# Separability as a Modeling Paradigm in Large Probabilistic Models

by

William J. Richoux

B.A., English, Case Western Reserve University (2004)
B.S., Systems & Control Engineering, Case Western Reserve University (2004)
M.S., Electrical Engineering, Case Western Reserve University (2005)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2011

Author ..............................................................
Department of Electrical Engineering and Computer Science
October 15, 2010

Certified by ..............................................................
George C. Verghese
Professor of Electrical Engineering
Thesis Supervisor

Accepted by ..............................................................
Terry P. Orlando
Professor of Electrical Engineering and Computer Science,
Chair, Committee for Graduate Students

# Separability as a Modeling Paradigm in Large Probabilistic Models

by

William J. Richoux

Submitted to the Department of Electrical Engineering and Computer Science
on October 15, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

Many interesting stochastic models can be formulated as finite-state vector Markov processes, with a state characterized by the values of a collection of random variables. In general, such models suffer from the curse of dimensionality: the size of the state space grows exponentially with the number of underlying random variables, thereby precluding conventional modeling and analysis. A potential cure to this curse is to work with models that allow the propagation of partial information, e.g. marginal distributions, expectations, higher-moments, or cross-correlations, as derived from the joint distribution for the network state.

This thesis develops and rigorously investigates the notion of *separability*, associated with structure in probabilistic models that permits exact propagation of partial information. We show that when partial information can be propagated exactly, it can be done so linearly. The matrices for propagating such partial information share many valuable spectral relationships with the underlying transition matrix of the Markov chain. Separability can be understood from the perspective of subspace invariance in linear systems, though it relates to invariance in a non-standard way. We analyze the asymptotic generality—as the number of random variables becomes large—of some special cases of separability that permit the propagation of marginal distributions.

Within this discussion of separability, we introduce the generalized influence model, which incorporates as special cases two prominent models permitting the propagation of marginal distributions: the influence model and Markov chains on permutations (the symmetric group). The thesis proposes a potentially tractable solution to learning informative model parameters, and illustrates many advantageous properties of the estimator under the assumption of separability. Lastly, we illustrate separability in the general setting without any notion of time-homogeneity, and discuss potential benefits for inference in special cases.

Thesis Supervisor: George C. Verghese
Title: Professor of Electrical Engineering

## Acknowledgments

I am deeply indebted to my advisor, George Verghese. His ability to offer immediate insight has been indispensable to my research, and his attention to detail has always improved its completeness and clarity. George has been exceedingly patient throughout our years working together, assuring that the years at MIT have been both fun and intellectually satisfying. I have been fortunate to have as an advisor someone who is both a terrific teacher and person.

I would like to thank my thesis committee, Devavrat Shah and John Tsitsiklis. Through their careful listening and reading, they have offered many valuable suggestions that have substantially improved the thesis.

Some of my favorite memories at MIT are experiences as a teaching assistant with John Wyatt, John Tsitsiklis, and Greg Wornell. I has been a pleasure to work alongside such amazing people.

There are countless people at MIT that have been immensely helpful. Fellow members of George's extended group, Thomas, Tushar, Victor, Laura, Carlos, Julien, and Faisal have offered countless hours of their time as both colleagues and friends. Robert, Brooke, Al, Warit, and Brandon have been great friends from lab. I must thank Demba for being instrumental in preparing me for interviews; Paco deserves many thanks for the hours we spent together writing. I will miss the camaraderie of my friends from intramural sports, especially Dave, Mike, Peter, and Jesse. I would especially like to thank Mark. I am fortunate to have spent high school, college, and graduate school together with such a great friend.

My family has been unconditionally supportive, especially my sisters, parents, and grandparents. I know that my in-laws have been as supportive as native Texans could be of a Massachusetts Yankee. My wife Cara deserves the greatest thank you, for her encouragement, patience, and love. I could not have achieved this without her.

This research was sponsored by the National Science Foundation under an NSF Graduate Research Fellowship and received additional support from the EECS department at MIT.

# Contents

# Chapter 1

# Introduction

This thesis investigates at a rigorous level the exact propagation of *partial information* in Markov chains, i.e., incomplete descriptions of the distribution for the Markov chain's state. As Markov chains for even moderately large collections of dependent stochastic processes are in general intractable, there is a natural inclination to consider propagating lower-order partial information. Markov chains describing large interconnected systems are common in biology, network theory, economics, and quantitative finance. All these areas could potentially benefit from a thorough analysis of exact propagation of partial information in Markov chains.

## 1.1 Motivation

Markov chains provide a rich, tractable mathematical framework to analyze discrete-time stochastic processes satisfying appropriate assumptions [1, 2]. One can determine many useful properties of a Markov chain from its transition matrix, such as its rate of mixing and its steady-state probability distribution. In general, computation of such properties for chains with $\eta$ states require algorithms of $\mathcal{O}(\eta^3)$ complexity.

Modeling a collection or *network* of interacting finite-state stochastic processes as a single Markov chain in general suffers computational limitations. Because the number of states in the Markov chain grows exponentially with the number of underlying processes, conventional analysis of such a network does not scale favorably with the network's size. Consequently, using a general Markovian framework to model and analyze large stochastic networks has severe limitations.

These computational limitations, a consequence of the explosion in the size of the state space, can be clearly demonstrated by some tangible examples. Suppose one has a deck of $n = 52$ cards that is repeatedly shuffled. If there is some independent randomness to the mechanics of each shuffle, we can model the shuffling probabilistically. One can define the state of the deck as the positions of each of the 52 cards, i.e., each card's position may assume one of $m = 52$ values. Its stochasticity upon shuffling can be modeled as a Markov chain. This model, however, may be rather complicated to analyze. There are $52! > 10^{67}$ possible states for the deck, exceeding the number of atoms on earth! It would be impossible to track the exact state occupancy probabilities over the course of several shuffles that fully randomize the deck.

Alternatively, for an example with more freedom in the modeling assumptions, consider the evolution of the weather. We focus on the daily weather in $n = 4$ different cities

(Pittsburgh, Toronto, Montreal, and Boston), and to maintain a simple model, the weather each day is classified as one of $m = 3$ possibilities: sun, clouds, or rain/snow. Assume that the dynamics of this weather model can be aggregately modeled as a single Markov chain, whose states correspond to unique weather combinations among the cities. Unless some weather combinations can be ruled out, the Markov chain will consist of $\eta = m^n = 81$ states. Double the number of cities to $n = 8$, and the number of possible states grows to $\eta = 6,561$. In general, analyzing such a network requires algorithms of $\mathcal{O}(m^{3n})$ complexity. The Markov chain not only requires storage that is exponential in $n$, but the computational complexity of analyzing such a chain is exponential in $n$. This paper seeks to identify constraints on the time-update structure of the network that enable analysis with algorithms that scale *polynomially* with $n$.

Many current research problems involve the analysis of network dynamics naturally modeled by large numbers of interacting stochastic processes. Models of this kind may be used to represent complex biological interactions [3], interactions among adjacent cells in a GSM/GPRS cellular communications network [4], or encounters among economic agents [5]. Some of these applications might benefit from stochastic network models whose analysis scales in a more friendly way with the number of underlying processes.

Returning to the shuffling example, the reader who is well-versed in group representation theory [6, 7], and in particular, representation theory of the symmetric group, may recall some powerful results regarding the analysis of shuffling a deck of cards [8]. Because the state of the Markov chain describing a deck's shuffling is a Markov chain on permutations, group representation theory for the symmetric group offers tractable techniques to analyze the performance of such chains, particularly, their convergence to steady-state. A product of this research is the notion that seven riffle shuffles fully randomizes a deck of cards [9].

The focus of [10] is how group representation theory for the symmetric group permits the propagation of marginal distributions. To explain what we mean by marginal distributions, consider the state of the deck of cards as $n$ random variables, the positions of each of the cards. The Markov chain representation permits one to propagate the joint distribution on these $n$ random variables, the vector process, as a probability vector. Representation theory for the symmetric group [11, 9] shows that the collection of $n$ univariate marginal distributions for each random variable can be propagated linearly, using matrices with $O(n)$ size. Computationally, this may be tractable for typical values for the number of cards in a deck. In addition, the collection of bivariate marginal distributions can be propagated in time with matrices of size $O(n^2)$. This propagation similarly holds for $r$th-order marginal distributions, for all values of $r$. This special property of Markov chains that maintain permutations offers an array of intermediate methods of analysis requiring computations that are only polynomial in $n$, as an alternative to the full analysis of the Markov chain and its state occupancy vector with a length that is exponential in $n$.

The influence model (IM) [12, 13] offers a parametric probabilistic model that, like the Markov chains on permutations, permits one to propagate $r$th-order marginal distributions. The IM also has the advantage of offering a compact representation requiring storage order $O(n^2m^2)$, and moreover, generating a realization of length $t$ requires computation of complexity only $O(nm^2t)$. These properties allow IMs to offer tractable models for networks when $n$ is large. Although IMs can illustrate many interesting behaviors, their modeling generality is restricted by requiring independent updates of the random variables.

The moment-linear stochastic systems (MLSS) of [14, 15] consider general parametric models where moments propagate linearly. 'Moments' include any linear function of the joint probability vector, including marginal distributions. Many interesting examples of

MLSS are illustrated within the aforementioned references. The flavor of MLSS is similar to many approaches in chemical kinetics [16, 17, 18, 19] that seek approximate methods to propagate moments, as an alternative to analyzing a Kolmogorov forward equation, the continuous-time analogue to a discrete time Markov chain's probability vector update.

What is missing in these analyses is a general theory to characterize and understand when particular moments can be propagated exactly. When one envisions the probability vector update equation of a finite state Markov chain as a discrete-time linear system, the notion of propagating moments, linear transformations of the state, suggests some kind of invariance. Invariant subspace theory for linear systems [20, 21] is well developed, providing an algebraic and geometric characterization of when dynamics remain isolated in subspaces, or effectively become isolated in the subspaces generated by the eigenvectors associated with the dominant eigenvalues. Lower-dimensional characterizations are made possible by subspace invariance. Invariance in Markov chains is a old idea: recurrent classes and stationary probability vectors are obvious examples. However, the idea of propagating moments exactly offers a new perspective on invariance. Rather than needing dynamics to be trapped in some subspace, what is necessary is that information orthogonal to the chosen moments be invariant.

Besides characterizing the models permitting the propagation of moments, it is equally important to have tractable representations of network models that could be used for simulations. A common assumption to ensure a tractable representation is to assume that transitions occur locally, thereby mandating a sparse transition matrix. However, there are other alternatives. Stochastic automata networks (SANs), [22, 23, 24, 25] offer a formalism for representing a transition matrix as a sum of simpler matrices. Each of these matrix terms can be represented as a sequence of Kronecker products of smaller matrices—and hence has a compact representation in memory, despite its size. When a transition matrix can be represented as a sum of a limited number of such simpler matrix terms, the SAN representation offers computational gains. It has also been shown that in some cases SANs can be analyzed tractably; in [26], iterative algorithms are developed to calculate the steady-state probability distribution. The IM also offers a solution to the problem of tractable representation of stochastic networks.

## 1.2 Thesis outline

The thesis is organized as follows. In Chapter 2, the notation used to represent networks of stochastic automata is introduced, and the idea of propagating moments is formalized in in the language of *separability*. We connect the notion of separability to many previously proposed models in the literature. A general constraint-based approach is developed. In a sense, separability continues the story of MLSS, by offering MLSS an algebraic characterization. In the language of algebra, MLSS proposed the parametrization, and this thesis identifies the equivalent constraints.

We rigorously develop several equivalent perspectives on separability in Chapter 3, and develop some measures of how general separable models are. The algebraic structure of separability is revealed, as is the new notion of invariance for Markov chains. We particularly focus on what we call the canonical examples of separability: when marginal distributions can be propagated, as is the case for both the IM and Markov chains over permutations. We reveal some interesting asymptotic properties of separable models when $n$ is large.

In Chapter 4, we introduce the generalized influence model (GIM), which generalizes both the IM and Markov chains over permutations. Several examples of GIMs are discussed. We derive the GIM's separability, and discuss some of its algebraic and geometric properties.

We consider the problem of parameter learning in Chapter 5. We discuss some illuminating perspectives on general parameter learning of Markov chains. We propose a tractable parameter learning algorithm under the assumptions of separability, and illustrate some of its interesting properties with computational examples.

The final chapter extends the notion of separability to cases without an underlying time-homogeneous Markov chain. We illustrate the connection of particular instances of separability to the familiar notions of Fisher sufficiency [27] and conditional independence. Inference on Bayesian networks with the additional structure of separability is discussed.

## 1.3 Contributions

- This thesis provides a rigorous characterization of when partial information, in the form of moments, can be propagated exactly in a Markov chain, i.e., when Markov chains exhibit separability. In doing so, this thesis offers a constraint-based analysis of moment-linear stochastic systems.

- The thesis algebraically connects the dynamics of partial information propagation and the dynamics of the underlying Markov chain. We build a general framework to analyze invariance in Markov chains, offering a new perspective on invariance in linear systems.

- We offer measures of generality for many important examples exhibiting the propagation of partial information, and consider the asymptotic properties of these measures of generality.

- The thesis presents a parametric model exhibiting the propagation of partial information, generalizing both the influence model and Markov chains on permutations. In many respects, it permits both of the latter models to be understood, developed, and analyzed more simply.

- We offer an algorithm for parameter learning for networks that permits the propagation of partial information.

- We investigate the propagation of partial information in spatial models without a notion of time-homogeneity. We make abstract connections between separability and conditional independence, as well as Fisher sufficiency. We illustrate potential benefits of separability for inference.

# Chapter 2

# Introducing Separability

In this chapter, we formally introduce separability. Before doing so, we introduce our perspective and its notation for networks of stochastic automata that allows us to characterize various kinds of separability in a general, yet succinct, mathematical framework. As will be shown, the notion of separability is effectively equivalent to the moment-linear stochastic systems (MLSS) of [14, 15] that are derived from discrete-time, finite-state Markov chains. However, we introduce separability (and thus MLSS) from a new perspective that becomes the essential foundation from which many of the results of subsequent chapters are derived. In effect, we develop a constraint-based characterization, the algebraic analogue to the parametric description of MLSS. We demonstrate several equivalent characterizations and offer comparative insight into instances of this special class of stochastic models. We hope this aids in subsequent identification of systems satisfying these special properties.

We introduce several notions of separability, including several examples that have been proposed and discussed in previously published work. Many of these examples will be referenced in subsequent chapters, especially, the instances that we refer to as the canonical examples of separability. The chapter also introduces some algebraic characterizations that will be expanded upon in subsequent chapters.

## 2.1 General setup for a network of automata

Our analysis involves a collection of $n$ interacting finite-state, discrete-time stochastic processes $\mathsf{x}_j[t]$, for $j = 1, \ldots, n$, which we refer to individually as automata and collectively as a network. At each time $t$, automaton $j$ expresses a particular value $\mathsf{x}_j[t]$ drawn from a finite set of possible values $\mathcal{X}_j$ (its alphabet); this value will be referred to as its *status* at that time. The number of possible statuses for automaton $j$ is denoted as $m_j \triangleq |\mathcal{X}_j|$, which is the $j$th element of the vector $\boldsymbol{m} \triangleq (m_1, m_2, \ldots, m_n)$. If $\boldsymbol{m}$ is a vector with constant entries, $m$ refers to any of its identical entries. In our weather example of Chapter 1, each automaton corresponds to a different city, and $m = 3$ as $\boldsymbol{m} = (3, 3, 3, 3)$. Typically we assume that $m_j > 1$, as otherwise we have degenerate random variables. We will later consider the complexity of algorithms for a network of stochastic automata as $n$ grows. For such cases, it will be assumed that $m_j \leq \overline{m}$ for all $j$ and some finite $\overline{m}$.

We assume that the automata collectively evolve as a time-homogeneous Markov chain, that is, the vector process $\mathsf{x}[t] = (\mathsf{x}_1[t], \mathsf{x}_2[t], \ldots, \mathsf{x}_n[t])$ is time-homogeneous and Markovian (a listing of scalars separated by commas and enclosed by parentheses, e.g., $(0, 1, 0)$, indicates a column vector). Under these assumptions, the probability mass function (PMF) of the network state $\mathsf{x}[t]$ and the underlying Markov chain's transition matrix are of interest. As our state is defined in terms of several random variables, we will also wish to represent the marginal PMF for a single automaton's status, as well as the marginal PMF for the statuses of a subset of automata. Our aim is to represent all PMFs as probability vectors, not as tables or matrices. This standard framework for representing all PMFs as probability vectors will simplify the development and description of general forms of separability. To represent any PMF as a probability vector, we employ Kronecker products and expectations of indicator vectors. Equivalently, we will show how all such probability vectors can be obtained via special marginalizing matrices that operate on the probability vector representation of the PMF for the network state. We begin by defining indicator vectors for random variables.

### 2.1.1 Automaton status indicator vector $\mathsf{s}_j[t]$

For a random variable $\mathsf{x}$ taking values in alphabet $\mathcal{X}$, its indicator vector is denoted as $\mathsf{s}_\mathsf{x}$, which is a $|\mathcal{X}| \times 1$ vector consisting of all 0s except for a single 1. When $\mathsf{s}_\mathsf{x} = e_k$, the $k$th standard unit vector (with a 1 in its $k$th position and 0s elsewhere), the random variable $\mathsf{x}$ assumes the $k$th element in its alphabet $\mathcal{X}$. The expected value of $\mathsf{s}_\mathsf{x}$ is a probability vector representation for the PMF of $\mathsf{x}$.

In the case of a network of stochastic automata, a simplified notation is employed for each automaton's indicator vector. The status of automaton $j$ at time $t$ is represented by a *status indicator vector* $\mathsf{s}_j[t]$, which is shorthand notation for $\mathsf{s}_{\mathsf{x}_j[t]}$. Returning to the example of Chapter 1, $\mathsf{s}_1[t]$ is the random indicator vector corresponding to the particular weather of the first city, Pittsburgh, at time $t$, and would be defined as follows:

$$\mathsf{s}_1[t] = \begin{cases} (1,\ 0,\ 0) & \text{if Pittsburgh has sun at time } t \\ (0,\ 1,\ 0) & \text{if Pittsburgh has clouds at time } t \\ (0,\ 0,\ 1) & \text{if Pittsburgh has rain/snow at time } t \end{cases} \quad .$$

The expected value of the status indicator vector is a probability vector for the associated automaton being in its various statuses. If $\mathbb{E}[\,\mathsf{s}_1[t]\,] = (\alpha_1,\ \alpha_2,\ \alpha_3)$ in our weather example, then

$$\begin{aligned} \mathbb{P}\,(\text{Pittsburgh has sun at time } t) &= \alpha_1 \\ \mathbb{P}\,(\text{Pittsburgh has clouds at time } t) &= \alpha_2 \\ \mathbb{P}\,(\text{Pittsburgh has rain/snow at time } t) &= \alpha_3\ , \end{aligned}$$

where $\mathbb{E}[\,\cdot\,]$ denotes the expected value of a random variable/vector and $\mathbb{P}\,(\,\cdot\,)$ denotes the probability of an event.

### 2.1.2 Joint status indicator vector $\mathsf{s}_{\boldsymbol{j}}[t]$

Each automaton's random indicator vector indicates its status. Suppose one wants to consider the statuses of automata $j_1, j_2, \ldots, j_r$ together at time $t$. Define the vector $\mathsf{x}_{\boldsymbol{j}}[t] \triangleq (\mathsf{x}_{j_1}[t],\ \mathsf{x}_{j_2}[t], \ldots,\ \mathsf{x}_{j_r}[t])$ involving all such automata, characterized by its vector subscript $\boldsymbol{j} = (j_1, j_2, \ldots, j_r)$. An indicator vector for $\mathsf{x}_{\boldsymbol{j}}[t]$, i.e., $\mathsf{s}_{\mathsf{x}_{\boldsymbol{j}}[t]}$, will be denoted in shorthand form as $\mathsf{s}_{\boldsymbol{j}}[t]$, and can be defined as

$$\mathsf{s}_{\boldsymbol{j}}[t] \triangleq \mathsf{s}_{j_1}[t] \otimes \mathsf{s}_{j_2}[t] \otimes \ldots \otimes \mathsf{s}_{j_r}[t] \quad , \tag{2.1}$$

where $\otimes$ indicates the Kronecker product [28].[1] Such an indicator vector for several automata will be referred to as a *joint status indicator vector.*

Returning to our weather example for a concrete illustration, suppose that

$$\mathbf{s}_{(1,2)}[t] \triangleq \mathbf{s}_1 \otimes \mathbf{s}_2 = (0,1,0,0,0,0,0,0,0) \quad . \tag{2.2}$$

One can deduce from such a joint status indicator vector that Pittsburgh is sunny and Toronto is cloudy at time $t$, as

$$\begin{aligned} \mathbf{s}_1[t] &= (1,0,0) \\ \mathbf{s}_2[t] &= (0,1,0) \quad . \end{aligned} \tag{2.3}$$

Since the joint status indicator vector for automata $j_1, j_2, \ldots, j_r$ indicates the simultaneous statuses of $r$ automata, we say that the joint status indicator vector indicates a *joint status.* A joint status indicator vector involving $r$ automata is referred to as an *$r$th-order joint status indicator vector.*

The expected value of a joint status indicator vector is a probability vector representing the PMF for the joint status of the corresponding automata. For example, by representing the expected value of the joint status indicator vector for Pittsburgh and Toronto at time $t$ as

$$\mathbb{E}\left[\, \mathbf{s}_{(1,2)}[t] \,\right] = (\alpha_1,\, \alpha_2,\, \alpha_3,\, \alpha_4,\, \alpha_5,\, \alpha_6,\, \alpha_7,\, \alpha_8,\, \alpha_9) \quad , \tag{2.4}$$

we can conclude that the probability that both Toronto and Pittsburgh are simultaneously sunny is $\alpha_1$. Probabilities for all other simultaneous weather combinations appear as other entries of the probability vector. Thus, status vectors and Kronecker products permit one to define an indicator vector whose expected value represents the PMF for any desired $r$-tuple of automata at a given time instant. By using status indicator vectors and Kronecker products, we can compactly express the joint PMFs for any subset of our random variables as a probability vector: it is the expectation of the appropriate joint status indicator vector.

