc.gov/auctions/conferences/combinatorial-bidding-conference-2001>]. Queenstown, Maryland, Federal Communications Commission. Retrieved October 12, 2021, from [https://wireless.fcc.gov/auctions/conferences/combin2001/papers/Wye\\_river\\_DAC.pdf](https://wireless.fcc.gov/auctions/conferences/combin2001/papers/Wye_river_DAC.pdf)
- Hoffmann-Jørgensen, J. (1994). *Probability with a view toward statistics* (Vol. 1). Chapman & Hall.
- Holland, P. W., & Leinhardt, S. (1977). A dynamic model for social networks. *Journal of Mathematical Sociology*, 5(1), 5–20. <https://doi.org/10.1080/0022250X.1977.9989862>
- Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373), 33–50. <https://doi.org/10.2307/2287037>

- Holte, R. C. (2001). Combinatorial auctions, knapsack problems, and hill-climbing search. In E. Stroulia & S. Matwin (Eds.), *Advances in artificial intelligence: 14th biennial conference of the canadian society for computational studies in intelligence* [Proceedings] (pp. 57–66). Springer. [https://doi.org/10.1007/3-540-45153-6\\_6](https://doi.org/10.1007/3-540-45153-6_6)
- Hudson, I. L. (1982). Large sample inference for markovian exponential families with application to branching processes with immigration. *Australian & New Zealand Journal of Statistics*, 24(1), 98–112.
- Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481), 248–258. <https://doi.org/10.1198/016214507000000446>
- Hunter, J. K., & Nachtergaele, B. (2005). *Applied analysis* (2nd ed.). World Scientific. <https://www.math.ucdavis.edu/~hunter/book/pdfbook.html>
- Hwang, S. Y., & Basawa, I. V. (1994). Large sample inference for conditional exponential families with applications to nonlinear time series. *Journal of Statistical Planning and Inference*, 38(2), 141–157.
- Illinois Institute of Technology Office of the General Counsel. (n.d.). *Faculty handbook*. Retrieved November 12, 2021, from <https://web.iit.edu/general-counsel/faculty-handbook>
- IBM Corporation. (2019, December). *ILOG CPLEX optimization studio*. Comp. software. Version 12.10.0. Retrieved July 28, 2021, from <https://www.ibm.com/docs/en/icos/12.10.0>
- Incentive Auction Closing and Channel Reassignment Public Notice; Incentive Auction Closes; Reverse Auction and Forward Auction Results Announced; Final Television Band Channel Assignments Announced; Post-Auction Deadlines Announced Docket no. AU 14-252, GN 12-258, WT 12-269, and MB 16-306, US (2017, April 13). Retrieved October 11, 2021, from <https://www.fcc.gov/document/fcc-announces-results-worlds-first-broadcast-incentive-auction-o>
- Industry Canada. (2015, October 2). *700 MHz auction (2014)*. Government of Canada. Retrieved August 2, 2021, from [https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/h\\_sf10598.html](https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/h_sf10598.html)
- Innovation, Science and Economic Development Canada. (2019a, May 31). *All supplementary round bids*. Government of Canada. Retrieved July 29, 2021, from [https://www.ic.gc.ca/eic/site/smt-gst.nsf/vwapj/600\\_b\\_supp\\_en.csv/\\$file/600\\_b\\_supp\\_en.csv](https://www.ic.gc.ca/eic/site/smt-gst.nsf/vwapj/600_b_supp_en.csv/$file/600_b_supp_en.csv). Available from *Auction of spectrum licences in the 600 MHz band: Auction results* [Bidding information]. (2019, May 31). Government of Canada. Retrieved July 29, 2021, from [https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/h\\_sf11331.html](https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/h_sf11331.html)

- Innovation, Science and Economic Development Canada. (2019b, May 31). *Auction of spectrum licences in the 600 MHz band*. Government of Canada. Retrieved July 29, 2021, from [https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/h\\_sf11331.html](https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/h_sf11331.html)
- Jacod, J., & Protter, P. (2004). *Probability essentials* (2nd ed.) [Corrected second printing]. Springer. <https://doi.org/10.1007/978-3-642-55682-1>
- Karwa, V., Pati, D., Petrović, S., & Schwartz, W. K. (2021–present). *Hypothesis tests for mixed membership stochastic block models* [authors are in alphabetical order] [Manuscript in preparation].