### 2.1.3 State indicator vector $\mathbf{s}_\mathbf{x}[t]$

A rather important joint status indicator vector is $\mathbf{s}_{\boldsymbol{j}}[t]$ for $\boldsymbol{j} = (1, 2, \ldots, n)$. As $\mathbf{x}[t]$ is Markovian, such a joint status indicator vector (denoted as $\mathbf{s}_\mathbf{x}[t]$ rather than $\mathbf{s}_{(1,\ldots,n)}[t]$) is an indicator vector for the network state. The length of $\mathbf{s}_\mathbf{x}[t]$ is $\eta \triangleq \prod_j m_j$; for the weather example, its length is $|\, \mathbf{s}_\mathbf{x}[t] \,| = 3^4 = 81$.

The expected value of the state indicator vector $\mathbf{s}_\mathbf{x}[t]$ is a probability vector for the joint PMF over the statuses of all automata at time $t$. As $\mathbf{x}[t]$ is assumed to be time-homogeneous and Markovian, there exists an $\eta \times \eta$ matrix $G$ that updates the state-occupancy probability vector of the underlying Markov chain, i.e.,

$$\mathbb{E}\left[\, \mathbf{s}_\mathbf{x}[t]' \,|\, \mathbf{s}_\mathbf{x}[\tau] \,\right] = \mathbf{s}_\mathbf{x}[\tau]' G^{t-\tau} \ , \tag{2.5}$$

for any $t > \tau$ (we use $'$ to denote the transpose of a vector or matrix). The row-stochastic matrix $G$ is referred to as the *joint transition matrix,* and its corresponding Markov chain as the *master Markov chain* or *underlying Markov chain.*

[1] Refer to Appendix A for a condensed introduction to Kronecker products.

### 2.1.4 Probability vectors represented via $\boldsymbol{\pi}_{\mathsf{x}}$ notation

Rather than representing each probability vector via an expectation of a (joint) status indicator vector, we will often represent PMFs using a more compact form,

$$\boldsymbol{\pi}_{\boldsymbol{j}}[t] \triangleq \mathbb{E}\left[\,\mathbf{s}_{\boldsymbol{j}}[t]\,\right] \quad , \tag{2.6}$$

where $\boldsymbol{j}$ may be either a scalar (when representing a PMF for a single automaton's status) or as it is written, a vector (when representing a PMF for a subset of automata's statuses). Oftentimes, the expectation in (2.6) is a conditional expectation, with $\boldsymbol{\pi}_{\mathbf{x}}[t]$ representing a conditional probability. When $j$ is a scalar, we refer to $\boldsymbol{\pi}_j[t]$ is a univariate PMF. When $\boldsymbol{j}$ is a vector of length $r$, for $r = 2$ we refer to $\boldsymbol{\pi}_{\boldsymbol{j}}[t]$ as a bivariate PMF, and for general $r > 2$ as an $r$th-order marginal PMF. When $\boldsymbol{j} = (1, 2, \ldots, n)$, as in the case of the state indicator vector, we will denote the $\eta \times 1$ probability vector that is propagated by the Markov chain's transition matrix as $\boldsymbol{\pi}_{\mathbf{x}}[t] \triangleq \boldsymbol{\pi}_{(1,\ldots,n)}[t]$.

In the case of Markov chains, such notation can be abused, as the probability vector $\boldsymbol{\pi}_{\mathbf{x}}[t]$ often represents conditional probabilities given some implicit event. This fact is easily overlooked, as the following example illustrates.

Suppose that we have a probabilistic description of a finite-state stochastic process $\mathsf{x}[t]$. Without making any assumptions regarding its behavior, we can always express the conditional PMF for $\mathsf{x}[t]$ given $\mathsf{x}[t-1]$ via some matrix $G_t$ and indicator vectors as follows:

$$\mathbb{E}\left[\,\mathbf{s}_{\mathsf{x}}[t]' \,|\, \mathsf{x}[t-1]\,\right] = \mathbf{s}_{\mathsf{x}}[t-1]' G_t \quad . \tag{2.7}$$

The $p$th row of the matrix $G_t$ therefore displays the conditional PMF for the current value of the stochastic process, given that the process assumes its $p$th value at the preceding time instant. Thus, $G_t$ is a *row-stochastic* matrix, i.e., the entries of each row are nonnegative and sum to 1. By total expectation applied to (2.7),

$$\mathbb{E}\left[\,\mathbf{s}_{\mathsf{x}}[t]'\,\right] = \mathbb{E}\left[\,\mathbf{s}_{\mathsf{x}}[t-1]'\,\right] G_t \quad . \tag{2.8}$$

If we assume that there exists a matrix $G = G_t$ for all $t$ (thereby assuming time-homogeneity), we would have

$$\boldsymbol{\pi}_{\mathsf{x}}[t]' = \boldsymbol{\pi}_{\mathsf{x}}[t-1]' G \tag{2.9}$$

where $\boldsymbol{\pi}_{\mathsf{x}}[t] \triangleq \mathbb{E}\left[\,\mathbf{s}_{\mathsf{x}}[t]\,\right]$. By repeatedly invoking (2.9), it would then follow that

$$\boldsymbol{\pi}_{\mathsf{x}}[t]' = \boldsymbol{\pi}_{\mathsf{x}}[\tau]' G^{t-\tau} \quad , \tag{2.10}$$

for all $t > \tau$. As (2.10) holds for all $t > \tau$, some may claim that it directly follows that $\mathsf{x}[t]$ must be Markovian: by (2.10) its state occupancy probabilities can be updated in a linear fashion. Obviously, this is not true: time-homogeneity does not imply Markovianity, in general. The confusion lies in the fact that in order to be Markovian, (2.10) must hold for *any* initial condition $\boldsymbol{\pi}_{\mathbf{x}}[\tau]$, with $\boldsymbol{\pi}_{\mathbf{x}}[t]$ understood to represent the conditional probability of the state given this initial condition. In other words, (2.10) must hold when $\boldsymbol{\pi}_{\mathbf{x}}[t]$ and $\boldsymbol{\pi}_{\mathbf{x}}[\tau]$ are interpreted as conditional probabilities given an event identifiable by time $\tau$, that is, an event in the $\sigma$-field generated by the stochastic process at or before time $\tau$. One often thinks of being able to vary the initial condition $\boldsymbol{\pi}_{\mathbf{x}}[\tau]$, with $\boldsymbol{\pi}_{\mathbf{x}}[t]$ being affected as a consequence; this idea can be more clearly stated as varying $\epsilon$ in the aforementioned $\sigma$-field, with $\boldsymbol{\pi}_{\mathbf{x}}[\tau] \triangleq \mathbb{E}\left[\,\mathbf{s}_{\mathbf{x}}[\tau] \,|\, \epsilon\,\right]$ and $\boldsymbol{\pi}_{\mathbf{x}}[t] \triangleq \mathbb{E}\left[\,\mathbf{s}_{\mathbf{x}}[t] \,|\, \epsilon\,\right]$.

Evidently, using notation like $\boldsymbol{\pi}_{\mathbf{x}}[\tau]$ and $\boldsymbol{\pi}_{\mathbf{x}}[t]$ for Markov chains can often be misleading, as the probability vectors often represent conditional probabilities. It is important to be mindful of such implicit conditioning events. In instances where we feel it is important to maintain full clarity, the precise notation involving conditional expectations of indicator vectors will be used, thereby permitting explicit tracking of conditioning events. However, in other cases, we will commit such notational abuses by not explicitly indicating an assumed conditioning event, and the reader must be mindful of this warning.

### 2.1.5 Marginalizing matrix $M_{\mathbf{s}_j}$

As marginalization is a linear operation, there exists a marginalizing matrix $M_{\mathbf{s}_j}$ that maps any valid probability vector for the network state, $\boldsymbol{\pi}_{\mathbf{x}}$, to the probability vector for the $j$th automaton's status, $\boldsymbol{\pi}_j$ (the seemingly peculiar subscript $_{\mathbf{s}_j}$ will make sense momentarily):

$$\boldsymbol{\pi}_j' = \boldsymbol{\pi}_{\mathbf{x}}' M_{\mathbf{s}_j} \quad . \tag{2.11}$$

Note that we have omitted a time index '$[t]$' on the probability vectors $\boldsymbol{\pi}_{\mathbf{x}}$ and $\boldsymbol{\pi}_j$ so as to simplify notation. We will commonly omit such time indices provided that all indicator vectors and PMFs in the expression share a common time index, and the expressed relationship holds irrespective of the specific time.

In essence, (2.11) states that the expected value of $\mathbf{s}_{\mathbf{x}}$ linearly maps to the expected value of $\mathbf{s}_j$, and hence such a mapping holds in the absence of expectations. Consequently,

$$\mathbf{s}_j' = \mathbf{s}_{\mathbf{x}}' M_{\mathbf{s}_j} \quad , \tag{2.12}$$

and the choice of subscript $_{\mathbf{s}_j}$ should now be clear—$M_{\mathbf{s}_j}$ maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to the status indicator vector $\mathbf{s}_j$. The structure of $M_{\mathbf{s}_j}$ is evident from (2.12): the $p$th row of $M_{\mathbf{s}_j}$ must be the indicator vector for automaton $j$ when the network is in its $p$th state. As the state indicator vector $\mathbf{s}_{\mathbf{x}}$ is defined as a Kronecker product of the status indicator vectors (2.1), $M_{\mathbf{s}_j}$ can be compactly represented as follows: [2]

$$M_{\mathbf{s}_j} \triangleq \mathbb{1} \otimes \mathbb{1} \otimes \ldots \otimes \underbrace{\quad I \quad}_{j\text{th position}} \otimes \ldots \otimes \mathbb{1} \quad , \tag{2.15}$$

with $I$ being the $m_j \times m_j$ identity matrix, and the $k$th $\mathbb{1}$ being the $m_k \times 1$ all-1s vector. Note that $M_{\mathbf{s}_j}$ is an $\eta \times m_j$ matrix, and thus is usually a 'skinny' matrix. Moreover, $M_{\mathbf{s}_j}$ consists of only 0s and 1s, and is row-stochastic.

[2] To show why (2.12) holds for $M_{\mathbf{s}_j}$ defined as in (2.15), we require the mixed-product property of Kronecker products (A.3), namely,

$$(AB) \otimes (CD) = (A \otimes C)(B \otimes D) \,, \tag{2.13}$$

for appropriately-sized matrices $A$, $B$, $C$, and $D$. With this,

$$\begin{aligned} \mathbf{s}_{\mathbf{x}}' M_j &= \left( \bigotimes_{j=1}^{n} \mathbf{s}_j' \right) (\mathbb{1} \otimes \ldots \otimes I \otimes \ldots \otimes \mathbb{1}) \\ &= \mathbf{s}_1' \mathbb{1} \otimes \ldots \otimes \mathbf{s}_j' I \otimes \ldots \otimes \mathbf{s}_n' \mathbb{1} \\ &= \mathbf{s}_j' \,. \end{aligned} \tag{2.14}$$

One can generalize this notation and define a matrix $M_{\mathbf{s}_j}$ that permits one to derive the joint status indicator vector $\mathbf{s}_j$ from $\mathbf{s}_\mathbf{x}$:

$$\mathbf{s}_j' = \mathbf{s}_\mathbf{x}' M_{\mathbf{s}_j} \ . \tag{2.16}$$

The matrix $M_{\mathbf{s}_j}$ also serves as the marginalizing matrix that maps the network state probability vector $\boldsymbol{\pi}_\mathbf{x}$ to the probability vector $\boldsymbol{\pi}_j$. As long as the dimension of $\boldsymbol{j}$ is much less than $n$, $M_{\mathbf{s}_j}$ will be a 'skinny' matrix. This discussion of marginalization matrices will be continued after defining a few additional vectors.

### 2.1.6 State array vector $\mathbf{s}^{(1)}[t]$

While we have defined status indicator vectors and joint status indicator vectors whose expected values provide one with a means of representing the PMF for an automaton's status or several automata's joint status, respectively, we have yet to consider concatenating these indicator vectors into a larger vector whose expected value contains several pieces of marginal information. For example, consider the *state array vector* $\mathbf{s}^{(1)}[t]$, defined in terms of the individual status indicator vectors as follows:

$$\mathbf{s}^{(1)}[t] = \left[ \ \mathbf{s}_1[t]' \ \mathbf{s}_2[t]' \ \ldots \ \mathbf{s}_n[t]' \ \right]' \ . \tag{2.17}$$

The state array vector $\mathbf{s}^{(1)}[t]$ has as many 1's as the number of automata in the network, $n$. For $n > 2$, its length is much less than the length of $\mathbf{s}_\mathbf{x}[t]$, namely $\sum_{j=1}^n m_j$ instead of $\prod_{j=1}^n m_j$. For our weather example, the length of $\mathbf{s}^{(1)}[t]$ is just 12 (compared to a length of 81 for $\mathbf{s}_\mathbf{x}[t]$).

The expected value of $\mathbf{s}^{(1)}[t]$, typically denoted as $\boldsymbol{\pi}^{(1)}[t]$, is a concatenation of the probability vectors for each automaton's status:

$$\begin{aligned}\boldsymbol{\pi}^{(1)}[t] &\triangleq \mathbb{E}\left[\mathbf{s}^{(1)}[t]\right] \\ &= \left[ \ \mathbb{E}\left[\mathbf{s}_1[t]'\right] \ \mathbb{E}\left[\mathbf{s}_2[t]'\right] \ \ldots \ \mathbb{E}\left[\mathbf{s}_n[t]'\right] \ \right] \quad . \\ &= \left[ \ \boldsymbol{\pi}_1[t]' \ \boldsymbol{\pi}_2[t]' \ \ldots \ \boldsymbol{\pi}_n[t]' \ \right]\end{aligned} \tag{2.18}$$

Note that $\boldsymbol{\pi}^{(1)}$ does not tell us anything specific regarding the probabilities of *simultaneous* statuses of multiple automata, it only gives us partial information in the form of univariate marginal PMFs. This is in contrast to $\boldsymbol{\pi} = \mathbb{E}\left[\mathbf{s}_\mathbf{x}\right]$, which provides a *joint* PMF over the statuses of all automata.

### 2.1.7 2nd-order state array vector $\mathbf{s}^{(2)}[t]$

One can define a vector whose expected value provides the marginal PMF for the joint status of any $r$ of automata. We focus on the $r = 2$ case, where the expected value of the vector provides all bivariate marginal PMFs. Generalizing to $r > 2$ is straightforward.

The *2nd-order state array vector* $\mathbf{s}^{(2)}[t]$ is defined in terms of the state array vector $\mathbf{s}^{(1)}[t]$ as follows:

$$\mathbf{s}^{(2)}[t] = \mathbf{s}^{(1)}[t] \otimes \mathbf{s}^{(1)}[t] \ . \tag{2.19}$$

This 2nd-order state array vector $\mathbf{s}^{(2)}[t]$ can be partitioned into a sequence of $n^2$ Kronecker products involving two status indicator vectors:

$$\mathbf{s}^{(2)}[t] = \left[ \underbrace{\mathbf{s}_1[t]' \otimes \mathbf{s}_1[t]'}_{\text{1st pair}} \; \underbrace{\mathbf{s}_1[t]' \otimes \mathbf{s}_2[t]'}_{\text{2nd pair}} \ldots \underbrace{\mathbf{s}_n[t]' \otimes \mathbf{s}_n[t]'}_{n^2\text{th pair}} \right]' . \tag{2.20}$$

Each of the above blocks is an indicator vector, which the reader should recognize as a unique 2nd-order joint status indicator vector. The expected value of the 2nd-order state array vector $\mathbf{s}^{(2)}[t]$ provides the joint PMFs for the joint statuses of any two automata at time $t$, i.e., all bivariate marginal PMFs. By extending these definitions to $r > 2$, we can define an $r$th-order state array vector $\mathbf{s}^{(r)}[t]$ that is composed of blocks of $r$th-order joint status indicator vectors,

$$\mathbf{s}^{(r)}[t] = \left(\mathbf{s}^{(1)}[t]\right)^{\otimes r} \quad , \tag{2.21}$$

where a superscript $^{\otimes r}$ denotes the *rth Kronecker power*, which is obtained by taking Kronecker products of $r$ instances of the base vector/matrix [28]. The expected value of $\mathbf{s}^{(r)}[t]$, often denoted as $\boldsymbol{\pi}^{(r)}$, provides partial information about the full PMF of the network state, in the form of all PMFs for joint statuses of $r$ automata. We say that $\boldsymbol{\pi}^{(r)}$ contains all $r$th-order marginal PMFs. The expected value of an $r$th-order state array vector consists of $n^r$ marginal PMFs for $r$th-order joint status vectors. Note that such a representation is redundant. For example, consider the 2nd-order state array vector. The PMF for the joint status of automata 1 and 2 appears twice, in blocks $\boldsymbol{\pi}_{(1,2)} = \mathbb{E}\left[\mathbf{s}_1 \otimes \mathbf{s}_2\right]$ and $\boldsymbol{\pi}_{(2,1)} = \mathbb{E}\left[\mathbf{s}_2 \otimes \mathbf{s}_1\right]$. Note that the marginal PMF for automaton 1 is given by the sparse probability vector $\boldsymbol{\pi}_{(1,1)} = \mathbb{E}\left[\mathbf{s}_1 \otimes \mathbf{s}_1\right]$, and can also be obtained by marginalizing (linear operations) either $\boldsymbol{\pi}_{(1,2)}$ or $\boldsymbol{\pi}_{(2,1)}$, is also offered . Evidently, there are only $\binom{n}{2}$ marginal PMFs that are of interest. Consequently, for general $r$, we will refer to $\boldsymbol{\pi}^{(r)}$ as containing all $\binom{n}{r}$ $r$th-order marginal PMFs (even though the expected value of such an $r$th-order state array vector actually contains $n^r$ probability vectors).

### 2.1.8 General marginalizing matrices

One should note that the state array vector, as suggested by its name, is a state vector, for which there is a bijective mapping between its value and the value of the state $\mathbf{x}$. This bijective mapping can be represented via a matrix $M_{\mathbf{s}^{(1)}}$, whose $p$th row provides the value of the state array vector $\mathbf{s}^{(1)}$ when the network is in its $p$th state. As the mapping is bijective, the rows of $M_{\mathbf{s}^{(1)}}$ must be unique. It should be evident that $M_{\mathbf{s}^{(1)}}$ directly maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to the corresponding state array vector $\mathbf{s}^{(1)}$,

$$\mathbf{s}^{(1)'} = \mathbf{s}_{\mathbf{x}}' M_{\mathbf{s}^{(1)}} \quad . \tag{2.22}$$

Such notation for $M_{\mathbf{s}^{(1)}}$ should be familiar to the reader; recall that $M_{\mathbf{s}_j}$ maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to the status indicator vector $\mathbf{s}_j$ (2.12). In fact, we will employ this notation repeatedly—a matrix $M_{\mathbf{v}}$ maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to the vector $\mathbf{v}$. We think of $\mathbf{v}[t]$ as a stochastic process, whose value at any time $t$ must be expressible as a function of the network state at time $t$, i.e., $\mathbf{v}[t]$ is in the $\sigma$-field generated by the random variable that is the network state at time $t$. We should note that $\mathbf{v}$ need not be a state vector, which is indeed the case when $n \geq 2$ and $\mathbf{v}$ is only a single automaton's status indicator vector, e.g., $\mathbf{v} = \mathbf{s}_1$.

The notation for such marginalizing matrices $M_{\mathbf{v}}$ will be simplified in two cases, when $\mathbf{v}$ is either a (joint) status indicator vector—the special case discussed in Section 2.1.5—or when $\mathbf{v}$ is an $r$th-order state array vector. In the former case, we denote the matrix that maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to the joint status indicator vector $\mathbf{s}_{\boldsymbol{j}}$ as $M_{\boldsymbol{j}} \triangleq M_{\mathbf{s}_{\boldsymbol{j}}}$. Note that by linearity of expectation, $M_{\boldsymbol{j}}$ is the *marginalizing* matrix that takes the probability vector for the network state and maps it to the probability vector for the joint status for automata $\boldsymbol{j}$. In the later case, we denote the matrix that maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to the $r$th-order state array vector $\mathbf{s}^{(r)}$ as $M^{(r)} \triangleq M_{\mathbf{s}^{(r)}}$.

As $\mathbf{s}^{(1)}$ consists of a sequence of concatenated status indicator vectors (2.17), we can represent $M^{(1)}$ in block form as follows:

$$\begin{aligned} M_{\mathbf{s}^{(1)}} &= \begin{bmatrix} M_{\mathbf{s}_1} \; M_{\mathbf{s}_2} \; \ldots \; M_{\mathbf{s}_n} \end{bmatrix} \\ M^{(1)} &= \begin{bmatrix} M_1 \; M_2 \; \ldots \; M_n \end{bmatrix} \quad \text{(simplified notation)} \quad . \end{aligned} \tag{2.23}$$

By linearity of expectation, the mapping of (2.22) must hold under expectations, and thus,

$$\mathbb{E}\left[\mathbf{s}^{(1)\prime}\right] = \mathbb{E}\left[\mathbf{s}_{\mathbf{x}}'\right] M^{(1)} \quad . \tag{2.24}$$

If we denote $\boldsymbol{\pi}^{(1)} \triangleq \mathbb{E}\left[\mathbf{s}^{(1)}\right]$, what we have have in (2.24) is

$$\left[\boldsymbol{\pi}^{(1)}\right]' = \left[\boldsymbol{\pi}_1' \; \boldsymbol{\pi}_2' \; \ldots \; \boldsymbol{\pi}_n'\right] = \boldsymbol{\pi}_{\mathbf{x}}' M^{(1)} \tag{2.25}$$

Thus $M^{(1)}$ is the matrix that converts the probability vector for the network state (representing a multi-variate PMF) into a sequence of probability vectors, each of which represents the univariate marginal PMF for a different automaton's status. We say that the matrix $M^{(1)}$ takes the full probabilistic state information $\boldsymbol{\pi}_{\mathbf{x}}$ and maps it to the *partial information* in the form of the marginal PMFs for each of $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$. Although the mapping between the state vectors $\mathbf{s}_{\mathbf{x}}$ and $\mathbf{s}^{(1)}$ via $M^{(1)}$ is lossless, the mapping between the expected values of such state vectors via $M^{(1)}$ is lossy, in general.

The matrix $M_{\mathbf{v}}$ is referred to as a marginalizing matrix even in cases when $\mathbf{v}$ is neither an indicator vector nor a concatenation of indicator vectors. However, in the majority of examples we consider, $M_{\mathbf{v}}$ is a marginalizing matrix in the strict sense, so the term *marginalizing matrix* in only infrequently abused.

We will often want to consider an unspecified marginalizing matrix, which will be denoted as $M_*$. By following our simplified notation, $M_*$ maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to the vector $\mathbf{s}_*$, a linear transformation of the state indicator vector, i.e.,

$$\mathbf{s}_*' = \mathbf{s}_{\mathbf{x}}' M_* \quad . \tag{2.26}$$

Moreover, $M_*$ maps the probability vector for the network state to the *partial information vector*,

$$\boldsymbol{\pi}_*' = \boldsymbol{\pi}_{\mathbf{x}}' M_* \quad . \tag{2.27}$$

As was noted in Section 2.1.7, the expected value of the $r$th-order state array vector $\boldsymbol{\pi}^{(r)}$ provides redundant information for $r > 1$ Equivalently, the columns of $M^{(r)}$ satisfy linear equalities, i.e., $M^{(r)}$ does not have full column rank (strict linear equalities are of consideration here for the existence of a null space, not affine linear equalities). In fact, this is true even for $r = 1$, as each of its partitions as probability vectors will have the same sum. By considering general matrices $M_*$ and the linear transformations of the state indicator vectors that they provide, one may consider in place of $M^{(1)}$ a matrix with the same column space but full column rank. This notion is central to our subsequent discussion, and is the basis of our approach to remove the redundancies in the representation of $\boldsymbol{\pi}^{(r)}$, thereby reducing computational burdens.

## 2.2 $M_*$-separability

Consider a network of stochastic automata that evolves as a time-homogeneous Markov chain. As long as the size of the state space is of 'reasonable' size, there exists a tractable means of estimating the joint transition matrix $G$ that permits one to compute the conditional PMF for the network state at some future time $t$, given the current network state at time $\tau$, as in (2.5). This will not be the case when $n$, the number of automata, is large, as the size of the state space, $\eta$ is exponential in $n$. $M_*$-separability is a general framework that often offers tractable analysis in such regimes when $n$ is sufficienty large that traditional analysis via (2.5) is intractable.

The PMF for the network state is captured by the $\eta \times 1$ probability vector $\mathbb{E}\left[\,\mathbf{s}_{\mathbf{x}}[t]\,\right]$, a vector whose size is exponential in $n$. Such detailed information as provided by the PMF for the network state may be unnecessary. Under many situations, one may only need to know the probabilities of the joint statuses for a relatively small number of $r << n$ automata, meaning that knowing the expected value of the $r$th-order state array vector, $\mathbb{E}\left[\,\mathbf{s}^{(r)}[t]\,\right]$, a vector whose size is $O(n^r)$ would be sufficient. This observation, combined with the computational limitations already discussed in the large $n$ regime, motivate the idea of $M_*$-separability.

Under $M_*$-separability, $M_*$ serves as a matrix that transforms the probability vector for the network state at time $t$, $\boldsymbol{\pi}_{\mathbf{x}}[t]$, into some partial information $\boldsymbol{\pi}_*[t]' \triangleq \boldsymbol{\pi}_{\mathbf{x}}[t]'M_*$. By partial information we mean that $M_*$ is a lossy linear transformation, or equivalently, $M_*$ fails to have full row rank. Examples of lossy $M_*$ include $M^{(r)}$ for any $r < n$, which transforms a joint PMF on $n$ random variables into a sequence of marginal PMFs involving only $r$ variables.

**Definition 1.** *A network of stochastic automata with transition matrix $G$ is $M_*$-**separable** provided that $G$ exhibits a special property: $\boldsymbol{\pi}_*[t]' \triangleq \boldsymbol{\pi}_{\mathbf{x}}[t]'M_*$ can be expressed as a linear function of $\boldsymbol{\pi}_*[t-1]$, when $\boldsymbol{\pi}_*[t-1]$ can assume any of it possible values (meaning that both vectors should be thought of as conditional expectations of instances of the random vector $\mathbf{s}_*' = \mathbf{s}_{\mathbf{x}}'M_*$ at different times). Mathematically, there exists a matrix $H_*$ such that*

$$\boldsymbol{\pi}_*[t]' = \boldsymbol{\pi}_*[t-1]'H_* \tag{2.28}$$

*for any initial condition $\boldsymbol{\pi}_*[t-1]$.*

**Corollary 1.** *A network of stochastic automata with transition matrix $G$ is $M_*$-separable if and only if there exists a matrix $H_*$ such that*

$$\mathbb{E}\left[\,\mathbf{s}_*[t]'\,|\,\mathbf{s}_*[t-1]\,\right] = \mathbf{s}_*[t-1]'H_* \tag{2.29}$$

*for every possible value of $\mathbf{s}_*[t-1] \triangleq \mathbf{s}_{\mathbf{x}}[t-1]'M_*$.*

This directly follows from Definition 1 by setting the initial condition $\boldsymbol{\pi}_*[t-1]$ equal to $\mathbf{s}_*[t-1]$, the linear transformation of the state vector $\mathbf{s}_\mathbf{x}[t-1]$.

**Corollary 2.** *A network of stochastic automata with transition matrix $G$ is $M_*$-separable if and only if there exists a matrix $H_*$ such that*

$$GM_* = M_*H_* \quad . \tag{2.30}$$

This directly follows by applying the definition of $M_*$-separability.

When $M_\mathbf{v}$-separability is assumed, we will represent the corresponding $H_*$ matrix as $H_\mathbf{v}$. Furthermore, we will employ the same shorthand notations for $H_\mathbf{v}$ as we have for $M_\mathbf{v}$, i.e., under $M_\mathbf{j}$-separability, there exists a matrix $H_\mathbf{j}$, and under $M^{(r)}$-separability there exists a matrix $H^{(r)}$. Moreover, it should be clear from (2) that any network of stochastic automata is $M_*$-separable for $M_* = \mathbb{1}$.

**Corollary 3.** *A network exhibits $M_*$-separability if and only if it exhibits $\overline{M}_*$-separability for every $\overline{M}_*$ such that $\mathcal{R}(M_*) = \mathcal{R}(\overline{M}_*)$.*

Note that $\mathcal{R}(\cdot)$ denotes the range/column space of a matrix. Corollary 3 follows upon realizing that the existence of a matrix $H_*$ in (2.30) is equivalent to $G$ being $\mathcal{R}(M_*)$-invariant [29]. Such algebraic equivalences for $M_*$-separability will be discussed at length in Section 3.2.1.

**Corollary 4.** *For a network of stochastic automata that is $M_*$-separable,*

$$\boldsymbol{\pi}_*[t]' = \boldsymbol{\pi}_*[\tau]' H_*^{t-\tau} \tag{2.31}$$

*for every possible initial condition $\boldsymbol{\pi}_*[\tau]' \triangleq \boldsymbol{\pi}_\mathbf{x}[\tau]' M_*$.*

This follows by repeatedly applying Corollary 2 to (2.28). Note that Corollary 4 allows one to propagate the partial information $\boldsymbol{\pi}_*[\tau]$ via a matrix $H_*$, whose size equals the number of columns of $M_*$. Recall that when $M_* = M^{(r)}$, the number of columns of $M^{(r)}$ (and hence the size of $H^{(r)}$) is polynomial in $n$, and thus computations involving $H^{(r)}$ may be tractable when $G$ of (2.5) is intractable.

**Corollary 5.** *For a network of stochastic automata that is $M_*$-separable,*

$$\mathbb{E}\left[\, \mathbf{s}_*[t]' \,|\, \mathbf{s}_*[\tau] \,\right] = \mathbf{s}_*[\tau]' H_*^{t-\tau} \tag{2.32}$$

*for any possible value of $\mathbf{s}_*[\tau]' \triangleq \mathbf{s}_\mathbf{x}[\tau]' M_*$.*

This follows from Corollary 4 in the same way that Corollary 1 follows from Definition 1.

### 2.2.1 Origins of separability

$M_*$-separability is primarily inspired by two sources, the influence model (IM) [12, 13] and Pfeffer's notion of separability [30]. Pfeffer's separability is, in essence, a local characterization of $M^{(1)}$-separability in a Bayesian network without any notion of time. See Section 4.1.1 for a clearer explanation of this relationship. As shown in [31], one can envision the influence model as Pfeffer's separability applied to a dynamic Bayesian network (DBN) [32, 33, 34, 35], or in other words, an influence model is a DBN that exhibits $M^{(1)}$-separability. One can show that it is a consequence of the conditional independencies of the DBN that the influence model also exhibits $M^{(r)}$-separability for all $r \leq n$. It is the computational advantages offered by these different forms of $M_*$-separability that make the influence model attractive.

The Bayesian network backbone that is characteristic of both Pfeffer's separability and the IM simplifies the parametric definition of the probabilistic models; nevertheless, such conditional independencies are not necessary for partial information to be propagated. The moment-linear stochastic systems (MLSS) defined in [14, 15] extend the propagation of partial information beyond Bayesian networks. As a class of probabilistic models, MLSS overlap substantially with $M_*$-separable systems; MLSS include all of the cases of $M_*$-separability when $\mathbb{1} \in \mathcal{R}(M_*)$, and $M_*$ has unique rows or, equivalently, $\mathbf{s}_*$ is a state vector. Despite this overlap, we elect to adopt different terminology to emphasize our constraint-based perspective for $M_*$-separability. Naturally, our discussion on $M_*$-separability offers much insight into MLSS examples, and vice-versa.

It is also important to note the generalized influence model (GIM) [36], a specific class of parametric probabilistic models that propagates partial information without the constraints of conditional independence. As its name suggests, the GIM generalizes the parametric form detailing how updates occur in a IM to permit richer dynamics, and thereby moves beyond the DBN restrictions of the IM (the GIM is still Markovian, however) while maintaining a compact parametric representation. The GIM is presented in Chapter 4.

### 2.2.2 $M_*$-sufficiency

An obvious consequence of assuming $M_*$-separability is that for any two probability vectors for the network state at time $t-1$, $\boldsymbol{\pi}_{\mathbf{x}}[t-1]$ and $\overline{\boldsymbol{\pi}}_{\mathbf{x}}[t-1]$, that produce the same partial information, i.e.,

$$\boldsymbol{\pi}_{\mathbf{x}}[t-1]'M_* = \overline{\boldsymbol{\pi}}_{\mathbf{x}}[t-1]'M_* \quad , \tag{2.33}$$

the resulting partial information at time $t$ will be equal, should the network be initialized by either probability vector at time $t-1$ (as clear by (2.28) ). We call this $M_*$-sufficiency.[3]

**Definition 2.** *A network of stochastic automata with transition matrix $G$ exhibits $M_*$-* ***sufficiency*** *provided that for every pair of probability vectors $\boldsymbol{\pi}_{\mathbf{x}}[t-1]$ and $\overline{\boldsymbol{\pi}}_{\mathbf{x}}[t-1]$, such that $\boldsymbol{\pi}_{\mathbf{x}}[t-1]'M_* = \overline{\boldsymbol{\pi}}_{\mathbf{x}}[t-1]'M_*$, the partial information at time $t$ will be equal regardless of which of the two probability vectors at time $t-1$ is chosen to initialize the network.*

Although it should be clear that $M_*$-separability implies $M_*$-sufficiency, the converse, which is also true, may not be as obvious. These facts are summarized in the following theorem.

**Theorem 1.** *A network of stochastic automata with transition matrix $G$ is $M_*$-separable if and only if it is $M_*$-sufficient.*

*Proof.* That $M_*$-separability implies $M_*$-sufficiency should be obvious. Necessity is a consequence of basic properties of linear algebra, which will be explained in what follows. One can define an open ball $\mathcal{B} \subset \mathbb{R}^{\eta}$ such that for any $\tilde{\boldsymbol{\pi}} \in \mathcal{B}$, there exist probability vectors $\boldsymbol{\pi}_{\mathbf{x}}[t-1]$ and $\overline{\boldsymbol{\pi}}[t-1]$ such that $\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi}_{\mathbf{x}}[t-1] - \overline{\boldsymbol{\pi}}[t-1]$. Note that $\mathcal{B}$ must be small. By the assumption of $M_*$-sufficiency (Definition 1), for all $\tilde{\boldsymbol{\pi}} \in \mathcal{B}$ such that $\tilde{\boldsymbol{\pi}}M_* = 0$, it follows

[3] We borrow the term 'sufficiency' from Pfeffer [30], for whom sufficiency is a local characterization of $M^{(1)}$-sufficiency.

that $\tilde{\pi} G M_* = 0$. As $\mathcal{B}$ is an open ball with full dimensionality in $\mathbb{R}^\eta$, it follows that

$$\mathcal{N}(M_*') \subset \mathcal{N}\big((GM_*)'\big) \quad . \tag{2.34}$$

where $\mathcal{N}(\cdot)$ indicates the null space of a matrix. Consequently,

$$\mathcal{R}(GM_*) \subset \mathcal{R}(M_*) \quad . \tag{2.35}$$

Thus by (2.35) there must exist a matrix $H_*$ such that $GM_* = M_* H_*$. Note that the existence of a matrix $H_*$ satisfying such an equation is equivalent to $M_*$-separability (Corollary 2), and thus $M_*$-sufficiency implies $M_*$-separability.[4] □

Although it is now clear that $M_*$-separability and $M_*$-sufficiency are equivalent, we will henceforth go back to referring only to $M_*$-separability.

### 2.2.3 Examples of $M_*$-separability

The simplest form of $M_*$-separability is when $M_* = \mathbb{1}$. All networks of stochastic automata exhibit $\mathbb{1}$-separability, as the joint transition matrix $G$ always has constant row sums (and in fact always sums to 1, and thus the associated $H_*$ matrix is simply 1). As $\mathbb{1}$-separability is not illuminating, we will mostly ignore it. Occasionally we will consider it when characterizing the constraints of separability.

One of the simplest forms of separability is $M_{\boldsymbol{j}}$-separability. When $M_{\boldsymbol{j}}$-separability is assumed, the distribution of the joint status for the automata identified by $\boldsymbol{j}$ can be determined by linear recursion. This fact, in conjunction with the assumed Markovianity of the network state and the tower property of conditional probabilities [37], necessitates that the subprocess $\mathbf{x}_{\boldsymbol{j}}$ itself be Markovian. However, the Markovianity of $\mathbf{x}_{\boldsymbol{j}}$ is not equivalent to $M_{\boldsymbol{j}}$-separability. $M_{\boldsymbol{j}}$-separability demands not only that $\mathbf{x}_{\boldsymbol{j}}$ be Markovian, but that given the value of $\mathbf{x}_{\boldsymbol{j}}$ at time $t$, future values of $\mathbf{x}_{\boldsymbol{j}}$ be conditionally independent not only of past values of $\mathbf{x}_{\boldsymbol{j}}$, but also of past values of the other automata not enumerated in $\boldsymbol{j}$. That $M_{\boldsymbol{j}}$-separability imposes this additional constraint beyond Markovianity of $\mathbf{x}_{\boldsymbol{j}}$ is easily seen by example. Suppose there are two automata. Let the first automaton be Markovian, and let the second always be a delayed replica of the first (delayed by one time unit). Clearly, the network state $\mathbf{x}_{(1,2)}$ is Markovian, and $\mathrm{x}_2$ is individually Markovian. However, the network does not exhibit $M_2$ separability, as the aforementioned conditional independence does not hold.

A second interesting example of separability is $M^{(r)}$-separability, for $1 \leq r < n$. Examples of $M^{(r)}$-separability are collectively referred to as *canonical* examples of separability. In the $r = 1$ case, it is assumed that the univariate marginal PMFs can be propagated in time; the bivariate marginal PMFs can be propagated in time for $r = 2$. It is desirable to be able to propagate such partial information, particularly when $n$ is large (as the probability vector capturing the PMF for the network state would be exponential in $n$, while the vector capturing the marginal PMFs would only be an $r$th-order polynomial in $n$). A common misconception is that for $r_2 > r_1$, $M^{(r_2)}$-separability implies $M^{(r_1)}$-separability. As illustrated within the discussion of the parametrization of $\mathcal{M}$-separability (Section 3.4.3), this is not true; neither implies the other, in general. However, when $r_2 > r_1$, one can always obtain the lower-order marginal information $\boldsymbol{\pi}^{(r_1)}$ from $\boldsymbol{\pi}^{(r_2)}$. In examples when $M^{(r_1)}$-separability

[4] This proof is a generalization of the proof offered in [31] that equates $M^{(1)}$-separability to $M^{(1)}$-sufficiency. We should note that in [31] there is an additional restriction that $H^{(1)}$ be *nonnegative.*

does not hold but $M^{(r_2)}$-separability does, one may elect to propagate $\boldsymbol{\pi}^{(r_2)}$ with a matrix $H^{(r_2)}$ of size $O(n^{r_2})$ and marginalize $\boldsymbol{\pi}^{(r_2)}$ to obtain the coarser information $\boldsymbol{\pi}^{(r_1)}$.

We can illustrate these ideas for our weather example. When $M^{(1)}$-separability is satisfied, the marginal distributions for the weather in each of the four cities can be determined by a linear recursion (2.31). Similarly, when $M^{(2)}$-separability is satisfied, the joint distributions for the weather in pairs of cities can be determined by a linear recursion. Under $M^{(r)}$-separability, the joint distributions for the weather in $r$ cities can be determined by a linear recursion.