- Karwa, V., Pati, D., Petrović, S., Solus, L., Alexeev, N., Raič, M., Wilburne, D., Williams, R., & Yan, B. (2016, December 19). *Monte Carlo goodness-of-fit tests for degree corrected and related stochastic blockmodels*.
- Katz, L., & Proctor, C. H. (1959). The concept of configuration of interpersonal relations in a group as a time-dependent stochastic process. *Psychometrika*, 24(4), 317–327. <https://doi.org/10.1007/BF02289814>
- Klamroth, K., & Wiecek, M. M. (2002). Dynamic programming approaches to the multiple criteria knapsack problem. *Naval Research Logistics*, 47(1), 57–76. <http://www2.math.uni-wuppertal.de/~klamroth/publications/klwidpoo.pdf>
- Klemperer, P. (1999). Auction theory: A guide to the literature. *Journal of Economic Surveys*, 13(3), 227–286. <https://doi.org/10.1111/1467-6419.00083>. Reprinted as A survey of auction theory. (2004). In *Auctions: Theory and practice* (pp. 9–65). Princeton University Press. Added an afterword and exercises not in the original
- Klemperer, P. (2004a). *Auctions: Theory and practice*. Princeton University Press.
- Klemperer, P. (2004b). A survey of auction theory. *Auctions: Theory and practice* (pp. 9–65). Princeton University Press. Added an afterword and exercises not in the original. (Orig. pub. as Auction theory: A guide to the literature by Blackwell Publishers Ltd., Oxford, GB).
- Kolaczyk, E. D. (2017, June 5). *Topics at the frontier of statistics and network analysis: (re)visiting the foundations* (E. C. Wit, Ed.). Cambridge University Press. <https://doi.org/10.1017/9781108290159>
- Kolaczyk, E. D., & Csárdi, G. (2020a). *Statistical analysis of network data with R* (R. Gentleman, K. Hornik, & G. Parmigiani, Eds.; 2nd ed.). Springer. <https://doi.org/10.1007/978-3-030-44129-6> (Original work published 2014)

- Kolaczyk, E. D., & Csárdi, G. (2020b, June 3). Statistical models for network graphs. In R. Gentleman, K. Hornik, & G. Parmigiani (Eds.), *Statistical analysis of network data with R* (2nd ed., pp. 87–113). Springer. [https://doi.org/10.1007/978-3-030-44129-6\\_6](https://doi.org/10.1007/978-3-030-44129-6_6) (Original work published 2014)
- König, D. (1936). *Theorie der endlichen und unendlichen Graphen: Kombinatorische topologie der streckenkomplexe*. Akademische Verlagsgesellschaft.
- Krivitsky, P.N., & Handcock, M.S. (2013). A separable model for dynamic networks. *Journal of the Royal Statistical Society B: Statistical Methodology*, 76(1), 29–46. <https://doi.org/10.1111/rssb.12014>
- Küchler, U. (1982). Exponential families of Markov processes: Part I. general results. *Series Statistics*, 13(1), 57–69. <https://doi.org/10.1080/02331888208801628>
- Küchler, U., & Sørensen, M. (1997). *Exponential families of stochastic processes*. Springer. <https://doi.org/10.1007/b98954>
- Küchler, U., & Sørensen, M. (1998). On exponential families of markov processes. *Journal of Statistical Planning and Inference*, 66(1), 3–19.
- Lahtonen, J. <https://math.stackexchange.com/users/11619/jyrki-lahtonen>. (2011, October 26). *Subspace generated by permutations of a vector in a vector space*. Mathematics Stack Exchange. Retrieved September 28, 2017, from <https://math.stackexchange.com/q/75665>
- Latouche, P., Robin, S., & Ouadah, S. (2018). Goodness of fit of logistic regression models for random graphs. *Journal of Computational and Graphical Statistics*, 27(1), 98–109. <https://doi.org/10.1080/10618600.2017.1349663>
- Laurent, S. (2010–April 15). Further comments on the representation problem for stationary processes. *Statistics & Probability Letters*, 80, 592–596. <https://doi.org/10.1016/j.spl.2009.12.015>
- Lauritzen, S. L., Rinaldo, A., & Sadeghi, K. (2018). Random networks, graphical models and exchangeability. *Journal of the Royal Statistical Society B: Statistical Methodology*, 80(3), 481–508. <https://doi.org/10.1111/rssb.12266>
- Lauritzen, S. L., Rinaldo, A., & Sadeghi, K. (2019). On exchangeability in network models. *Journal of Algebraic Statistics*, 10(1). Issue in honor of S. E. Fienberg, 85–114. <https://doi.org/10.18409/jas.v10i1.73>
- Lehmann, D., Müller, R., & Sandholm, T. (2006). The winner determination problem [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 297–317). MIT Press.

- Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer. <https://doi.org/10.1007/b98854>
- Lei, J. (2016). A goodness-of-fit test for stochastic block models. *Annals of Statistics*, *44*(1), 401–424. <https://doi.org/10.1214/15-AOS1370>
- Levin, D. A., & Peres, Y. (2017). *Markov chains and mixing times* (E. L. Wilmer, J. G. Propp, & D. B. Wilson, Collaborators; 2nd ed.). American Mathematical Society. Retrieved March 5, 2018, from <http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf>. URL is for the 1st ed.
- Li, W., Ahn, S., & Welling, M. (2016, May). Scalable MCMC for mixed membership stochastic blockmodels. In A. Gretton & C. C. Robert (Eds.), *Proceedings of the 19th international conference on artificial intelligence and statistics* (pp. 723–731). Journal of Machine Learning Research. Retrieved September 4, 2021, from <https://proceedings.mlr.press/v51/li16d.html>
- Licensing Framework for Mobile Broadband Services (MBS) — 700 MHz Band Notice DGSA-001-13, Canada (2013, March 7). Retrieved September 13, 2021, from <https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/sf10583.html#AnnexB>
- Lindsey, J. K. (2004, August 2). *The statistical analysis of stochastic processes in time*. Cambridge University Press. Retrieved February 9, 2018, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.4796&rep=rep1&type=pdf>
- Berkelaar, M., Eikland, K., & Notebaert, P. (2020, December 31). *lp\_solve* (comp. software; Version 5.5.2.11). Retrieved July 28, 2021, from <http://lpsolve.sourceforge.net/5.5/>
- Maldoom, D. (2007, December). *Winner determination and second pricing algorithms for combinatorial clock auctions* (DotEcon Discussion Paper No. 07/01). DotEcon Ltd. London. Retrieved July 26, 2021, from <https://www.dotecon.com/assets/images/dp0701.pdf>
- Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of majorization and its applications* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-68276-1>
- Minoux, M. (1986). *Programmation mathématique: Théorie et algorithmes* [Mathematical programming: Theory and algorithms] (S. Vajda, Trans.). Wiley-Interscience. (Original work published 1983)
- Mitrofanova, N. M. (1971). Families of transient densities in Markov chains admitting nontrivial sufficient statistics (I. G. Petrovskii & S. M. Nikol'skii, Eds.; S. Kotz, Trans.) [UDC 519.73]. *Proceedings of the Steklov Institute of Mathematics*, *104*. Studies in Mathematical Statistics, 219–228. (Original work published 1968)

- Muller, J.-M., Brisebarre, N., de Dinechin, F., Jeannerod, C.-P., Lefèvre, V., Melquiond, G., Revol, N., Stehlé, D., & Torres, S. (2010). *Handbook of floating-point arithmetic*. Birkhäuser. <https://doi.org/10.1007/978-0-8176-4705-6>
- Müller, R. (2006). Tractable cases of the winner determination problem [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 319–336). MIT Press.
- Nagaoka, H. (2005–November 22). The exponential family of Markov chains and its information geometry. *Proceedings of the 28th Symposium on Information Theory and its Applications*, 601–604. Retrieved February 23, 2018, from <http://www.ieice.org/ess/sita/SITA2005/advance/>. I cannot locate the *Proceedings* itself, but the arXiv has the article and the link lists the corresponding talk at the symposium.
- Nemhauser, G. L., & Wolsey, L. A. (1999). *Integer and combinatorial optimization*. Wiley-Interscience.
- Nielsen, F., & Garcia, V. (2011, May 13). *Statistical exponential families: A digest with flash cards*.
- Nisan, N. (2006). Bidding languages for combinatorial auctions [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 215–231). MIT Press.