$M_*$-separability is not limited to cases that involve the propagation of marginal distributions. Consider the matrix $M_{\mathbf{x}}$ that maps the state indicator vector to $\mathbf{x}$. Under the assumption of $M_{\mathbf{x}}$-separability, one assumes that the expected value of $\mathbf{x}$ can be propagated linearly in time. By selectively choosing the columns of $M_*$, one can define a vector $\mathbf{s}_*$ whose expectation yields any desired sequence of moments, i.e., any quantity expressible as a linear function of the PMF for the network state (recall that the term 'moment' will be used to refer to any such quantity, including marginal distributions). $M_*$-separability in such a case characterizes the models for which such moments can be linearly propagated.

### Affine linear propagation

Rather than limiting ourselves to linear models, we may want to consider when the expected value of $\mathbf{s}_*$ can be propagated in time by an *affine* linear function, i.e.,

$$\mathbb{E}\left[\,\mathbf{s}_*[t]' \,|\, \mathbf{s}_*[t-1]\,\right] = \mathbf{s}_*[t-1]'H_* + \boldsymbol{h}_*' \quad , \tag{2.36}$$

for some matrix $H_*$ and vector $\boldsymbol{h}_*$. Note that this was the setup considered in MLSS. Mirroring how Corollary 2 was derived from Definition 1, we can show that the expected value of $\mathbf{s}_*$ can be propagated in time by a linear function if and only if there exists a matrix $\widetilde{H}_*$ such that

$$GM_* = [M_* \;\; \mathbb{1}]\, widetildeH_* \quad , \tag{2.37}$$

or in simple terms, $GM_*$ is in the column space of $[M_* \;\; \mathbb{1}]$. It follows that $GM_*$ is in the column space of $[M_* \;\; \mathbb{1}]$ if and only if $G\,[M_* \;\; \mathbb{1}]$ is in the column space of $[M_* \;\; \mathbb{1}]$ (because $\mathbb{1}$ is a right eigenvector of $G$). This realization leads us to the following corollary.

**Corollary 6.** *The expected value of* $\mathbf{s}_*$ *can be propagated in time by an* affine *linear function if and only if the network is* $[M_* \;\; \mathbb{1}]$*-separable. When* $\mathbb{1} \in \mathcal{R}(M_*)$*, the expected value of* $\mathbf{s}_*$ *can be propagated by an affine linear function if and only if it can also be propagated by a linear function.*

The second observation in Corollary 6 is a consequence of Corollary 3.

By definition, a network is a moment-linear stochastic system of degree 1 with respect to the state vector $\mathbf{s}_*$ if the expected value of $\mathbf{s}_*$ can be propagated in time by an affine linear function [14, 15]. Because for *any* state vector $\mathbf{s}_*$ there exists a matrix $M_*$ that maps the state indicator vector $\mathbf{s}_{\mathbf{x}}$ to $\mathbf{s}_*$, i.e.,

$$\mathbf{s}_*' = \mathbf{s}_{\mathbf{x}}' M_* \quad , \tag{2.38}$$

it follows from Corollary 6 that an MLSS of degree 1 with respect to the state vector $\mathbf{s}_*$ that is derived from a finite-state Markov chain is equivalent to $[M_* \;\; \mathbb{1}]$-separability.

### Separability in chemical kinetics

Mass action kinetics (MAK) [38] can be thought of as an example of $M_{\mathbf{x}}$-separability that is extended to continuous-time Markov chains (extending the ideas of $M_*$-separability to continuous-time models is straightforward). Consider a large collection of chemical species in a closed container. The species will react with one another, and the counts of the number of molecules of each species will fluctuate. Provided that the temperature remains constant, one can argue that such a system can be well modeled as a time-homogeneous Markov process whose state vector $\mathbf{x}$ is a vector of species counts. This mathematical description, the Kolmogorov forward equation, in chemical kinetics is commonly referred to as the chemical master equations (CME). MAK is the assumption of $M_{\mathbf{x}}$-separability in continuous-time; MAK is the mathematical description of the evolution of the expected *concentrations* of the chemical species, i.e., scaled counts. Note that because of conservation of mass, there is a linear equation involving the counts of the various chemical species that remains constant over time; this fact ensures that the all 1s vector $\mathbb{1} \in \mathcal{R}(M_{\mathbf{x}})$ (the coefficients of a conservation of mass equation specify how to combine the columns of $M_{\mathbf{x}}$ to define a constant column, a scaled instance of $\mathbb{1}$).

Alternatively, rather than just propagating the moments of a state vector, one can consider propagating the moments and correlations; this would be equivalent to $M_{[\mathbf{x}\ \mathbf{x}^{\otimes 2}]}$-separability. For chemical kinetics, $M_{[\mathbf{x}\ \mathbf{x}^{\otimes 2}]}$-separability would allow one to linearly propagate the means and correlations of the various species counts. In the language of MLSS, this is an example of a system that is assumed to be 2nd-moment linear. One can consider other instances of $M_*$-separability involving even higher moments. Note that central moments cannot be incorporated into examples of $M_*$-separability, as central moments cannot be obtained by a linear function of a joint distribution $\boldsymbol{\pi}_{\mathbf{x}}$. Mass fluctuation kinetics (MFK), which intends to generalize MAK by presenting a means via moment-closure to approximately propagate not only the expected counts of the chemical species, but also the covariances [16], is not an instance of $M_*$-separability, because of the incorporation of central moments. If, however, MFK was altered to include only non-central moments and was exact, it would be an instance of $M_*$-separability generalized to continuous-time.

Other models in chemical kinetics intermediate between the complete description offered by the CME and the coarse description of MAK have been proposed, e.g., [17] proposes a method to propagate bivariate marginals by nonlinear recursions. There has also been a concerted effort to develop further moment closure recursions [18] and approximately propagating higher moments [19]. Although the theory of $M_*$-separability does not pertain to such approximate schemes, it does offer specific guidance when partial information can be propagated *exactly.* This is summarized in the following corollary.

**Corollary 7.** *Consider any partial information $\boldsymbol{\pi}_*$, and any bijective function $f(\cdot)$ on such partial information $\boldsymbol{\pi}_*$. For example, $\boldsymbol{\pi}_*$ is a sequence of moments, e.g., mean and second moment, and $f(\boldsymbol{\pi}_*)$ is a sequence of corresponding central moments, e.g., mean and variance. Then $f(\boldsymbol{\pi}_*)$ can be propagated exactly if and only if $\boldsymbol{\pi}_*$ can be propagated linearly.*

This corollary follows from Theorem 1. Note that Corollary 7 does not make any statements regarding approximate propagation. The correct approach to gauging the fidelity of approximate propagation of moments depends on the situation and costs. If for example, one wished to derive a function that could provide the exact value of the moment $\boldsymbol{\pi}_*[t]$ given $\mathbf{s}[t-1]$, this could be done exactly, and unless $M_*$-separability is satisfied, *must* be done by nonlinear means. On the other hand, suppose one wanted to propagate $\boldsymbol{\pi}_*[t-1]$

to time $t$. Unless $M_*$-separability is satisfied, the information orthogonal to $\boldsymbol{\pi}_*[t-1]$ in part determines $\boldsymbol{\pi}_*[t]$. To determine the fidelity of an approximate method to propagate moments, one must place some measure on $\boldsymbol{\pi}_*[t-1]$ to evaluate a cost, and how this is done is subject to interpretation.

### 2.2.4 $\mathcal{M}$-separability

The idea of $\mathcal{M}$-separability is motivated by networks of stochastic automata that exhibit multiple kinds of $M_*$-separability.

**Definition 3.** *A network of stochastic automata exhibits $\mathcal{M}$-*separability*, with $\mathcal{M}$ being a set of matrices, provided that it exhibits $M_*$-separability for every $M_* \in \mathcal{M}$.*

Recall that every network of stochastic automata is $\mathbb{1}$-separable. Hence, for any network that is $M_*$-separable, we can envision it as being $\mathcal{M}$-separable, with $\mathcal{M} = \{\mathbb{1}, M_*\}$. Typically, we will not incorporate $\mathbb{1}$ as a element of $\mathcal{M}$, as it can always be assumed.

When there is a nesting of the range spaces of the matrices constituting $\mathcal{M}$, this special case of $\mathcal{M}$-separability is termed *regular separability.* It is assumed that under regular separability, $\mathbb{1} \in \mathcal{R}(M_*)$ for each $M_* \in \mathcal{M}$. Our particular interest in regular separability is twofold: many of the important examples of $\mathcal{M}$-separability that we consider are regular, and as we will show in Section 3.2.2, characterizing the parametrization induced by regular separability is straightforward. The generic terms 'separability' is intended to encompass both $M_*$-separability and $\mathcal{M}$-separability.

Although there are endless versions of $\mathcal{M}$-separability to consider, there are a few examples that we will focus on primarily. The first is when a network of automata satisfies $M^{(r)}$-separability for all $1 \leq r < n$, which is referred to as *total network separability* in [36]. Here we will refer to such networks as being $\mathcal{M}^{(n)}$-separable, where the set of matrices $\mathcal{M}^{(n)}$ is defined as

$$\mathcal{M}^{(n)} \triangleq \left\{M^{(r)} : 1 \leq r \leq n\right\} \quad . \tag{2.39}$$

Note that $M^{(n)}$-separability follows from the network state being Markovian. The IM [12, 13] and GIM [36] exhibit $\mathcal{M}^{(n)}$-separability. When a network is $M^{(r)}$-separable for all $1 \leq r \leq \overline{r}$ with $\overline{r} < n$, we classify it as being $\mathcal{M}^{(\overline{r})}$-separable. Such examples of regular separability are also referred to as canonical examples of separability.

A second kind of $\mathcal{M}$-separability that will be of interest is $\mathcal{M}_n$-separability, where the set of matrices $\mathcal{M}_n$ is defined as

$$\mathcal{M}_n \triangleq \{M_j \, : \, 1 \leq j \leq n\} \quad . \tag{2.40}$$

Under this particular kind of $\mathcal{M}$-separability, each automaton is itself Markovian. Each automaton also exhibits an additional property: given its current status, its future statuses do not depend on the past statuses of *any* of the automata. This may seem to imply that all of the automata are independent, but this is not the case. A GIM with all automata always selecting themselves as their influencers achieves this kind of $\mathcal{M}$-separability, without the automata necessarily being independent. For this special case, there is an independent and identically distributed (IID) exogenous input process, which then specifies how each Markov chain will transition for each of its possible current statuses. This framework provides a natural means to model coupled Markov chains.

MLSS [14, 15] provide an entire class of examples of $\mathcal{M}$-separability. Before proceeding to these examples of $\mathcal{M}$-separability, we first need to define when a state vector $\mathbf{v}$ is $r$th-moment linear. In order for a state vector $\mathbf{v}$ to be $r$th-moment linear, the conditional expected value of $\mathbf{v}[t]^{\otimes r}$ (sometimes referred to as the $r$th-moment of $\mathbf{v}[t]$) given $\mathbf{v}[t-1]$, must be expressible as an affine linear function of $\left\{\mathbf{v}[t-1]^{\otimes k}\right\}$ for $k \leq r$, that is,

$$\begin{aligned}&\mathbb{E}\left[\mathbf{v}[t]^{\otimes r'} \,|\, \mathbf{v}[t-1]\right] \\ &\quad = [\mathbf{v}[t-1]^{\otimes r}]' H_{r,r} + [\mathbf{v}[t-1]^{\otimes r-1}]' H_{r,r-1} + \ldots + \mathbf{v}[t-1]' H_{r,1} + \boldsymbol{h}'_{r,0} \quad , \qquad (2.41)\end{aligned}$$

for some matrices $H_{r,r}$, $H_{r,r-1}$, $\ldots$, $H_{r,1}$ and vector $\boldsymbol{h}_{r,0}$ [14, 15]. Equivalently, there exists a matrix $\widetilde{H}_r$ such that

$$G M_{\mathbf{v}^{\otimes r}} = [M_{\mathbf{v}^{\otimes r}} \;\; M_{\mathbf{v}^{\otimes r-1}} \ldots M_{\mathbf{v}} \;\; \mathbb{1}]\, \widetilde{H}_r \quad , \qquad (2.42)$$

One should recall that in the $r = 1$ case, by Corollary 6, the existence of a matrix $\widetilde{H}_1$ in (2.37) is equivalent to $[M_{\mathbf{v}} \;\; \mathbb{1}]$-separability. However, when $r > 1$, being $r$th-moment linear does not equate to any form of separability. However, it does equate to a form of $\mathcal{M}$-separability when we assume that the state vector $\mathbf{v}$ is $r$th-moment linear for all $r \leq \overline{r}$; this is referred to in [14, 15] as an MLSS of degree $\overline{r}$. To see this, we argue by induction. First note that in the $r = 1$ case, we have $[M_{\mathbf{v}} \;\; \mathbb{1}]$-separability. Next as our inductive step, assume that $[M_{\mathbf{v}^{\otimes r-1}} \;\; M_{\mathbf{v}^{\otimes r-2}} \;\; \ldots \;\; \mathbb{1}]$-separability holds. Suppose that the network is $r$th-moment linear with respect to $\mathbf{v}$. As $[M_{\mathbf{v}^{\otimes r-1}} \;\; M_{\mathbf{v}^{\otimes r-2}} \;\; \ldots \;\; \mathbb{1}]$-separability is already assumed, our equation for being $r$th-moment linear, (2.42), can be equivalently expressed in terms of the existence of a matrix $\overline{H}_r$ such that

$$G\,[M_{\mathbf{v}^{\otimes r}} \;\; M_{\mathbf{v}^{\otimes r-1}} \ldots M_{\mathbf{v}} \;\; \mathbb{1}] = [M_{\mathbf{v}^{\otimes r}} \;\; M_{\mathbf{v}^{\otimes r-1}} \ldots M_{\mathbf{v}} \;\; \mathbb{1}]\, \overline{H}_r \quad . \qquad (2.43)$$

One should now recognize (2.43) as being equivalent to $[M_{\mathbf{v}^{\otimes r}} \;\; M_{\mathbf{v}^{\otimes r-1}} \;\; \ldots \;\; \mathbb{1}]$-separability. Consequently, an MLSS of degree $\overline{r}$ is equivalent to being $[M_{\mathbf{v}^{\otimes r}} \;\; M_{\mathbf{v}^{\otimes r-1}} \ldots M_{\mathbf{v}} \;\; \mathbb{1}]$-separable for all $r \leq \overline{r}$, which we refer to compactly as $\mathcal{M}_{\mathbf{v}^{\otimes r}}$-separability. Note that this is an instance of regular separability.

We have already discussed an instance of $\mathcal{M}_{\mathbf{v}^{\otimes r}}$-separability: $\mathcal{M}^{(\overline{r})}$-separability, which is satisfied when the network exhibits $M^{(r)}$-separability for all $r \leq \overline{r}$, is equivalent to $\mathcal{M}_{[\mathbf{s}^{(1)}]^{\otimes \overline{r}}}$-separability. This can be demonstrated on the basis of Corollary 3. First, note that $M^{(r)}$ is shorthand notation for $M_{[\mathbf{s}^{(1)}]^{\otimes r}}$, and thus $\left[M_{[\mathbf{s}^{(1)}]^{\otimes r}} \;\; M_{[\mathbf{s}^{(1)}]^{\otimes r-1}} \;\; \ldots \;\; \mathbb{1}\right]$-separability is equivalent to $\left[M^{(r)} \;\; M^{(r-1)} \;\; \ldots \;\; \mathbb{1}\right]$-separability. Next, as $\mathbb{1} \in \mathcal{R}(M^{(r)})$ for all $r$, and $\mathcal{R}(M^{(r')}) \subset \mathcal{R}(M^{(r)})$ whenever $r' \leq r$, it follows that $M^{(r)}$-separability is equivalent to $\left[M^{(r)} \;\; M^{(r-1)} \;\; \ldots \;\; \mathbb{1}\right]$-separability.

The majority of the examples in [14, 15] are MLSS of degree $\overline{r}$. We refer the reader to these references for these interesting examples of regular $\mathcal{M}_{[\mathbf{s}^{(1)}]^{\otimes \overline{r}}}$-separability.

## 2.3 Conclusions

This chapter serves as the foundation for the remainder of the thesis. The general framework characterizing separability that is introduced in this chapter is further analyzed in the next chapter, which examines our general description of separability algebraically and geometrically. In future chapters, we will elaborate further on several examples of separability introduced in this chapter.

# Chapter 3

# Algebraic and Geometric Perspectives on Separability

In order for a network of stochastic automata to exhibit $M_*$-separability, its transition matrix $G$ must satisfy two conditions: not only must $G$ be row-stochastic, but by Corollary 2, there also must exist a matrix $H_*$ such that

$$GM_* = M_*H_* \quad . \tag{3.1}$$

Although these conditions define when a transition matrix satisfies $M_*$-separability, they fail to offer intuition into the algebraic and geometric structure of separability. This chapter attempts to reveal the inherent structure in networks exhibiting $M_*$-separability, as characterized by their transition matrices, and to likewise do the same for networks exhibiting $\mathcal{M}$-separability. The set of transition matrices satisfying $M_*$-separability will be denoted as $\mathcal{G}(M_*)$. Similarly, we define $\mathcal{G}(\mathcal{M})$ as the set of transition matrices exhibiting $\mathcal{M}$-separability. Note that as $M_*$ or $\mathcal{M}$ effectively fix $n$, the number of automata/random variables, and $\boldsymbol{m}$, the sizes of their alphabets, both $\mathcal{G}(M_*)$ and $\mathcal{G}(\mathcal{M})$ are thought of as a set of transition matrices that exhibit a particular form of separability for a network of automata with *fixed* $n$ and $\boldsymbol{m}$ .

The contributions of this chapter extend to the models that have motivated our discussion of separability, including the influence model [12, 13], moment-linear stochastic systems [14, 15], and the generalized influence model [36]. Most results of this chapter provide new insights into these specific examples of separability, while a few are generalizations of what is already known about some of these models, e.g., the spectral relationships of $G$ and $H$ for the IM (discussed in [12]).

We begin in Section 3.1 by discussing the implications of (3.1) on the transition matrix $G$, and proceed then to its implications on $H_*$. In both cases, our perspectives will draw heavily on geometry and linear algebra. Then we develop a unified algebraic parametrization of separability in Section 3.2, from which we will derive expressions for the dimensions of $\mathcal{G}(M_*)$ and $\mathcal{G}(\mathcal{M})$, i.e., which can be interpreted as the number of free parameters in a probabilistic model exhibiting $M_*$-separability or $\mathcal{M}$-separability, respectively. In doing so, it will become clear the relative degree to which different forms of separability constrain the underlying transition matrix of a network. Although we will develop equations for the dimension of $\mathcal{G}(M_*)$ and $\mathcal{G}(\mathcal{M})$ under general settings, we do so with the plan to apply such findings in Section 3.4 to canonical examples of separability: $M^{(r)}$-separability and $\mathcal{M}^{(\overline{r})}$-separability. In Section 3.5, we will analyze the evolution of the dimension of $\mathcal{G}(M^{(r)})$

as $r$, $m$, and $n$ vary, and then subsequently do the same for $\mathcal{G}(\mathcal{M}^{(\overline{r})})$. Several interesting insights into the asymptotic properties of separability will be illustrated. We conclude the chapter in Section 3.6 by offering an interpretation of the asymptotic results of Section 3.5, namely, that many canonical forms of separability have increasingly high relative dimension as $n$ grows.

## 3.1 Geometric implications of $M_*$-separability

We begin by discussing some special structure in the dynamics of the update equation for the probability vector for the network state (2.5) when $M_*$-separability is satisfied. Without loss of generality, we assume that $M_*$ has full column rank (recall Theorem 16). Our arguments will rely heavily on the the complementary orthogonal subspaces $\mathcal{R}(M_*)$ and $\mathcal{R}^\perp(M_*)$ ($\mathcal{R}^\perp(\cdot)$ denotes the orthogonal complement to the range space of the specified matrix). To simplify notation, we define $M_*^\perp$ as a matrix whose orthonormal columns serve as a basis for $\mathcal{R}^\perp(M_*)$.

By multiplying $\left[M_*^\perp\right]'$ to both sides of (3.1), we obtain an equation that equivalently characterizes the existence of a matrix $H_*$ in (3.1):

$$\left[M_*^\perp\right]' G M_* = 0 \quad . \tag{3.2}$$

We can also express the linear constraint on $G$ given in (3.2) as a requirement that $G$ be upper triangular under the linear transformation specified by $\left[M_* \;\; M_*^\perp\right]$, i.e.,

$$\left[M_* \;\; M_*^\perp\right]^{-1} G \left[M_* \;\; M_*^\perp\right] = \begin{bmatrix} H_* & H_{(12)} \\ 0 & H_{(22)} \end{bmatrix} \quad , \tag{3.3}$$

where $\left[M_* \;\; M_*^\perp\right]^{-1}$ can be rewritten as

$$\left[M_* \;\; M_*^\perp\right]^{-1} = \begin{bmatrix} M_*^{-L} \\ [M_*^\perp]' \end{bmatrix} \quad , \tag{3.4}$$

where

$$M^{-L} \triangleq \left(M_*' M_*^\perp\right)^{-1} M_*' \tag{3.5}$$

is the Moore-Penrose generalized inverse of $M_*$ [39], or equivalently, the only left inverse of $M_*$ whose row space coincides with the column space of $M_*$. It should be clear how the constraints of (3.2) manifest themselves on the right-hand side of (3.3), as the $(2,1)$ block consisting of 0s.

Those familiar with linear dynamical systems may recognize the existence of a matrix $H_*$ in (3.1), or the block triangular form of $G$ under a linear transformation in (3.3), as a statement concerning *invariant subspaces* [20, 21], namely, $\mathcal{R}(M_*)$ being $G$-invariant, or, as we may alternatively state, $G$ leaving $\mathcal{R}(M_*)$ invariant. Invariant subspaces are a well-developed topic in linear dynamical systems theory; however, at least to this author's knowledge, invariant subspaces beyond steady-state vectors or recurrent classes have not been explored in the context of Markov chains. We feel it is instructive to expose the ramifications of such invariance from a variety of perspectives.

Mathematically, this invariance can be expressed as

$$Gv \in \mathcal{R}(M_*) \quad \forall\, v \in \mathcal{R}(M_*) \quad . \tag{3.6}$$

As we already noted, $\mathcal{R}(M_*)$ being $G$-invariant is equivalent to $G$ having an upper block triangular form under the linear transformation specified by $\left[M_* \;\; M_*^{\perp}\right]$, as well as equivalent to the existence of a matrix $H_*$ in (3.1). Such invariance can also be equivalently expressed in terms of eigenvectors. $M_*$-separability is equivalent to the existence of a basis for $\mathcal{R}(M_*)$ consisting only of right eigenvectors (and generalized right-eigenvectors) of $G$.[1]

### 3.1.1 Left-invariance

As $G$ is the transition matrix of a Markov chain, it is more illuminating to discuss such invariance in terms of row vectors left-multiplying $G$; we call this *left-invariance.* As evident from (3.3), $M_*$-separability is equivalent to $G$ leaving $\mathcal{R}(M_*^{\perp})$ left-invariant, i.e.,

$$v'G \in \mathcal{R}(M_*^{\perp}) \quad \forall\, v \in \mathcal{R}(M_*^{\perp}) \quad , \tag{3.7}$$

and similarly equivalent to the existence of a matrix $H_{(22)}$ in (3.51) such that

$$\left[M_*^{\perp}\right]' G = H_{(22)} \left[M_*^{\perp}\right]' \quad . \tag{3.8}$$

In terms of left eigenvectors, such left-invariance is equivalent to the existence of a basis for $\mathcal{R}(M^{\perp})$ consisting only of left eigenvectors (and generalized left-eigenvectors) of $G$.

In linear systems theory, one leverages invariance when the dynamics are initialized in the invariant subspace. Because the subspace is invariant, the dynamics are guaranteed to remain in the subspace, thereby permitting a lower-order description of the dynamics in the subspace. Separability offers an alternate perspective on invariance. There is no precondition that the dynamics must be initialized in the invariant subspace to take advantage of such left-invariance. Under $M_*$-separability, left-invariance is leveraged by allowing one to track projections of the dynamics (evolution of probability vectors) into $\mathcal{R}(M_*)$, the orthogonal complement of the assumed left-invariant subspace, with a lower-order description.

To see how this works, let's return to our weather example. Suppose that the network is initialized with a probability vector at time 0, $\boldsymbol{\pi}_{\mathbf{x}}[0]$, which provides the joint probability of weather combinations for each of the four cities. Any probability vector $\boldsymbol{\pi}_{\mathbf{x}}$ can be represented as a sum of its projections onto two complementary orthogonal subspaces, $\mathcal{R}(M_*)$ and $\mathcal{R}^{\perp}(M_*)$, as follows:

$$\begin{aligned} \boldsymbol{\pi}_{*\parallel} &\triangleq \boldsymbol{\pi}_{\mathbf{x}}' M_* M_*^{-L} \in \mathcal{R}(M_*) \\ \boldsymbol{\pi}_{*\perp} &\triangleq \boldsymbol{\pi}_{\mathbf{x}}' M_*^{\perp} [M_*^{\perp}]' \in \mathcal{R}^{\perp}(M_*) \quad . \end{aligned} \tag{3.9}$$

with

$$\boldsymbol{\pi}_{\mathbf{x}} = \boldsymbol{\pi}_{*\parallel} + \boldsymbol{\pi}_{*\perp} \quad . \tag{3.10}$$

[1] As $G$ is not necessarily symmetric, it is possible for $G$ to have complex eigenvectors. However, since $G$ is real, such complex eigenvectors will always come in complex conjugate pairs and will be associated with eigenvalues that are complex conjugates. When the matrix $G$ is multiplied by a real vector and such complex eigenvectors are excited, the result must be real (obviously) and will be a linear combination of the real part and imaginary part of such complex eigenvectors. Thus, it is not the complex eigenvectors that will contribute to the basis for $\mathcal{R}(M_*)$, but rather, their real and imaginary parts.

If $M^{(1)}$-separability was assumed for our weather example, $\boldsymbol{\pi}_{*\parallel}$ would contain all information about the univariate marginal distributions for the weather in each of the four cities, i.e., it would provide information equivalent to the expected value of the state array vector, $\mathbb{E}\left[\mathbf{s}^{(1)}\right]$. Alternatively, $\boldsymbol{\pi}_{*\perp}$ would contain all 'orthogonal' information, the additional information from which one could construct the joint distribution (the distribution for the network state) from the univariate marginals. In a sense, $\boldsymbol{\pi}_{*\perp}$ would provide the information about how the statuses of the different automata would be coupled. Under $M^{(1)}$-separability, it would be assumed that such coupling information was invariant, meaning it would be irrelevant as far as updating the univariate marginals.

Because of linearity, the updated probability vector at time 1 can be computed by individually propagating the projections of $\boldsymbol{\pi}_{\mathbf{x}}[0]$ onto the complementary orthogonal spaces and then summing such propagated projections, i.e.,

$$\boldsymbol{\pi}_{\mathbf{x}}[1]' = \boldsymbol{\pi}_{*\parallel}[0]'G + \boldsymbol{\pi}_{*\perp}[0]'G \quad . \tag{3.11}$$

However, if one was only interested in the projection of $\boldsymbol{\pi}_{\mathbf{x}}[1]$ onto $\mathcal{R}(M_*)$ (in the context of our weather example, one would only be interested in univariate marginal distributions for each city's weather), because of the assumption of $M_*$-separability and its left-invariance,

$$\boldsymbol{\pi}_{*\parallel}[1] = \boldsymbol{\pi}_{*\parallel}[0]G \quad , \tag{3.12}$$

as by the assumption of left-invariance, $\boldsymbol{\pi}_{*\parallel}[0]GM_*M_*^{-L} = 0$. By induction, it follows that when initializing the network with probability vector $\boldsymbol{\pi}_{\mathbf{x}}[\tau]$,

$$\boldsymbol{\pi}_{*\perp}[t]' = \boldsymbol{\pi}_{*\perp}[\tau]'G^{t-\tau} \quad . \tag{3.13}$$

By then multiplying both sides of (3.13) by $M_*$, we obtain our familiar update equation (2.31) for our partial information vector $\boldsymbol{\pi}_*$. It is because the dynamics of $\boldsymbol{\pi}_{*\perp}$ remain trapped in $\mathcal{R}^{\perp}(M_*)$, it permits low-order representations of the projected dynamics of $\boldsymbol{\pi}_{\mathbf{x}}$ in $\mathcal{R}(M_*)$. Fig. 3-1 attempts to schematically illustrate this phenomena in three dimensions.

For typical cases of $M_*$-separability, the Markov chain exhibits a left-invariant subspace with meaningful structure, $\mathcal{R}(M_*^{\perp})$, whose orthogonal complement, $\mathcal{R}(M_*)$, is of low-order, i.e., $p << \eta$, where $p$ denotes the dimension of $\mathcal{R}(M_*)$, and $\eta$ is the dimension of the state space. This allows one to monitor dynamics tractably in the orthogonal complement $\mathcal{R}(M_*)$. Consider again the case of $M^{(1)}$-separability. Rather than tracking the dynamics of the $\eta \times 1$ vector $\boldsymbol{\pi}_{*\parallel}$ directly, which offers no computational advantages, one tracks the dynamics of the substantially smaller vector $\boldsymbol{\pi}^{(1)'} \triangleq \boldsymbol{\pi}_{\mathbf{x}}' M^{(1)}$ that parameterizes vectors in $\mathcal{R}(M^{(1)})$. As a function of the number of automata $n$, the size of $\boldsymbol{\pi}_{\mathbf{x}}$ is growing exponentially, while $\boldsymbol{\pi}^{(1)}$ is growing linearly. Considering our weather example where $m = 3$ and $n = 4$, we can monitor the dynamics of the evolving univariate marginals via a $12 \times 12$ matrix $H^{(1)}$, under the assumption of $M^{(1)}$-separability, which is substantially smaller than the $81 \times 81$ transition matrix $G$. These apparent computation advantages offered by $M_*$-separability motivate our interest in its development.

### 3.1.2 Correspondence of eigenvalues/eigenvectors of $G$ and $H_*$

We have already noted that under $M_*$-separability, $G$ has a set of $p$ right eigenvectors (and generalized right eigenvectors) spanning the $p$-dimensional subspace $\mathcal{R}(M_*)$. Let's visualize this subset of right eigenvectors (and generalized right eigenvectors) of $G$ in Jordan canonical

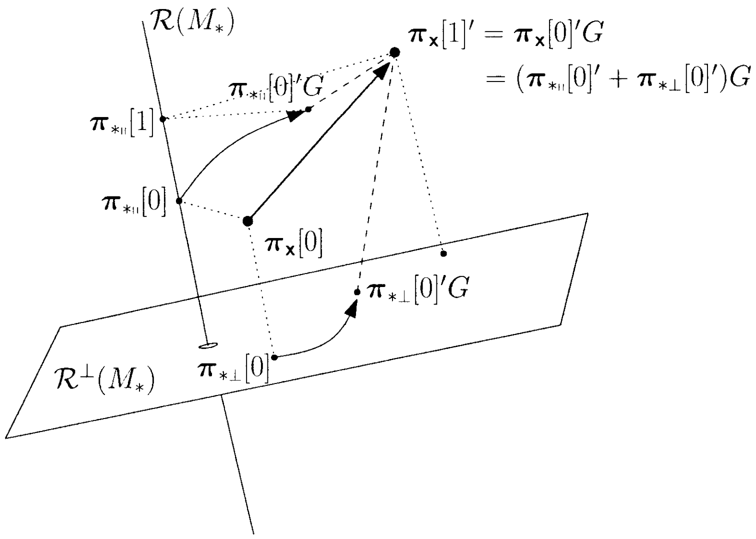

$\mathcal{R}(M_*)$
$\boldsymbol{\pi}_{\mathsf{x}}[1]' = \boldsymbol{\pi}_{\mathsf{x}}[0]'G$
$= (\boldsymbol{\pi}_{*_{\parallel}}[0]' + \boldsymbol{\pi}_{*\perp}[0]')G$
$\boldsymbol{\pi}_{*_{\parallel}}[0]'G$
$\boldsymbol{\pi}_{*_{\parallel}}[1]$
$\boldsymbol{\pi}_{*_{\parallel}}[0]$
$\boldsymbol{\pi}_{\mathsf{x}}[0]$
$\boldsymbol{\pi}_{*\perp}[0]'G$
$\mathcal{R}^{\perp}(M_*)$
$\boldsymbol{\pi}_{*\perp}[0]$

Figure 3-1: This schematic attempts to illustrate the invariance of $M_*$-separability. Two complementary subspaces are pictured, the line $\mathcal{R}(M_*)$, and the plane, $\mathcal{R}^{\perp}(M_*)$. An initial probability vector $\boldsymbol{\pi}_{\mathsf{x}}[0]$ is shown, and its projections onto the two subspaces, $\boldsymbol{\pi}_{*_{\parallel}}[0]$ and $\boldsymbol{\pi}_{*\perp}[0]$, respectively, are indicated by dotted lines (all dotted lines in the picture indicate projections). By linearity, $\boldsymbol{\pi}_{\mathsf{x}}[0]$ can be propagated by $G$ to define $\boldsymbol{\pi}_{\mathsf{x}}[1]$, or its propagated projections can be summed. Because $G$ is left-invariant with respect to $\mathcal{R}^{\perp}(M_*)$, the projection of $\boldsymbol{\pi}_{\mathsf{x}}[1]$ onto $\mathcal{R}(M_*)$ is the same as the projection of $\boldsymbol{\pi}_{*_{\parallel}}[0]'G$ onto $\mathcal{R}(M_*)$.

form [39]:

$$G = \begin{bmatrix} X & \widetilde{X} \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \widetilde{\Lambda} \end{bmatrix} \begin{bmatrix} V' \\ \widetilde{V}' \end{bmatrix} \quad , \tag{3.14}$$

where $X$ is an $\eta \times p$ matrix whose columns are the right eigenvectors (and generalized right eigenvectors) spanning $\mathcal{R}(M_*)$; $\widetilde{X}$ is an $\eta \times (\eta - p)$ matrix of additional right eigenvectors (and generalized right eigenvectors) as its columns; $\Lambda$ and $\widetilde{\Lambda}$ are $p \times p$ and $(\eta - p) \times (\eta - p)$ block diagonal matrices, respectively, consisting of Jordan blocks with the eigenvalues of $G$ on their diagonals; $V'$ and $\widetilde{V}'$ are matrices whose rows are left eigenvectors (and generalized left eigenvectors) of $G$.[2] Note that all matrices are partitioned commensurately.

By nature of being in Jordan canonical form, the matrix of right eigenvectors must be the inverse of the matrix of left eigenvectors, and hence,

$$\begin{bmatrix} V' \\ \widetilde{V}' \end{bmatrix} \begin{bmatrix} X\widetilde{X} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \quad . \tag{3.15}$$

We will denote the $i$th column of $X$ as $\boldsymbol{x}_i$, the $i$th column of $V$ as $\boldsymbol{v}_i$, and the $i$th diagonal element of $\Lambda$ as $\lambda_i$. We will refer to $\boldsymbol{x}_i$ and $\boldsymbol{v}_i$ as a right/left eigenvector pair associated with the eigenvalue $\lambda_i$. It should be evident from (3.15) that as $\widetilde{V}'X = 0$, the rows of $\widetilde{V}'$ must be the $(\eta - p)$ left eigenvectors spanning $\mathcal{R}(M_*^{\perp})$ guaranteed under $M_*$-separability.

[2] Up to this point, we have been careful to include the possibility of generalized eigenvectors for many of our statements, when appropriate. For the remainder of this section, 'eigenvectors' will refer to both eigenvectors and generalized eigenvectors, collectively. If general eigenvectors shall be excluded when discussing 'eigenvectors,' this will be explicitly noted.

Thus it is evident from the Jordan canonical form that the right/left eigenvector pairs of $G$ can be partitioned into two sets: one set whose right eigenvectors span $\mathcal{R}(M_*)$ and the other set whose left eigenvectors span $\mathcal{R}(M_*^\perp)$.

As before, assume that $M_*$ has full column rank. Recall (3.1) and consider the equation obtained for $H_*$ when both sides are multiplied by $M_*^{-L}$:

$$H_* = M_*^{-L} G M_* \quad . \tag{3.16}$$

By substituting $G$'s Jordan canonical form (3.14) into (3.16), we obtain

$$H_* = M_*^{-L} \begin{bmatrix} X & \widetilde{X} \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \widetilde{\Lambda} \end{bmatrix} \begin{bmatrix} V' \\ \widetilde{V}' \end{bmatrix} M_* \quad . \tag{3.17}$$

By (3.15), it follows that $\widetilde{V}'X = 0$, and as $M_*$ and $X$ have the same range space (the columns of $X$ consist of the right eigenvectors of $G$ spanning $\mathcal{R}(M_*)$), it also follows that $\widetilde{V}'M_* = 0$. Using this fact, we can simplify (3.17) as follows:

$$\begin{aligned} H_* &= \begin{bmatrix} M_*^{-L}X & M_*^{-L}\widetilde{X} \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \widetilde{\Lambda} \end{bmatrix} \begin{bmatrix} V'M_* \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} M_*^{-L}X & M_*^{-L}\widetilde{X} \end{bmatrix} \begin{bmatrix} \Lambda V'M_* \\ 0 \end{bmatrix} \\ &= M_*^{-L} X \Lambda V' M_* \quad . \end{aligned} \tag{3.18}$$

Note that the $p \times p$ matrix $\Lambda$ is block diagonal consisting of Jordan blocks. Moreover, observe that

$$\begin{aligned} \left(V'M_*\right)\left(M_*^{-L}X\right) &= V'X \\ &= I \quad , \end{aligned} \tag{3.19}$$

where the first equality follows from the fact that $M_*M_*^{-L}$ acts as an identity when right-multiplied by column vectors in $\mathcal{R}(M_*)$ (the columns of $X$ span $\mathcal{R}(M_*)$), and the second equality follows from (3.15). Combining these two facts, it follows that $H_*$ is represented in Jordan canonical form in (3.18). Evidently, the eigenvalues and eigenvectors of $H_*$ are directly inherited from $G$. For each $\boldsymbol{x}_i$ (a right eigenvector of $G$ in the column space of $M_*$) and its paired left eigenvector $\boldsymbol{v}_i$, both being associated with an eigenvalue of $\lambda_i$,

- $M_*^{-L}\boldsymbol{x}_i$ is a right eigenvector of $H_*$, and
- $M_*'\boldsymbol{v}_i$ is a left eigenvector of $H_*$,

both associated with an eigenvalue of $\lambda_i$. Conversely, [3] for any right/left eigenvector pair

[3] The astute reader may be concerned with the fact that, in general, a matrix's Jordan canonical form need not be unique (when there exist eigenvalues with geometric multiplicity greater than 1, i.e., when *eigenspaces* of dimension greater than 1 exist, the choice of a linearly independent set of normalized eigenvectors and generalized eigenvectors is not unique). As we have only derived a particular Jordan canonical form for $H_*$, the natural question to ask is whether the eigenvector pairs suggested by *any* Jordan canonical form for $H_*$ share such a correspondence with eigenvector pairs of $G$. The answer, thankfully, is yes. Although the Jordan canonical form may not necessarily be unique, the eigenspaces and their associated eigenvalues (as well as any generalized eigenspaces and their associated eigenvalues) are unique. The eigenvectors and eigenvalues of $G$ and $H_*$ could just as well have been related in terms of their unique eigenspaces, instead of relating their potentially not unique left/right eigenvector pairs.

of $H_*$, $\overline{\boldsymbol{x}}_i$ and $\overline{\boldsymbol{v}}_i$, associated with an eigenvalue of $\lambda_i$,

- $M_*\overline{\boldsymbol{x}}_i$ is a right eigenvector of $G$ associated with the eigenvalue $\lambda_i$, and
- although a left eigenvector of $G$ cannot be uniquely identified from a left eigenvector of $H_*$, it is known that $\overline{\boldsymbol{v}}_i = M_*'\boldsymbol{v}_i$, for some $\boldsymbol{v}_i$ that is a left eigenvector of $G$ associated with the eigenvalue of $\lambda_i$.

As evident from (3.19), we have defined a linearly independent set of left eigenvectors of $H_*$ as well as a linearly independent set of right eigenvectors of $H_*$. Thus, all left/right eigenvectors of $H_*$, as well as the associated eigenvalues, are inherited from $G$. The spectrum of $H_*$ (its eigenvalues) is the subset of the spectrum of $G$, and specifically, consists of the eigenvalues corresponding to the right eigenvectors of $G$ spanning $\mathcal{R}(M_*)$.

This has important implications for the dynamics of our network as monitored in $\mathcal{R}(M_*)$, via (2.31). Such dynamics will only be affected by eigenvalues inherited from $G$. From the spectrum of $H_*$, one can obtain lower bounds on the second largest eigenvalue of the network's transition matrix $G$, as well as lower bounds on settling times for the network [40].

We note that in the case of $\mathcal{M}$-separability, for each $M_* \in \mathcal{M}$, we have such left/right invariance and a correspondence of eigenvalues and eigenvectors between $G$ and $H_*$.