- NumPy* (comp. software; Version 1.21.0). (2021, June 22). Retrieved September 3, 2021, from <https://numpy.org>. See also Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Wireless Telegraphy (Licence Award) Regulations 2012 (SI 2012/2817), GB (2012, November 9). Retrieved September 17, 2021, from <https://www.legislation.gov.uk/uksi/2012/2817/contents/made>
- Office of Communications. (2013a, March 1). *800 MHz & 2.6 GHz combined award*. GB. Retrieved August 2, 2021, from <https://www.ofcom.org.uk/spectrum/spectrum-management/spectrum-awards/awards-archive/800mhz-2.6ghz>
- Office of Communications. (2013b, March 15). *Auction data: Details of bids made in the auction* (Zip archive) [800\_2.6\_auction\_bid\_data\_files.zip]. Zip archive. Retrieved October 11, 2021, from [http://static.ofcom.org.uk/static/spectrum/800\\_2.6\\_auction\\_bid\\_data\\_files.zip](http://static.ofcom.org.uk/static/spectrum/800_2.6_auction_bid_data_files.zip). Available from *800 MHz & 2.6 GHz combined award*. (2013, March 1). GB. Retrieved August 2, 2021, from <https://www.ofcom.org.uk/spectrum/spectrum-management/spectrum-awards/awards-archive/800mhz-2.6ghz>

- Ogawa, M., Hara, H., & Takemura, A. (2013). Graver basis for an undirected graph and its application to testing the beta model of random graphs. *Annals of the Institute of Statistical Mathematics*, 65(1), 191–212. <https://doi.org/10.1007/s10463-012-0367-8>
- Ogryczak, W. (1988). The simplex method is not always well behaved. *Linear Algebra and Its Applications*, 109, 41–57. [https://doi.org/10.1016/0024-3795\(88\)90197-8](https://doi.org/10.1016/0024-3795(88)90197-8)  
Accepted 2 November 1987, Available online 15 July 2002. Submitted by Gene H. Golub
- Øksendal, B. (2003). *Stochastic differential equations: An introduction with applications* (6th ed.) [5th corrected printing 2010]. Springer. <https://doi.org/10.1007/978-3-642-14394-6>
- O'Neill, B. (2020). The classical occupancy distribution: Computation and approximation. *The American Statistician*. <https://doi.org/10.1080/00031305.2019.1699445>
- Park, J., & Newman, M. E. J. (2004). Solution of the two-star model of a network. *Physical Review E*, 70(6). <https://doi.org/10.1103/PhysRevE.70.066146>
- Parkes, D. C. (2006). Iterative combinatorial auctions [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 41–77). MIT Press.
- Pekeč, A., & Rothkopf, M. H. (2003). Combinatorial auction design. *Management Science*, 49(11), 1485–1503. <https://doi.org/10.1287/mnsc.49.11.1485.20585>
- Pekeč, A., & Rothkopf, M. H. (2006). Noncomputational approaches to mitigating computational problems in combinatorial auctions [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 395–411). MIT Press.
- Petrović, S. (2015, January 11). A survey of discrete methods in (algebraic) statistics for networks. In H. A. Harrington, M. Omar, & M. Wright (Eds.), *Algebraic and geometric methods in discrete mathematics* (pp. 251 sqq.). American Mathematical Society. <https://doi.org/10.1090/conm/685>
- Petrović, S., Rinaldo, A., & Fienberg, S. E. (2010). Algebraic statistics for a directed random graph model with reciprocation. In M. A. G. Viana & H. P. Wynn (Eds.), *Algebraic methods in statistics and probability II* (pp. 261–283). American Mathematical Society. <https://doi.org/10.1090/conm/516>
- Power Auctions, LLC. (2019, January 18). *Mathematical formulations for winner and price determination for the combinatorial clock auction in the 600 MHz band* (tech. rep.). Innovation, Science and Economic Development Canada. <http://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/sf11449.html>



- Pratelli, L. (1990). Sur le lemme de mesurabilité de Doob. In J. Azéma, P. A. Meyer, & M. Yor (Eds.), *Séminaire de probabilités. Vol. 24. Séminaire de probabilités XXIV 1988/89* (pp. 46–51). Springer. <https://doi.org/10.1007/BFb0083752>. For an English summary, see Dominique R.F. <https://mathoverflow.net/users/90906/dominique-r-f>. (2017, March 7). *Does the Doob-Dynkin lemma hold for any measurable space that separates points?* MathOverflow. Retrieved April 3, 2018, from <https://mathoverflow.net/a/263994/39652>