### When $M_*$ fails to have full rank

Up to this point, it has been assumed that $M_*$ has full column rank. Suppose that $M_*$-separability is assumed for a particular $M_*$ that lacks full column rank. Such is the case for $M^{(r)}$-separability, for any value of $r \geq 1$. Let $p$ denote the rank of $M_*$ as before, and $\overline{p} > p$ its number of columns. Nothing changes regarding $G$: it will still have a basis of right eigenvalues spanning $\mathcal{R}(M_*)$ and $\eta - p$ left-eigenvalues in $\mathcal{R}(M_*^\perp)$. What may change are the characteristics of an $H_*$ that satisfies (3.1). Because $M_*$ lacks full column rank, it will have a null space of dimension $\overline{p} - p$. Therefore, an $H_*$ that solves (3.1) will not be unique. For a given $G$, an entire affine subspace of matrices will satisfy (3.1). In contrast, recall that $H_*$ is unique when $M_*$ has full column rank (3.16).

Let's look at this affine subspace of $\overline{p} \times \overline{p}$ matrices. A particular matrix satisfying (3.1) can be determined by solving (3.1) column by column, and minimizing the 2-norm for the selected solution for each column. What we obtain is the matrix of minimal Frobenius norm [39] amongst all matrices solving (3.1). We denote such a matrix as $H_0$; this matrix $H_0$ can be expressed as

$$H_0 = M_*^\dagger G M_* \quad , \tag{3.20}$$

where $M_*^\dagger$ denotes the Moore-Penrose generalized inverse of $M_*$ [39]. Note that the rank of $H_0$ cannot exceed $p$, as both its row space and column space are subsets of the $p$-dimensional subspace $\mathcal{N}^\perp(M_*)$ (the row space of $M_*$).

By the same arguments that allowed us to derive the Jordan canonical form for $H_*$ when $M_*$ has full column rank (3.18), we find that

$$H_0 = M_*^\dagger X \Lambda V' M_* \quad . \tag{3.21}$$

$H_0$ expressed as in (3.21) is almost in Jordan canonical form; by augmenting the decomposition to include $\mathcal{N}(M_*)$, $H_0$ will be expressed in Jordan canonical form. Specifically, let

$N_*$ be a $\overline{p} \times (\overline{p} - p)$ matrix with orthonormal columns that span $\mathcal{N}(M_*)$.[4] We can rewrite (3.21) as a product of $\overline{p} \times \overline{p}$ matrices as follows:

$$H_0 = \begin{bmatrix} M_*^\dagger X & N_* \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V'M_* \\ N_*' \end{bmatrix} \quad . \tag{3.22}$$

Note that the middle matrix on the right hand side of (3.22) is a block diagonal matrix consisting of Jordan blocks, and that the product of the outer matrices of (3.22) satisfies

$$\begin{aligned} \begin{bmatrix} V'M_* \\ N_*' \end{bmatrix} \begin{bmatrix} M_*^\dagger X & N_* \end{bmatrix} &= \begin{bmatrix} V'M_*M_*^\dagger X & 0 \\ 0 & N_*'N_* \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \quad , \end{aligned} \tag{3.23}$$

where the first equality follows from the fact that since the columns of $N_*$ form a basis for $\mathcal{N}(M_*)$, both $M_*N_* = 0$ and $N_*'M_*^\dagger = 0$. The second equality in (3.23) follows by definition of $N_*$ having orthonormal columns ($N_*'N_* = I$) and from the fact that as $X \in \mathcal{R}(M_*)$, $M_*M_*^\dagger X = X$, while $V'X = I$ by (3.19). From these observations, it follows that $H_0$ is in Jordan canonical form in (3.22), and thus, as in the case when $M_*$ has full column rank, $H_0$ inherits all of its eigenvectors (and generalized eigenvectors) associated with nonzero eigenvalues from $G$. For each $\boldsymbol{x}_i$ (a right eigenvector of $G$ in the column space of $M_*$) and its paired left eigenvector $\boldsymbol{v}_i$, both being associated with an eigenvalue of $\lambda_i$,

- $M_*^\dagger \boldsymbol{x}_i$ is a right eigenvector of $H_0$,
- $M_*'\boldsymbol{v}_i$ is a left eigenvector of $H_0$,

both being associated with an eigenvalue of $\lambda_i$. Conversely, for any right/left eigenvector pair of $H_0$, $\overline{\boldsymbol{x}}_i$ and $\overline{\boldsymbol{v}}_i$, associated with an eigenvalue of $\lambda_i \neq 0$,

- $M_*\overline{\boldsymbol{x}}_i$ is a right eigenvector of $G$ associated with the eigenvalue $\lambda_i$, and
- although a left eigenvector of $G$ cannot be uniquely identified from the left eigenvector of $H_0$, it is know that $\overline{\boldsymbol{v}}_i = M_*'\boldsymbol{v}_i$, for some $\boldsymbol{v}_i$ that is a left eigenvector of $G$ associated with the eigenvalue $\lambda_i$.

We have examined the spectral structure of a particular matrix $H_0$ that solves (3.1) (the one of minimal Frobenius norm), now we wish to examine the spectral structure of other matrices that solve (3.1). By linear algebra, any matrix solving (3.1) for a given $G$ and $M_*$ can be expressed as the summation of $H_0$ with a matrix whose columns are drawn from $\mathcal{N}(M_*)$, i.e.,

$$H_*(Z) = H_0 + N_*Z \quad , \tag{3.24}$$

where $N_*$ is defined as before (a $\overline{p} \times (\overline{p} - p)$ matrix whose columns form a basis for $\mathcal{N}(M_*)$), and $Z \in \mathbb{R}^{(\overline{p}-p)\times\overline{p}}$ is otherwise unconstrained. Each unique $Z$ corresponds to a unique matrix in the affine subspace of matrices satisfying (3.1) for a given $G$ and $M_*$, and thereby parameterizes the affine subspace. Note that the columns of $H_0$ are always orthogonal to the columns of $N_*Z$.

[4] The careful reader will notice that in representing such a matrix as $N_*$, we have departed from our traditional notation for matrices with orthonormal columns spanning the orthogonal complement to a matrix's range space. Rather than defining a new matrix $N_*$, our convention would be to express such a matrix as $(M_*')^\perp$. In this case, we found such notation overwhelming.

It can be shown that $H_*(Z)$ inherits the left eigenvectors of $H_0$ associated with nonzero eigenvalues. Consider any left eigenvector of $H_0$, $\overline{\boldsymbol{v}}_i$, in the row space of $M_*$. As the columns of $N_*$ span $\mathcal{N}(M_*)$,

$$\begin{aligned} \overline{\boldsymbol{v}}_i' H_*(Z) &= \overline{\boldsymbol{v}}_i' H_0 + \overline{\boldsymbol{v}}_i' N_* Z \\ &= \lambda_i \overline{\boldsymbol{v}}_i' \quad , \end{aligned} \tag{3.25}$$

and thus $\overline{\boldsymbol{v}}_i$ and its associated eigenvalue of $\lambda_i$ are a left eigenvector and eigenvalue of any $H_*$ satisfying (3.1). On the other hand, $\overline{\boldsymbol{v}}_i$'s paired right eigenvector in $H_*(Z)$ need not coincide with $\overline{\boldsymbol{v}}_i$'s paired right eigenvector in $H_0$, $\overline{\boldsymbol{x}}_i$. As $Z$ is unrestricted, $N_* Z \boldsymbol{x}_i$ may be nonzero, meaning that $\overline{\boldsymbol{x}}_i$ would not be a right eigenvector of $H_*(Z)$. The only restriction on any $\overline{\boldsymbol{v}}_i$'s paired right eigenvector in $H_*(Z)$ would be that it must have a nonzero projection onto the row space of $M_*$.

It can be similarly shown that $H_*(Z)$ acquires right eigenvectors and associated eigenvalues from $N_* Z$. As the matrix $N_* Z$'s columns span $\mathcal{N}(M_*)$, it can always be expressed in Jordan canonical form as

$$N_* Z = \begin{bmatrix} \overline{\overline{X}} & * \end{bmatrix} \begin{bmatrix} \overline{\overline{\Lambda}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \overline{\overline{V}}' \\ * \end{bmatrix} \quad , \tag{3.26}$$

where $\overline{\overline{X}}$ is a $\overline{p} \times (\overline{p} - p)$ matrix whose columns are right eigenvectors of $H_*(Z)$ spanning $\mathcal{N}(M_*)$, $\overline{\overline{\Lambda}}$ is a $(\overline{p} - p) \times (\overline{p} - p)$ Jordan block with the associated eigenvalues on the diagonal, and $\overline{\overline{V}}'$ would be a $p \times (\overline{p} - p)$ matrix with left eigenvectors as its rows, which are paired with the right eigenvectors of $N_* Z$ given in $\overline{\overline{X}}$. Consider any right eigenvector of $N_* Z$ in the null space of $M_*$, denoted as $\overline{\overline{\boldsymbol{x}}}_i$ and associated with the eigenvalue $\overline{\overline{\lambda}}_i$. As $H_0$ shares the same row space as $M_*$,

$$\begin{aligned} H_*(Z) \overline{\overline{\boldsymbol{x}}}_i &= H_0 \overline{\overline{\boldsymbol{x}}}_i + N_* Z \overline{\overline{\boldsymbol{x}}}_i \\ &= \overline{\overline{\lambda}}_i \overline{\overline{\boldsymbol{x}}}_i \quad , \end{aligned} \tag{3.27}$$

meaning that $\overline{\overline{\boldsymbol{x}}}_i$ is also a right eigenvector of $H_*(Z)$ associated with the eigenvalue $\overline{\overline{\lambda}}_i$. Because such right eigenvectors of $H_*(Z)$ lie in $\mathcal{N}(M_*)$, they cannot be the right eigenvector pairs to $H_*(Z)$'s known left eigenvectors in the row space of $M_*$ that are acquired from $H_0$. Hence we can conclude that $H_*(Z)$ can be expressed in Jordan canonical form as follows:

$$H_*(Z) = \begin{bmatrix} * & \overline{\overline{X}} \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \overline{\overline{\Lambda}} \end{bmatrix} \begin{bmatrix} V' M_* \\ * \end{bmatrix} \quad , \tag{3.28}$$

where the columns of $\overline{\overline{X}}$ are the $\overline{p} - p$ right eigenvectors that $H_*(Z)$ inherits from $N_* Z$, and the rows of $V' M_*$ are the $p$ left eigenvectors that $H_*(Z)$ inherits from $H_0$. Evidently, the eigenvector pairs of $H_*(Z)$ can be partitioned into two sets: one whose left eigenvectors and corresponding eigenvalues are copied from $H_0$, and a second whose right eigenvectors and corresponding eigenvalues are copied from $N_* Z$.

Clearly when $Z \neq 0$, $H_*(Z)$ will have additional left eigenvectors paired with the right eigenvectors it acquires from $N_* Z$. As these additional left eigenvectors will be paired with right eigenvectors in $\mathcal{N}(M_*)$, such left eigenvectors and their eigenvalues will not be excited

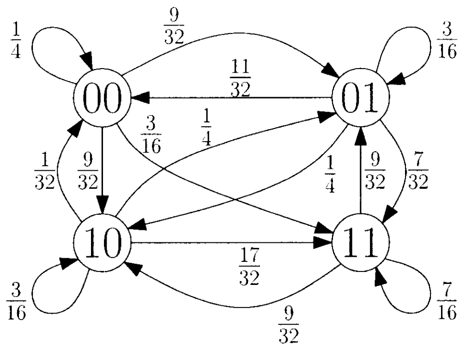


Figure 3-2: Markov Chain Example

when $H_*(Z)$ is left-multiplied by a row vector in the column space of $M_*$—and this will always be the case when we properly initialize the partial information dynamics in terms of $H_*(Z)$ (2.31) with the linear transformation of a probability vector, $\boldsymbol{\pi}'_{\mathbf{x}}M_*$. It is why any such additional left eigenvectors of $H_*(Z)$ not restricted to the row space of $M_*$ are referred to as "irrelevant" in [12] for the special case of the IM's $M^{(1)}$-separability.

By (3.24), any matrix $H_*$ that satisfies (3.1) can be expressed as a sum of two matrices, one ($H_0$) whose row and column spaces are subsets of the row space of $M_*$, and a second ($N_*Z$) whose column space is a subset of $\mathcal{N}(M_*)$ and row space is arbitrary.

### Example

We feel that it is instructive to illustrate these eigenvalue and eigenvector relationships via an example. Suppose we have two automata ($n = 2$), each with two possible statuses ($m = 2$): 0 ('off') or 1 ('on'). Assume these automata are modeled via the Markov chain illustrated in Fig. 3-2, where the state label '01' represents the state when automaton 1 is off and automaton 2 is on. The other network state labels can be similarly decoded.

The transition matrix for such a network would be

$$G = \begin{bmatrix} \frac{1}{4} & \frac{9}{32} & \frac{9}{32} & \frac{3}{16} \\ \frac{11}{32} & \frac{3}{16} & \frac{1}{4} & \frac{7}{32} \\ \frac{1}{32} & \frac{1}{4} & \frac{3}{16} & \frac{17}{32} \\ 0 & \frac{9}{32} & \frac{9}{32} & \frac{7}{16} \end{bmatrix} \quad , \tag{3.29}$$

and it is straightforward to show that such a network is $M^{(1)}$-separable by applying (3.2):

$$[1 \ -1 \ -1 \ 1]G \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} = [0 \ 0 \ 0 \ 0] \quad , \tag{3.30}$$

where evidently, $\left[M^{(1)}\right]^{\perp} = (1,\ -1,\ -1,\ 1)$.

By nature of being $M^{(1)}$-separable, we know that $G$'s left/right eigenvector pairs can be partitioned into two sets: one whose right eigenvectors span $\mathcal{R}(M^{(1)})$, and a second whose left eigenvectors span $\mathcal{R}([M^{(1)}]^{\perp}$. The first set's right eigenvectors spanning $\mathcal{R}(M^{(1)})$ are as follows:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} , \quad \boldsymbol{x}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} , \quad \boldsymbol{x}_3 = \begin{bmatrix} 5 \\ 5 \\ -3 \\ -3 \end{bmatrix} , \tag{3.31}$$

which are paired with the left eigenvectors

$$\boldsymbol{v}_1 = \begin{bmatrix} \frac{1}{8} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{3}{8} \end{bmatrix} , \quad \boldsymbol{v}_2 = \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix} , \quad \boldsymbol{v}_3 = \begin{bmatrix} \frac{1}{8} \\ 0 \\ 0 \\ -\frac{1}{8} \end{bmatrix} , \tag{3.32}$$

each associated with the eigenvalues $\lambda_1 = 1$, $\lambda_2 = -1/16$, and $\lambda_3 = 1/4$, respectively.

On the other hand, $G$ must also leave $\mathcal{R}(\,[M^{(1)}]^{\perp}\,)$ left-invariant, meaning it must have left eigenvectors spanning $[M^{(1)}]^{\perp}$. As $[M^{(1)}]^{\perp}$ is merely a 1-dimensional subspace, the requirement reduces to $G$ having a left eigenvector

$$\boldsymbol{v}_4 = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{bmatrix}' , \tag{3.33}$$

which will be paired with the right eigenvector

$$\boldsymbol{x}_4 = \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}' , \tag{3.34}$$

both associated with an eigenvalue $\lambda_4 = -1/8$ (note that the left and right eigenvectors of the pair are the same, apart from scaling). As there is only a single eigenvalue of modulus 1, $\boldsymbol{v}_1'$ must be the steady-state vector of the Markov chain.

Let's now consider a matrix $H^{(1)}$, satisfying (3.1) for $M_* = M^{(1)}$, which, like $G$, will be a $4 \times 4$ matrix (this a consequence of the anomaly $m_1 + m_2 = m_1 m_2$; typically $H^{(1)}$ is a much smaller matrix than $G$). As $M^{(1)}$ does not have full column rank, $H^{(1)}$ need not be unique. We can find a specific matrix $H_0^{(1)}$ via (3.20),

$$H_0^{(1)} = \left[M^{(1)}\right]^{\dagger} G M^{(1)} = \begin{bmatrix} \frac{21}{64} & \frac{11}{64} & \frac{23}{64} & \frac{9}{64} \\ \frac{5}{64} & \frac{27}{64} & \frac{3}{64} & \frac{29}{64} \\ \frac{13}{64} & \frac{19}{64} & \frac{11}{64} & \frac{21}{64} \\ \frac{13}{64} & \frac{19}{64} & \frac{15}{64} & \frac{17}{64} \end{bmatrix} . \tag{3.35}$$

As $H_0^{(1)}$ of (3.35) is the matrix of minimal Frobenius norm amongst the affine subspace of matrices satisfying (3.1) for $M_* = M^{(1)}$, all of its left and right eigenvectors associated with nonzero eigenvalues will correspond to the the subset of right/left eigenvector pairs of $G$ whose right eigenvectors span $\mathcal{R}(M^{(1)})$. In particular, by (3.22), $H_0^{(1)}$ will have right eigenvectors

$$\begin{aligned}\overline{\boldsymbol{x}}_1 &= \left[M^{(1)}\right]^{\dagger}\boldsymbol{x}_1 = \begin{bmatrix}\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}\end{bmatrix}' \\ \overline{\boldsymbol{x}}_2 &= \left[M^{(1)}\right]^{\dagger}\boldsymbol{x}_2 = \begin{bmatrix}\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2}\end{bmatrix}' \\ \overline{\boldsymbol{x}}_3 &= \left[M^{(1)}\right]^{\dagger}\boldsymbol{x}_3 = \begin{bmatrix}\frac{9}{2} & -\frac{7}{2} & \frac{1}{2} & \frac{1}{2}\end{bmatrix}' \end{aligned} \tag{3.36}$$

paired with left eigenvectors

$$\begin{aligned}\overline{\boldsymbol{v}}_1 &= \left[M^{(1)}\right]'\boldsymbol{v}_1 = \begin{bmatrix}\frac{3}{8} & \frac{5}{8} & \frac{3}{8} & \frac{5}{8}\end{bmatrix}' \\ \overline{\boldsymbol{v}}_2 &= \left[M^{(1)}\right]'\boldsymbol{v}_2 = \begin{bmatrix}0 & 0 & -1 & 1\end{bmatrix}' \\ \overline{\boldsymbol{v}}_3 &= \left[M^{(1)}\right]'\boldsymbol{v}_3 = \begin{bmatrix}\frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8}\end{bmatrix}' \end{aligned} \tag{3.37}$$

associated with eigenvalues $\overline{\lambda}_1 = 1$, $\overline{\lambda}_2 = -1/16$, and $\overline{\lambda}_3 = 1/4$, respectively. Note that both the left and right eigenvectors of $H_0^{(1)}$ given in (3.36) and (3.37) lie in the row space of $M^{(1)}$. It may seem peculiar that the rows of $H_0^{(1)}$ sum to 1, and moreover are nonnegative. Naturally the rows sum to 1 by (3.36), which states that $\overline{\boldsymbol{x}}_1 = \frac{1}{2}\mathbb{1}$ must be a right eigenvector of $H_0^{(1)}$ associated with an eigenvalue at 1. The fact that $H_0^{(1)}$ is nonnegative is special to our example and not guaranteed in general.

In addition, $H_0^{(1)}$ will have an additional right and left eigenvector pair associated with the eigenvalue $\overline{\lambda}_4 = 0$,

$$\overline{\boldsymbol{x}}_4 = \begin{bmatrix}\frac{1}{2}\\ \frac{1}{2}\\ -\frac{1}{2}\\ -\frac{1}{2}\end{bmatrix}, \quad \overline{\boldsymbol{v}}_4 = \begin{bmatrix}\frac{1}{2}\\ \frac{1}{2}\\ -\frac{1}{2}\\ -\frac{1}{2}\end{bmatrix}, \tag{3.38}$$

respectively. Note that such additional right eigenvectors (those not inherited from $G$) for a matrix satisfying (3.1) of minimal Frobenius norm will always be associated with eigenvalues of 0 and will span $\mathcal{N}(M^{(1)})$ (again, recall (3.22)). The same can be said for the additional left eigenvectors.

As $M^{(1)}$ does not have full column rank, there is an affine subspace of matrices that solve (3.1) for $M_* = M^{(1)}$. Using (3.24), let's consider another matrix in this affine subspace,

$$H^{(1)}(\boldsymbol{z}) = H_0^{(1)} + \begin{bmatrix}\frac{1}{2}\\ \frac{1}{2}\\ -\frac{1}{2}\\ -\frac{1}{2}\end{bmatrix}\boldsymbol{z}' \ . \tag{3.39}$$

As argued in (3.25), $\overline{\boldsymbol{v}}_1$, $\overline{\boldsymbol{v}}_2$, and $\overline{\boldsymbol{v}}_3$ must be left eigenvectors of $H^{(1)}(\boldsymbol{z})$ associated with eigenvalues $\overline{\lambda}_1$, $\overline{\lambda}_2$, and $\overline{\lambda}_3$, respectively, for any choice of $\boldsymbol{z}$. However, by choosing $\boldsymbol{z}$ strategically, any desired left eigenvector and eigenvalue can be introduced provided that the left eigenvector is not in the row space of $M^{(1)}$. As evident from (3.39), the right eigenvectors paired with such 'new' left eigenvectors will be restricted to being in $\mathcal{N}(M^{(1)})$, and hence will not be excited by left multiplication of $H^{(1)}(\boldsymbol{z})$ by row vectors in the row space of $M^{(1)}$.