- Python* (comp. software; Version 3.9.7). (2021, August 30). Retrieved September 3, 2021, from <https://www.python.org>
- Rassenti, S. J., Smith, V. L., & Bulfin, R. L. (1982). A combinatorial auction mechanism for airport time slot allocation. *The Bell Journal of Economics*, 13(2), 402–417. <https://doi.org/10.2307/3003463>
- Rastelli, R., Latouche, P., & Friel, N. (2018). Choosing the number of groups in a latent stochastic blockmodel for dynamic networks. *Network Science*, 6(4), 469–493. <https://doi.org/10.1017/nws.2018.19>
- Reimann, J. (2011, January). *Polish spaces* (lecture notes) [lecture no. 2]. Pennsylvania State University. Retrieved April 4, 2018, from [http://www.personal.psu.edu/jsr25/Spring\\_11/574\\_Sp11\\_Syllabus.html](http://www.personal.psu.edu/jsr25/Spring_11/574_Sp11_Syllabus.html)
- Rinaldo, A., Fienberg, S. E., & Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3, 446–484. <https://doi.org/10.1214/08-EJS350>
- Rinaldo, A., Petrović, S., & Fienberg, S. E. (2013). Maximum likelihood estimation in the  $\beta$ -model. *Annals of Statistics*, 41(3), 1085–1110. <https://doi.org/10.1214/12-AOS1078>
- Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). Springer-Verlag. <https://doi.org/10.1007/978-1-4757-4145-2>
- Robins, G. L., & Pattison, P. E. (2001). Random graph models for temporal processes in social networks. *Journal of Mathematical Sociology*, 25(1), 5–41. <https://doi.org/10.1080/0022250X.2001.9990243>
- Robinson, D. J. S. (2003). *An introduction to abstract algebra*. Walter de Gruyter. Retrieved April 2, 2018, from <https://p4mriunpat.files.wordpress.com/2011/10/derek-j-s-robinson-an-introduction-to-abstract-algebra.pdf>
- Rosenblatt, M. (1959). Stationary processes as shifts of functions of independent random variables. *Journal of Mathematics and Mechanics*, 8(5), 665–681.

- Rosenblatt, M. (1960). Stationary Markov chains and independent random variables. *Journal of Mathematics and Mechanics*, 9(6), 945–949. Corrected in Addendum to “Stationary Markov chains and independent random variables” M. Rosenblatt, *Journal of Mathematics and Mechanics*, 1960, p. 945. (1962). *Journal of Mathematics and Mechanics*, 11(2), 317.
- Rosenblatt, M. (1963). The representation of a class of two state stationary processes in terms of independent random variables. *Journal of Mathematics and Mechanics*, 12(5), 721–730.
- Rubshtein, B.-Z. (2004, June 3). *On a class of one-sided Markov shifts*.
- Rudin, W. (1976). *Principles of mathematical analysis* (A. A. Arthur & S. L. Langman, Eds.; 3rd ed.). McGraw-Hill.
- Sandholm, T. (2006). Optimal winner determination algorithms [First MIT Press paperback edition, 2010]. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 337–368). MIT Press.
- Sandholm, T., Suri, S., Gilpin, A., & Levine, D. (2005). CABOB: A fast optimal algorithm for winner determination in combinatorial auctions. *Management Science*, 51(3), 374–. <https://doi.org/10.1287/mnsc.1040.0336>
- Schwartz, W. K. (2016, October 21). *The broadcaster-repacking problem*.
- Schwartz, W. K. (2019, October 22). *The lexicographic winner determination problem (WDP) for real world combinatorial auctions: Tie breaking, side constraints, and dynamic programming* (poster No. 48). Retrieved October 1, 2021, from <https://www.abstractsonline.com/pp8/#!/6818/presentation/12274>. Presented at the (2019, October 23). Retrieved October 1, 2021, from <http://meetings2.informs.org/wordpress/seattle2019>
- Schwartz, W. K., Ingraham, A. T., Ravi, S., & Sojourner, C. T. (2019, October 22). *Explaining and validating second prices for combinatorial auctions* (presentation No. 2). Retrieved October 1, 2021, from <https://www.abstractsonline.com/pp8/#!/6818/presentation/12350>. Presented at the (2019, October 23). Retrieved October 1, 2021, from <http://meetings2.informs.org/wordpress/seattle2019>
- Schwartz, W. K., Petrović, S., & Kaul, H. (2021, August 12). *Longitudinal network models and permutation-uniform Markov chains* [Manuscript submitted for publication].