Suppose that we want

$$\overline{\boldsymbol{v}}_4 = \begin{bmatrix} 2 & 0 & 0 & 0 \end{bmatrix}' \tag{3.40}$$

to be a left eigenvector of $H^{(1)}(\boldsymbol{z})$ associated with an eigenvalue of $\overline{\lambda}_4 = 2$. As $\mathcal{N}(M^{(1)})$ has a dimension of 1, $\overline{\boldsymbol{v}}_4$'s paired right eigenvector must be

$$\overline{\boldsymbol{x}}_4 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}' \quad . \tag{3.41}$$

After some algebra, we find that the necessary $\boldsymbol{z}$ to achieve such a left eigenvector (3.40) and associated eigenvalue $\overline{\lambda}_4 = 2$ is

$$\boldsymbol{z}' = \begin{bmatrix} \frac{107}{32} & -\frac{11}{32} & -\frac{23}{32} & -\frac{9}{32} \end{bmatrix} \quad , \tag{3.42}$$

and the resulting $H^{(1)}(\boldsymbol{z})$ for $\boldsymbol{z}$ specified as in (3.42) is

$$H^{(1)}(\boldsymbol{z}) = \begin{bmatrix} 2 & 0 & 0 & 0 \\ \frac{7}{8} & \frac{1}{4} & -\frac{5}{16} & -\frac{5}{16} \\ -\frac{47}{32} & \frac{15}{32} & \frac{17}{32} & \frac{15}{32} \\ -\frac{47}{32} & \frac{15}{32} & \frac{19}{32} & \frac{13}{32} \end{bmatrix} \quad . \tag{3.43}$$

As a double-check, one can verify that for $M_* = M^{(1)}$ and the $G$ specified in (3.29), $H^{(1)}(\boldsymbol{z})$ of (3.43) satisfies (3.1). Note that $H^{(1)}(\boldsymbol{z})$ as given in (3.43) does not have rows summing to 1 (as was guaranteed for $H_0^{(1)}$).

$H^{(1)}(\boldsymbol{z})$ inherits the left eigenvectors $\overline{\boldsymbol{v}}_1$, $\overline{\boldsymbol{v}}_2$, and $\overline{\boldsymbol{v}}_3$ of $H_0^{(1)}$ given in (3.37). The choice of $\boldsymbol{z}$ in (3.42) results in the paired right eigenvectors for $H^{(1)}(\boldsymbol{z})$ becoming

$$\overline{\boldsymbol{x}}_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \; , \quad \overline{\boldsymbol{x}}_2 = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \; , \quad \overline{\boldsymbol{x}}_3 = \begin{bmatrix} 0 \\ -8 \\ 5 \\ 5 \end{bmatrix} \quad . \tag{3.44}$$

As evident on comparing (3.36) to (3.44), the particular choice of $\boldsymbol{z}$ in (3.42) results in a substantial shift in the paired right eigenvectors.

With Corollary 3 in mind, one could define a matrix $\widetilde{M}^{(1)}$ with full column rank such that $\mathcal{R}(M^{(1)}) = \mathcal{R}(\widetilde{M}^{(1)})$. The unique $3 \times 3$ matrix satisfying (3.1) for $M_* = \widetilde{M}^{(1)}$, denoted by $\widetilde{H}^{(1)}$, could be computed using (3.16) and replace $H^{(1)}(\boldsymbol{z})$ in computations. There would be no ambiguity regarding the spectral structure of $\widetilde{H}^{(1)}$. The partial information expressed in terms of $M^{(1)}$, i.e., $\left[\boldsymbol{\pi}^{(1)}\right]' = \boldsymbol{\pi}_{\mathsf{x}}' M^{(1)}$, which is a sequence of probability vectors representing univariate marginal PMFs, could always be recovered from the partial information expressed in terms of $\widetilde{M}^{(1)}$, as

$$\left[\boldsymbol{\pi}^{(1)}\right]' = \left[\widetilde{\boldsymbol{\pi}}^{(1)}\right]' \left[\widetilde{M}^{(1)}\right]^{-L} M^{(1)} \quad . \tag{3.45}$$

For small network such as this example, working with $\widetilde{H}^{(1)}$ in place of $H^{(1)}$ offers limited computational advantages. However in large networks, the benefits of using $\widetilde{H}^{(1)}$ can be substantial. After computing the rank of $M^{(1)}$ in Section 3.4, one can determine precisely what the benefits would be.

## 3.2 The algebraic parametrization of separability

The objective of this section is to develop an abstract characterization of $\mathcal{G}(M_*)$, the set of transition matrices exhibiting $M_*$-separability. One may note that we already have an algebraic characterization of $\mathcal{G}(M_*)$: it consists of matrices $G$ that are row-stochastic and satisfy (3.1). We will combine these constraints into a unified algebraic characterization, with the intention of developing equations for the dimension of $\mathcal{G}(M_*)$, i.e., the minimum number of scalar parameters needed to specify uniquely any probabilistic model exhibiting $M_*$-separability, or equivalently, the dimension of the affine subspace of smallest dimension containing $\mathcal{G}(M_*)$. By (3.3), which expresses the additional constraints of $M_*$-separability as an $(\eta - p) \times p$ block of 0s, we expect the dimension of $\mathcal{G}(M_*)$ to be roughly $\eta^2 - (\eta - p)p$. As clear by Corollary 3 and our discussions of $M_*$-separability in Section 3.1, any equations for the dimension of $\mathcal{G}(M_*)$ should depend only on $\mathcal{R}(M_*)$; it will be shown that the dimension of $\mathcal{G}(M_*)$ is determined by the dimension of $\mathcal{R}(M_*)$ as well as whether or not $\mathbb{1}$ is an element of $\mathcal{R}(M_*)$. We will then similarly analyze the dimension of $\mathcal{G}(\mathcal{M})$. After developing some general mathematical machinery, these results will be applied to the canonical examples of separability, $M^{(r)}$-separability and $\mathcal{M}^{(\overline{r})}$-separability, in Section 3.4.

### 3.2.1 The feasibility set $\mathcal{G}(M_*)$ under $M_*$-separability

The feasibility set $\mathcal{G}(M_*)$, the set of transition matrices exhibiting $M_*$-separability, is characterized by satisfying two sets of constraints: row-stochasticity (nonnegativity combined with affine linear constraints) and the existence of a matrix $H_*$ satisfying (3.1) (linear constraints). Thus the feasibility set $\mathcal{G}(M_*)$ can be identified as the intersection of the nonnegative orthant (a polyhedral cone) with an affine subspace that captures both the linear constraints of (3.1) and the affine linear constraints of row-stochasticity. This intersection must be bounded (as a subset of the bounded set of row-stochastic matrices), and as it is the intersection of a polyhedral cone with an affine subspace, it must be a finitely-generated convex set (have a finite number of extreme points). Rather than differentiating between linear and affine linear constraints, we will subsequently refer to both as linear constraints. On the other hand, we will continue to differentiate between subspaces and affine subspaces.

By rearranging (3.2), the additional linear constraints imposed by $M_*$-separability can be expressed as a null space constraint on vec$(G)$:

$$\left(M_*' \otimes \left[M_*^{\perp}\right]'\right)\mathrm{vec}(G) = 0 \quad , \tag{3.46}$$

where we have used the fact that $\mathrm{vec}(ABC) = (C' \otimes A)\mathrm{vec}(B)$ (A.5). Thus by (3.46), $\mathcal{G}(M_*)$ must be a subset of the the null space (kernel) of $M_*' \otimes \left[M_*^{\perp}\right]'$.

There are additional constraints on any $G \in \mathcal{G}(M_*)$, namely that the matrix be row-stochastic. This imposes nonnegativity constraints, as well as the linear constraints that its rows must sum to 1, i.e.,

$$G\mathbb{1} = \mathbb{1} \quad . \tag{3.47}$$

Combining all linear constraints into a single equation, we obtain

$$\begin{bmatrix} M_*' \otimes [M_*^{\perp}]' \\ \mathbb{1}' \otimes I \end{bmatrix} \mathrm{vec}(G) = \begin{bmatrix} 0 \\ \mathbb{1} \end{bmatrix} \quad . \tag{3.48}$$

The dimension of the affine subspace of matrices $G$ that satisfy (3.48) is determined by the dimension of the null space of the matrix on the left hand side of (3.48), provided that the constraints are consistent. The matrix on the left hand side of (3.48) is an example of a *block Kronecker matrix*, a block matrix whose blocks are expressed in terms of Kronecker products. Section A.2.1 discusses techniques to calculate the rank of block Kronecker matrices. Since $\mathcal{R}(\mathbb{1}) \subset \mathcal{R}(I)$, the dimension of the null space of the matrix on the left hand side of (3.48) can be determined by applying Corollary 19 to its transpose. Note that the dimension of this null space will depend on whether or not $\mathbb{1} \in \mathcal{R}(M_*)$. If $\mathbb{1} \in \mathcal{R}(M_*)$, the null space will be of dimension $\eta^2 - p(\eta - p) - p$, where as before, $p$ is the rank of $M_*$. On the other hand, if $\mathbb{1} \notin \mathcal{R}(M_*)$, then the null space will be of dimension $\eta^2 - p(\eta - p) - \eta$.

Returning to our objective, which is to characterize $\mathcal{G}(M_*)$, we see that it can be identified as the intersection of the nonnegative orthant with the affine subspace of $\eta \times \eta$ matrices satisfying (3.48). Suppose that the affine subspace of matrices satisfying (3.48) intersects the interior of the nonnegative orthant, i.e., the positive orthant. As the nonnegative orthant is of full dimension $\eta \times \eta$, we are assured then that the dimension of $\mathcal{G}(M_*)$ will equal the dimension of the null space of the matrix on the left hand side of (3.48) (note that this condition is sufficient, yet not necessary). For the majority of examples of $M_*$-separability that we have introduced, including both $M^{(r)}$-separability and $M_j$-separability, such an intersection is assured: for any $\eta \times 1$ probability vector $\boldsymbol{\pi}_{\mathsf{X}} > 0$, the strictly positive matrix $\mathbb{1}\boldsymbol{\pi}_{\mathsf{X}}'$ exhibits both $M^{(r)}$-separability and $M_j$-separability. These results are summarized in the following theorem.

**Theorem 2.** *If there exists a strictly positive matrix that exhibits $M_*$-separability, then the dimension of $\mathcal{G}(M_*)$, the minimum number of scalar parameters needed to specify uniquely the transition matrix of any probabilistic model exhibiting $M_*$-separability, is given as follows:*

$$\dim\left(\mathcal{G}(M_*)\right) = \begin{cases} \eta^2 - p(\eta - p) - p & \text{if } \mathbb{1} \in \mathcal{R}(M_*) \\ \eta^2 - p(\eta - p) - \eta & \text{if } \mathbb{1} \notin \mathcal{R}(M_*) \end{cases} \quad , \tag{3.49}$$

*where $p = \mathrm{rank}(M_*)$, and $\dim(\cdot)$ denotes the dimension of a set, defined as the dimension of the affine subspace of smallest dimension enclosing it.*

If one would like to highlight the number of free parameters 'lost' in assuming $M_*$-separability, (3.49) can be rearranged as

$$\dim\left(\mathcal{G}(M_*)\right) = \begin{cases} \dim(\mathcal{G}) - (p-1)(\eta - p) & \text{if } \mathbb{1} \in \mathcal{R}(M_*) \\ \dim(\mathcal{G}) - p(\eta - p) & \text{if } \mathbb{1} \notin \mathcal{R}(M_*) \end{cases} \quad , \tag{3.50}$$

where $\mathcal{G}$ is the set of otherwise unrestricted $\eta \times \eta$ row-stochastic transition matrices of dimension $\eta(\eta - 1)$.

Although there exist cases where $\mathcal{G}(M_*)$ is of lower dimension than the subspace satisfying (3.48) (implying that there does not exist a strictly positive matrix in $\mathcal{G}(M_*)$), this intersection can never be empty. As evident from (3.1), $I \in \mathcal{G}(M_*)$ for all $M_*$.

### Visualizing the free parameters of $M_*$-separability

After having meticulously computed the number of free parameters under $M_*$-separability, we recall our comment in the introduction to this section that the number of free parameters under $M_*$-separability is effectively illustrated in (3.3) by $G$'s upper triangular form under the linear transformation specified by $\left[M_* \;\; M_*^\perp\right]$ (we have assumed that $M_*$ has full column rank). This equation can be rewritten in terms of $G$ as follows:

$$G = \left[M_* \;\; M_*^\perp\right] \begin{bmatrix} H_* & H_{(12)} \\ 0 & H_{(22)} \end{bmatrix} \left[M_* \;\; M_*^\perp\right]^{-1} \quad . \tag{3.51}$$

Expressing $G$ as in (3.51) intuitively allows one to visualize the free parameters under $M_*$-separability. The $p(\eta - p)$ zeros in the $(2,1)$ block of the middle matrix of the right hand side of (3.51) are the constraints imposed by (3.46), and the number of potentially nonzero entries in the matrix, $\eta^2 - p(\eta - p)$ is the dimension of the null space of (3.46). If we further consider the linear constraints that the rows of $G$ must sum to 1 (3.47), we obtain

$$\left[M_* \;\; M_*^\perp\right] \begin{bmatrix} H_* & H_{(12)} \\ 0 & H_{(22)} \end{bmatrix} \left[M_* \;\; M_*^\perp\right]^{-1} \mathbb{1} = \mathbb{1} \quad . \tag{3.52}$$

By multiplying both sides of (3.52) by $\left[M_* \;\; M_*^\perp\right]^{-1}$ and recalling (3.4), we obtain,

$$\begin{aligned} \begin{bmatrix} H_* & H_{(12)} \\ 0 & H_{(22)} \end{bmatrix} \left[M_* \;\; M_*^\perp\right]^{-1} \mathbb{1} &= \left[M_* \;\; M_*^\perp\right]^{-1} \mathbb{1} \\ \begin{bmatrix} H_* & H_{(12)} \\ 0 & H_{(22)} \end{bmatrix} \begin{bmatrix} M_*^{-L} \\ \left[M_*^\perp\right]' \end{bmatrix} \mathbb{1} &= \begin{bmatrix} M_*^{-L} \\ \left[M_*^\perp\right]' \end{bmatrix} \mathbb{1} \quad . \end{aligned} \tag{3.53}$$

The seemingly convoluted (3.53) simplifies substantially when $\mathbb{1} \in \mathcal{R}(M_*)$:

$$\begin{aligned} \begin{bmatrix} H_* & H_{(12)} \\ 0 & H_{(22)} \end{bmatrix} \begin{bmatrix} M_*^{-L}\mathbb{1} \\ 0 \end{bmatrix} &= \begin{bmatrix} M_*^{-L}\mathbb{1} \\ 0 \end{bmatrix} \\ H_* M_*^{-L}\mathbb{1} &= M_*^{-L}\mathbb{1} \quad , \end{aligned} \tag{3.54}$$

where we have used the fact that $\left[M_*^\perp\right]' \mathbb{1} = 0$ if any only if $\mathbb{1} \in \mathcal{R}(M_*)$. When $\mathbb{1} \in \mathcal{R}(M_*)$, as should be evident from (3.54), the additional linear constraints of row-stochasticity ($G$'s rows summing to 1, i.e., $G$ having a right eigenvector of $\mathbb{1}$ associated with the eigenvalue 1) require that $M_*^{-L}\mathbb{1}$ be a right eigenvector of $H_*$ associated with the eigenvalue of 1. From our discussion of the correspondence of the eigenvalues and eigenvectors of $G$ and $H_*$ (Section 3.1.2), this fact should have already been deduced. The exact number of free parameters involved in $M_*$-separability, $\eta^2 - p(\eta - p) - p$, is illustrated in (3.51), taking into consideration the $p$ additional constraints imposed on $H_*$ when $\mathbb{1} \in \mathcal{R}(M_*)$. When $\mathbb{1} \notin \mathcal{R}(M_*)$, there is an additional linear constraint for each row of the middle matrix on the right hand side of (3.52), meaning the number of free parameters involved in $M_*$-separability must be $\eta^2 - p(\eta - p) - \eta$. The parametrization of $M_*$-separability is complete by recognizing that such free parameters of the middle matrix on the right hand side of (3.51) must lie in a particular cone, which is the nonnegative cone under the linear transformation specified by $\left[M_* \;\; M_*^\perp\right]$.

### 3.2.2 The feasibility set $\mathcal{G}(\mathcal{M})$ under $\mathcal{M}$-separability

In Section 3.2.1, it was shown that matrices exhibiting $M_*$-separability can be characterized as the nonnegative matrices satisfying the linear constraints of (3.48). Under $\mathcal{M}$-separability, a set of linear constraints will be imposed by nature of being $M_*$-separable for each $M_* \in \mathcal{M}$. All such linear constraints satisfied by a matrix $G$ exhibiting $\mathcal{M}$-separability can be illustrated as follows:

$$\begin{bmatrix} \mathbb{1}' \otimes I \\ M(1)' \otimes \left[M(1)^{\perp}\right]' \\ M(2)' \otimes \left[M(2)^{\perp}\right]' \\ \vdots \\ M(\overline{r})' \otimes \left[M(\overline{r})^{\perp}\right]' \end{bmatrix} \text{vec}(G) = \begin{bmatrix} \mathbb{1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad , \tag{3.55}$$

where

$$\mathcal{M} = \{M(1),\ M(2),\ \ldots,\ M(\overline{r})\} \quad . \tag{3.56}$$

The dimension of the affine subspace of matrices $G$ solving (3.55) equals the dimension of the left hand side matrix's null space, provided that the constraints are consistent. Note that this left hand side matrix of (3.55) is a block Kronecker matrix. In general, the rank of a block Kronecker matrix will not have a simple analytical solution. An exception is when the *common orthonormal bases assumption* is satisfied (discussed in Appendix A.2.1) for matrices $\mathbb{1}$, $M(1)$, $M(2)$, ..., $M(\overline{r})$, meaning that there is single orthonormal basis in $\mathbb{R}^{\eta}$ whereby the range of $\mathbb{1}$ and each $M(r)$ can be expressed as the span of a subset of the basis vectors. As discussed in Appendix A, this notion is the linear algebraic equivalent to the running intersection property of junction trees in Markov random fields [41], as well as sufficient conditions for the extendability of measures [42, 43, 44]. Obviously this necessitates that $\mathbb{1} \in \mathcal{R}(M(r))$ for all $r$. A special case of this is when

$$\mathbb{1} \subset \mathcal{R}(M(1)) \subset \mathcal{R}(M(2)) \subset \ldots \subset \mathcal{R}(M(\overline{r})) \quad , \tag{3.57}$$

which the reader should recognize as regular separability (Section 2.2.4). Corollary 19 offers an analytical expression for the dimension of the affine subspace of matrices $G$ satisfying (3.55) when regular separability holds,

$$\eta^2 - \left(\eta + (p_{M(1)} - 1)(\eta - p_{M(1)}) + \sum_{r=2}^{\overline{r}} (p_{M(r)} - p_{M(r-1)})(\eta - p_{M(r)})\right) \quad , \tag{3.58}$$

where $p_{M(r)}$ is the rank of $M(r)$. If there exists a strictly positive matrix exhibiting such an instance of $\mathcal{M}$-separability, then (3.58) also provides the dimension of $\mathcal{G}(\mathcal{M})$, as stated by the following theorem.

**Theorem 3.** *Consider an instance of regular separability, with*

$$\mathcal{M} = \{M(1),\ M(2),\ \ldots,\ M(\overline{r})\} \quad , \tag{3.59}$$

*and for each $r$, $M(r)$ has rank $p_{M(r)}$.*

*Provided that there exists a strictly positive transition matrix exhibiting $\mathcal{M}$-separability, then the dimension of $\mathcal{G}(\mathcal{M})$, the set of row-stochastic matrices exhibiting $\mathcal{M}$-separability,* *will be*

$$\begin{aligned}\dim\left(\mathcal{G}(\mathcal{M})\right) &= \eta^2 - \eta - \sum_{r=1}^{\overline{r}} (p_{M(r)} - p_{M(r-1)})(\eta - p_{M(r)}) \\ &= \dim(\mathcal{G}) - \sum_{r=1}^{\overline{r}} (p_{M(r)} - p_{M(r-1)})(\eta - p_{M(r)}) \\ &= \dim(\mathcal{G}(M(\overline{r})) - \sum_{r=1}^{\overline{r}} (p_{M(r)} - p_{M(r-1)})(p_{M(\overline{r})} - p_{M(r)}) \quad , \end{aligned} \tag{3.60}$$

*where we have defined* $p_{M(0)} = 1$.