- Shalizi, C. R., & Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *Annals of Statistics*, 41(2), 508–535. <https://doi.org/10.1214/12-AOS1044>
- Sharia, T. (2007). Rate of convergence in recursive parameter estimation procedures. *Georgian Mathematical Journal*, 14(4). <https://doi.org/10.1515/GMJ.2007.721>

- Sharia, T. (2010). Recursive parameter estimation: Asymptotic expansion. *Annals of the Institute of Statistical Mathematics*, 62(2), 343–362. <https://doi.org/10.1007/s10463-008-0179-z>
- Shiryaev, A. N. (2016). *Probability-1* (R. P. Boas & D. M. Chibisov, Trans.; 3rd ed.). Springer. <https://doi.org/10.1007/978-0-387-72206-1>
- Silvapulle, M. J. (1996). A test in the presence of nuisance parameters. *Journal of the American Statistical Association*, 91(436), 1690–1693. <https://doi.org/10.2307/2291597>
- Smith, J. D. H., & Romanowska, A. B. (1999). *Post-modern algebra*. John Wiley & Sons.
- Smith Institute. (2013, March 15). *Smith Institute to Ofcom auction team*. Retrieved October 12, 2021, from [https://www.ofcom.org.uk/\\_\\_data/assets/pdf\\_file/0022/92461/Letters-from-Smith-Institute-to-Ofcom.pdf](https://www.ofcom.org.uk/__data/assets/pdf_file/0022/92461/Letters-from-Smith-Institute-to-Ofcom.pdf). Available from Office of Communications. (2013, March 1). *800 MHz & 2.6 GHz combined award*. GB. Retrieved August 2, 2021, from <https://www.ofcom.org.uk/spectrum/spectrum-management/spectrum-awards/awards-archive/800mhz-2.6ghz>
- Sniedovich, M. (1992). *Dynamic programming*. Marcel Dekker, Inc.
- Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, 31, 361–395. <https://doi.org/10.1111/0081-1750.00099>
- Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1), 99–153. <https://doi.org/10.1111/j.1467-9531.2006.00176.x>
- Sørensen, A. B., & Hallinan, M. T. (1976). A stochastic model for change in group structure. *Sociological Review*, 24, 143–166. <https://doi.org/10.1111/j.1467-954X.1976.tb00050.x>
- Stefanov, V. T. (1984). Efficient sequential estimation in finite-state markov processes. *Stochastics*, 11(3-4), 291–300.
- Stefanov, V. T. (1991). Noncurved exponential families associated with observations over finite state markov chains. *Scandinavian Journal of Statistics*, 18(4), 353–356.
- Stefanov, V. T. (1995). Explicit limit results for minimal sufficient statistics and maximum likelihood estimators in some markov processes: Exponential families approach. *Annals of Statistics*, 23(4), 1073–1101. <https://doi.org/10.1214/aos/1176324699>
- Storer, B. E., & Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association*, 85(409), 146–155. <https://doi.org/10.2307/2289537>
- Takesaki, M. (1979). *Theory of operator algebras I*. Springer. Retrieved April 4, 2018, from <https://link.springer.com/content/pdf/bbm%3A978-1-4612-6188-9%2F1.pdf>

- Tao, T. (2011). *An introduction to measure theory*. American Mathematical Society. Retrieved August 2, 2021, from <https://terrytao.wordpress.com/books/an-introduction-to-measure-theory/>
- Taraldsen, G. (2018, January 3). *Optimal learning from the Doob-Dynkin lemma*.
- Technical, Policy and Licensing Framework for Spectrum in the 600 MHz Band Notice SLPB-002-18, Canada (2018, March 28). Retrieved July 26, 2021, from <https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/sf11374.html#sC>
- Tennenholtz, M. (2000). Some tractable combinatorial auctions. In H. Kautz & B. Porter (Organizers), *Proceedings of the seventeenth national conference on artificial intelligence* (pp. 98–103). Association for the Advancement of Artificial Intelligence. Retrieved October 12, 2021, from <https://www.aaai.org/Papers/AAAI/2000/AAAI00-015.pdf>
- Trefethen, L. N., & Bau, D., III. (1997). *Numerical linear algebra*. Society for Industrial; Applied Mathematics.
- Tsumura, Y. (2017, October 20). *Exponential functions are linearly independent*. Problems in Mathematics. Retrieved February 28, 2018, from <https://yutsumura.com/exponential-functions-are-linearly-independent/>
- user1551 <https://math.stackexchange.com/users/1551/user1551>. (2015, February 2). *Characterize stochastic matrices such that max singular value is less or equal one*. Mathematics Stack Exchange. Retrieved October 5, 2017, from <https://math.stackexchange.com/q/1129977>
- van Hoesel, S., & Müller, R. (2001). Optimization in electronic markets: Examples in combinatorial auctions. *Netnomics*, 3(1), 23–33. <https://doi.org/10.1023/A:1009940607600>
- Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1), 8–37. <https://doi.org/10.2307/2977633>
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.
- Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1, 1–305. <https://doi.org/10.1561/2200000001>
- Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, 3(2), 295–312. Retrieved September 7, 2021, from <http://www3.stat.sinica.edu.tw/statistica/j3n2/j3n23/j3n23.htm>

- Wasserman, S. S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75(370), 280–294. <https://doi.org/10.2307/2287447>
- Wasserman, S. S., & Iacobucci, D. (1988). Sequential social network data. *Psychometrika*, 53(2), 261–282. <https://doi.org/10.1007/BF02294137>
- Wooldridge, J. M. (2012, September 26). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning. Retrieved January 10, 2018, from [http://economics.ut.ac.ir/documents/3030266/14100645/Jeffrey\\_M.\\_Wooldridge\\_Introductory\\_Econometrics\\_A\\_Modern\\_Approach\\_\\_2012.pdf](http://economics.ut.ac.ir/documents/3030266/14100645/Jeffrey_M._Wooldridge_Introductory_Econometrics_A_Modern_Approach__2012.pdf) (Original work published 1999, August)
- Wu, W. B., & Mielniczuk, J. (2010). A new look at measuring dependence. In P. Doukhan, G. Lang, D. Surgailis, & G. Teyssi re (Eds.), *Dependence in probability and statistics* (pp. 123–142). Springer. [https://doi.org/10.1007/978-3-642-14104-1\\_7](https://doi.org/10.1007/978-3-642-14104-1_7)
- Yan, T., Jiang, B., Fienberg, S. E., & Leng, C. (2018). Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association*, 1–12. <https://doi.org/10.1080/01621459.2018.1448829>
- Yano, K., & Yasutomi, K. (2011a, May 5). *Random walk in a finite directed graph subject to a synchronizing road coloring*.
- Yano, K., & Yasutomi, K. (2011b). Realizations of an ergodic Markov chain as a random walk subject to a synchronizing road coloring. *Journal of Applied Probability*, 48(3), 766–777. <https://doi.org/10.1017/S0021900200008305>
- Ycart, B. (1988). A characteristic property of linear growth birth and death processes. *Sankhy  Series A*, 50(2), 184–189.
- Ycart, B. (1989). Markov processes and exponential families on a finite set. *Statistics & Probability Letters*, 8(4), 371–376. [https://doi.org/10.1016/0167-7152\(89\)90046-1](https://doi.org/10.1016/0167-7152(89)90046-1)
- Ycart, B. (1992a). Integer valued markov processes and exponential families. *Statistics & Probability Letters*, 14(1), 71–78. [https://doi.org/10.1016/0167-7152\(92\)90213-O](https://doi.org/10.1016/0167-7152(92)90213-O)
- Ycart, B. (1992b). Markov processes and exponential families. *Stochastic Processes and their Applications*, 41(2), 203–214. [https://doi.org/10.1016/0304-4149\(92\)90121-6](https://doi.org/10.1016/0304-4149(92)90121-6)
- Zhong, L. (2015, August 1). *Truncated stochastic approximation with moving bounds* (Doctoral dissertation). Dept. of Mathematics, Royal Holloway, University of London. Retrieved February 9, 2018, from [https://pure.royalholloway.ac.uk/portal/en/publications/truncated-stochastic-approximation-with-moving-bounds\(cee4d5d2-059d-467e-86e9-7e8798fb03ao\).html](https://pure.royalholloway.ac.uk/portal/en/publications/truncated-stochastic-approximation-with-moving-bounds(cee4d5d2-059d-467e-86e9-7e8798fb03ao).html)