Although there are other cases for which one can derive relatively simple analytical forms for the dimension of $\mathcal{G}(\mathcal{M})$, we do not pursue these other cases. Our canonical examples of $\mathcal{M}$-separability are covered by Theorem 3.

### Visualizing the free parameters of regular separability

The number of free parameters under regular separability can be visualized via a linear transformation on the matrix $G$. To show this, we define a new orthonormal matrix $\widehat{M}(r)$ for each $M(r) \in \mathcal{M}$. Let $\widehat{M}(1)$ be an orthonormal matrix with the same range space as $M(1)$, with its leading column being $\mathbb{1}$. We represent $\widehat{M}(1)$ as

$$\widehat{M}(1) = [\mathbb{1} \;\; M_{1\perp 0}] \quad , \tag{3.61}$$

where the columns of $M_{1\perp 0}$ serve as an orthonormal basis for $\mathcal{R}(M(1)) \cap \mathbb{1}^{\perp}$.

Then for each $1 < r \leq \overline{r}$, let

$$\widehat{M}(r) = \left[\widehat{M}(r-1) \;\; M_{r\perp r-1}\right] \tag{3.62}$$

where $M_{r\perp r-1}$ is a matrix with orthonormal columns spanning

$$\mathcal{R}(M(r)) \cap \mathcal{R}^{\perp}(M(r-1)) \quad . \tag{3.63}$$

Note that this process, an example of Gram-Schmidt orthogonalization, is constructing the orthonormal basis guaranteed by the common orthonormal basis assumption. By nature of our construction and the common orthonormal basis assumption, it follows that

$$\mathcal{R}(\widehat{M}(r)) = \mathcal{R}(M(r)) \quad . \tag{3.64}$$

By (3.64) and Corollary 3, $\widehat{M}(r)$-separability must be equivalent to $M(r)$-separability.

Consider the linear transformation of $G$ under the orthonormal matrix $\left[\widehat{M}(\overline{r}) \;\; \widehat{M}(\overline{r})^{\perp}\right]$:

$$G = \left[\widehat{M}(\overline{r}) \;\; \widehat{M}(\overline{r})^{\perp}\right] \begin{bmatrix} \widehat{H}(\overline{r}) & * \\ 0 & * \end{bmatrix} \left[\widehat{M}(\overline{r}) \;\; \widehat{M}(\overline{r})^{\perp}\right]' \quad , \tag{3.65}$$

where the $(2,1)$ block in the middle matrix on the right hand side of (3.65) is 0 by the assumption of $M(\overline{r})$-separability, and $\widehat{H}(\overline{r})$ is the matrix that propagates the partial information $\left[\widehat{\boldsymbol{\pi}}^{(r)}\right]' \triangleq \boldsymbol{\pi}_{\mathbf{x}}'\widehat{M}(\overline{r})$.

Because of the way that $\widehat{M}(\bar{r})$ is constructed (3.62), the matrix specifying the linear transformation of (3.65) can be equivalently expressed as

$$\begin{aligned}\left[\widehat{M}(\bar{r})\ \widehat{M}(\bar{r})^{\perp}\right] &= \left[\widehat{M}(\bar{r}-1)\ M_{\bar{r}\perp\bar{r}-1}\ \widehat{M}(\bar{r})^{\perp}\right] \\ &= \left[\widehat{M}(\bar{r}-1)\ \widehat{M}(\bar{r}-1)^{\perp}\right] \quad ,\end{aligned} \tag{3.66}$$

and by induction,

$$\begin{aligned}\left[\widehat{M}(\bar{r})\ \widehat{M}(\bar{r})^{\perp}\right] &= \left[\widehat{M}(\bar{r}-1)\ \widehat{M}(\bar{r}-1)^{\perp}\right] \\ &= \left[\widehat{M}(\bar{r}-2)\ \widehat{M}(\bar{r}-2)^{\perp}\right] \\ &\ \ \vdots \\ &= \left[\widehat{M}(1)\ \widehat{M}(1)^{\perp}\right] \\ &= \left[\mathbb{1}\ \mathbb{1}^{\perp}\right] \quad .\end{aligned} \tag{3.67}$$

What this means is that the middle matrix on the right hand side of (3.65) can illustrate all constraints of (3.55) simultaneously, that is, all linear constraints associated with $\mathcal{M}$-separability:

$$\begin{aligned}&\left[\mathbb{1}\ M_{1\perp 0}\ M_{2\perp 1}\ \ldots\ M_{\bar{r}\perp\bar{r}-1}\ \widehat{M}(\bar{r})^{\perp}\right]' G \left[\mathbb{1}\ M_{1\perp 0}\ M_{2\perp 1}\ \ldots\ M_{\bar{r}\perp\bar{r}-1}\ \widehat{M}(\bar{r})^{\perp}\right] \\ &= \left[\begin{array}{ccccccc} 1 & \widehat{H}_{0,1} & \widehat{H}_{0,2} & \ldots & \widehat{H}_{0,\bar{r}-1} & \widehat{H}_{0,\bar{r}} & * \\ 0 & \widehat{H}_{1,1} & \widehat{H}_{1,2} & \ldots & \widehat{H}_{1,\bar{r}-1} & \widehat{H}_{1,\bar{r}} & * \\ 0 & 0 & \widehat{H}_{2,2} & \ldots & \widehat{H}_{2,\bar{r}-1} & \widehat{H}_{2,\bar{r}} & * \\ 0 & 0 & 0 & \ldots & \widehat{H}_{3,\bar{r}-1} & \widehat{H}_{3,\bar{r}} & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & \widehat{H}_{\bar{r},\bar{r}} & * \\ & & & & & & \\ 0 & 0 & 0 & \ldots & 0 & 0 & * \\ \underbrace{\phantom{00}}_{1} & \underbrace{\phantom{00}}_{p_{M(1)}-1} & \underbrace{\phantom{00}}_{\substack{p_{M(2)} \\ -p_{M(1)}}} & & \underbrace{\phantom{00}}_{\substack{p_{M(\bar{r}-1)} \\ -p_{M(\bar{r}-2)}}} & \underbrace{\phantom{00}}_{\substack{p_{M(\bar{r})} \\ -p_{M(\bar{r}-1)}}} & \underbrace{\phantom{00}}_{\eta-p_{M(\bar{r})}} \end{array}\right] \quad ,\end{aligned} \tag{3.68}$$

where

$$\widehat{H}(r) = \left[\begin{array}{cccc} 1 & \widehat{H}_{0,1} & \ldots & \widehat{H}_{0,r} \\ 0 & \widehat{H}_{1,1} & \ldots & \widehat{H}_{1,r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \widehat{H}_{r,r} \end{array}\right] \tag{3.69}$$

is the matrix that propagates the partial information $\widehat{\pi}^{(r)}$. The numbers adjacent to the brackets along the bottom of the matrix in (3.68) indicate the number of columns of a given block. If one counts the number of free parameters in the matrix represented in (3.68), one finds that this is equal to the dimension of $\mathcal{G}(\mathcal{M})$. There are different approaches to counting the fixed parameters (and thereby the free parameters), and these different approaches correspond to the different expressions for the dimension of the set of matrices exhibiting $\mathcal{M}$-separability as given in Theorem 3. For example, if one counts the entries of the matrix that are fixed as 0 or 1 (in fact all are fixed as 0 except for the (1,1) entry) column by column, one will obtain the expression in the first line of (3.60). Alternatively, if one counts first the fixed parameters associated with assuming $M(\bar{r})$-separability, which includes the $(\eta - p_{M(\bar{r})}) \times p_{M(\bar{r})}$ block of zeros as well as the entries of the first column of (3.68), and then counts the remaining entries that must be 0, column by column, we would arrive at the expression in the third line of (3.60). It is important to be reminded that the free parameters are constrained to be in a cone, which is the nonnegative orthant under the given rotation.

As noted in Section 2.2.4, an MLSS of degree $\bar{r}$ that is derived from a finite state Markov chain is an example of regular separability. One can show that the propagation matrix of an MLSS of degree $\bar{r}$ will be upper block triangular, much like the propagation matrix of the right hand side of (3.68).[5] However, in the case of MLSS, the number of 0s will not correspond to actual number of free parameters lost under the assumption of an MLSS of degree $\bar{r}$, because of the redundancies of the representation. Effectively, the partial information that is propagated in an MLSS is obtained via a matrix lacking full column rank.

## 3.3 Illustrating the restricted dynamics of $M_*$-separability

A transition matrix $G$ exhibiting $M_*$-separability has restricted dynamics, as evident from the constraints that must be satisfied in (3.2). Such constraints provide a mathematical characterization of $M_*$-separability, but fail to offer visual or intuitive cues. In Section 3.1, it was shown that $M_*$-separability requires that particular subspaces be invariant (either by left or right multiplication), but again, no visual intuition is offered. This section hopes to offer such visual intuition into $M_*$-separability. A model with two automata ($n = 2$), each having two possible statuses ($m = 2$) is assumed, meaning that $\eta = m^2 = 4$. We consider a particular $4 \times 4$ transition matrix for this set-up that exhibits $M^{(1)}$-separability. We then illustrate the restrictions on the dynamics as a consequence of $M^{(1)}$-separability.

A discussion regarding the dynamics must begin with the joint transition matrix $G$, which is given in this example as

$$G = \begin{bmatrix} \frac{2}{5} & \frac{1}{4} & \frac{1}{5} & \frac{3}{20} \\ \frac{1}{10} & \frac{1}{2} & \frac{3}{20} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{5} & \frac{9}{20} & \frac{1}{10} \\ \frac{3}{20} & \frac{1}{4} & \frac{1}{5} & \frac{2}{5} \end{bmatrix} . \tag{3.70}$$

The matrix $G$ is a representation of the time-homogeneous function $\mathbb{E}\left[\, \mathsf{s}_{\mathsf{x}}[t+1]' \,|\, \mathsf{s}_{\mathsf{x}}[t] \,\right]$ (recall (2.5) for $t - \tau = 1$), or abstractly, $G : \mathcal{S} \to \Delta$,[6] where the domain $\mathcal{S} \subset \mathbb{R}^4$ is the set

[5] One may note that in an MLSS, propagation of the moments, i.e., partial information, is accomplished via right multiplication of the propagation matrix, meaning that by switching from left multiplication to right multiplication, such propagation matrices should become lower block triangular. However, this switch is reversed in an MLSS, because its partial information vector (or state vector prior to taking expectations) is in reverse order compared to the ordering induced in our example by the process of constructing $\widehat{M}^{(r)}$ in (3.61) and (3.62).

[6] Using the same notation for the function and the matrix that represents the function is sloppy. At the same time, representing the matrix and the function with different notation, we feel, would be confusing.

of possible values for the state indicator vector, i.e.,

$$\mathcal{S} = \{\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \boldsymbol{e}_3,\ \boldsymbol{e}_4\} \quad , \tag{3.71}$$

and $\Delta \subset \mathbb{R}^4$ is the probability simplex, i.e., the convex hull of $\mathcal{S}$, the set of probability vectors of length 4. Fig. 3-3 illustrates such a mapping using arrows, where both the domain and range of the function are superimposed. Each arrow originates at a standard unit vector in the domain $\mathcal{S}$ and points to the element in the range $\Delta$ to which it is associated by the function. Note that the probability simplex for an alphabet of size 4 can be visualized in 3-dimensions as a regular tetrahedron, as all probability vectors satisfy the linear constraint that the entries sum to 1.

Typically we think of the domain of the function $G$ as being extended to the entire probability simplex $\Delta$, which is accomplished by taking an expectation on both sides of (2.5), where we obtain the new function

$$\boldsymbol{\pi}_{\mathbf{x}}[t+1]' = \boldsymbol{\pi}_{\mathbf{x}}[t]'G \quad , \tag{3.72}$$

defined for any initial condition $\boldsymbol{\pi}_{\mathbf{x}}[t]$, i.e., any element of $\Delta$. Iterated expectation extends $G$ as a linear function that is defined over the extended domain $\Delta$, the convex hull of the original domain, $\mathcal{S}$. Such an extension is well defined because each probability vector in $\Delta$ has a unique convex expansion in terms of the extreme points of the probability simplex, the standard unit vectors.

In order to visualize $M^{(1)}$-separability, the dynamics specified by $G$ must be considered under a transformation by $M^{(1)}$. Define

$$\mathcal{S}^{(1)} = \left\{\boldsymbol{e}'M^{(1)} : \boldsymbol{e} \in \mathcal{S}\right\} \quad , \tag{3.73}$$

and

$$\Delta^{(1)} = \left\{\boldsymbol{\pi}_{\mathbf{x}}'M^{(1)} : \boldsymbol{\pi}_{\mathbf{x}} \in \Delta\right\} \quad . \tag{3.74}$$

Both $\mathcal{S}^{(1)}$ and $\Delta^{(1)}$ are subsets of $\mathbb{R}^4$ because $M^{(1)}$ has $\overline{p} = 4$ columns. However, as the rank of $M^{(1)}$ is $p = 3 < 4$ and $\mathbb{1} \in M^{(1)}$, the dimension of both sets is $p - 1 = 2$. An additional dimension beyond the rank of $M^{(1)}$ is lost when $\mathbb{1} \in M^{(1)}$ because the linear constraint that a probability vector's entries sum to 1 is transformed into an additional linear constraint on the transformed spaces $\mathcal{S}^{(1)}$ and $\Delta^{(1)}$.

As each row of $M^{(1)}$ is unique (or equivalently, $\mathbf{s}_{\mathbf{x}}'M^{(1)}$ is a state vector for the network), we can define the function $H^{(1)} : \mathcal{S}^{(1)} \to \Delta^{(1)}$, which is the transformed version of our $G$ function, defined as follows:

$$H^{(1)}(\boldsymbol{e}^{(1)}) = \left[M^{-1}(\boldsymbol{e}^{(1)})\right]' GM^{(1)} \qquad \forall\, \boldsymbol{e}^{(1)} \in \mathcal{S}^{(1)} \tag{3.75}$$

where $M^{-1}(\cdot)$ is the inverse mapping from $\mathcal{S}^{(1)}$ to $\mathcal{S}$ (well defined as it only defined on extreme points). This function $H^{(1)}$ is illustrated in 2-dimensions in Fig. 3-4, once again using arrows with the domain and range superimposed. In the projection shown, the first and third coordinates of the vectors serve as the $x$ and $y$ Cartesian coordinates.

$M^{(1)}$-separability holds if and only if $H^{(1)}$ can be extended as a linear function over $\Delta^{(1)}$. The fact that $H^{(1)}$ cannot necessarily be extended as a linear function over $\Delta^{(1)}$ should be clear. As $\Delta^{(1)}$ is a convex set in $p-1 = 2$ dimensions with $\eta = 4$ extreme points, an element

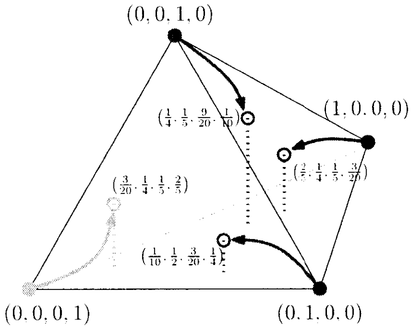

(0, 0, 1, 0)
(1, 0, 0, 0)
$(\frac{1}{4}, \frac{1}{5}, \frac{9}{20}, \frac{1}{10})$
$(\frac{2}{5}, \frac{1}{4}, \frac{1}{5}, \frac{3}{20})$
$(\frac{3}{20}, \frac{1}{4}, \frac{1}{5}, \frac{2}{5})$
$(\frac{1}{10}, \frac{1}{2}, \frac{3}{20}, \frac{1}{4})$
(0, 0, 0, 1)
(0, 1, 0, 0)

Figure 3-3: $G : \mathcal{S} \to \Delta$.

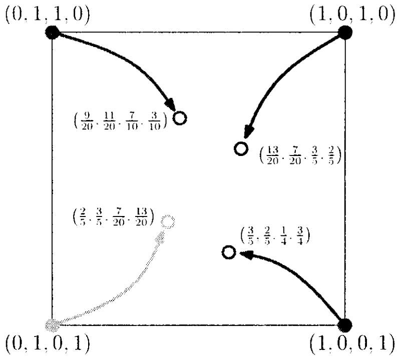

(0, 1, 1, 0)
(1, 0, 1, 0)
$(\frac{9}{20}, \frac{11}{20}, \frac{7}{10}, \frac{3}{10})$
$(\frac{13}{20}, \frac{7}{20}, \frac{3}{5}, \frac{2}{5})$
$(\frac{2}{5}, \frac{3}{5}, \frac{7}{20}, \frac{13}{20})$
$(\frac{3}{5}, \frac{2}{5}, \frac{1}{4}, \frac{3}{4})$
(0, 1, 0, 1)
(1, 0, 0, 1)

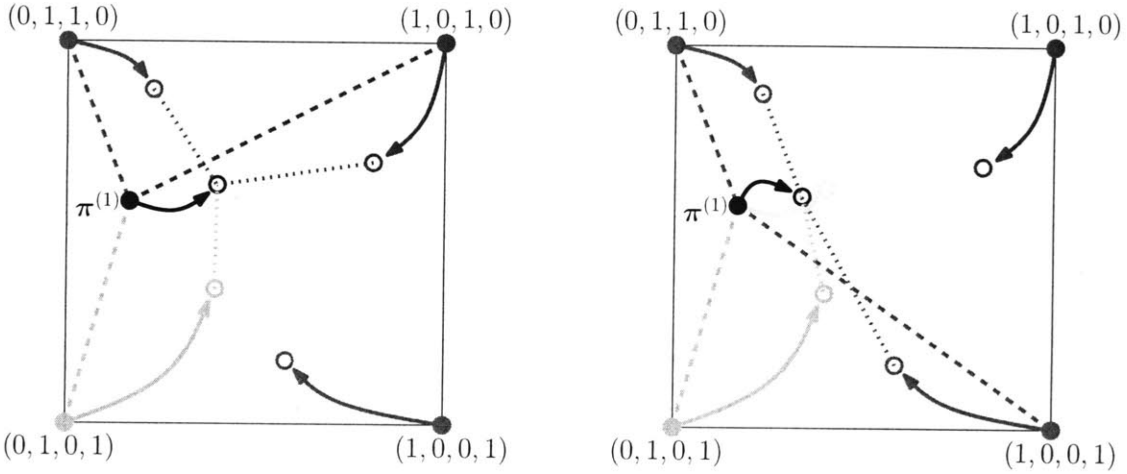
Figure 3-4: $H^{(1)} : \mathcal{S}^{(1)} \to \Delta^{(1)}$.

$\boldsymbol{\pi}^{(1)} \in \Delta^{(1)}$ will not necessarily have a unique representation as a convex combination of the extreme points $\mathcal{S}^{(1)}$. By Carathéodory's theorem, any element $\boldsymbol{\pi}^{(1)}$ in a convex set in $p - 1 = 2$ dimensions can be expressed as a convex sum of $p = 3$ extreme points [45]. For many points in the convex set, their convex expansion in terms of extreme points will not be unique. This may result in the linear extension of $H^{(1)}$ over $\Delta^{(1)}$ to be ill-defined; $M^{(1)}$-separability is the property that assures that such an extension is well-defined for all elements $\boldsymbol{\pi}^{(1)} \in \Delta^{(1)}$. Fig. 3-5 illustrates this problem of extending $H^{(1)}$ over $\Delta^{(1)}$ for an underlying transition matrix $G$ that is *not* $M^{(1)}$-separable.

There are several visual ways to verify whether or not $H^{(1)}$ can be extended as a linear function over $\Delta^{(1)}$. When $n = 2$ and $m = 2$ as in our example, one must verify that the four points comprising the image of $H^{(1)}$ form a parallelogram, with neighboring vertices being neighboring extreme points in their preimage. This notion extends to general $M_*$-separability: when the image of $\mathcal{S}_*$ under $H^{(1)}$ is a linear transformation of $\mathcal{S}_*$, $M_*$-separability is satisfied. A specific example of this illustrated in Fig. 3-4, as the vertices of the parallelogram are a linear transformation of the vertices of the square.

Alternatively, as $H^{(1)}$ is a vector-valued function, one can visualize in 3-dimensions the individual mappings for each coordinate of the function. $H^{(1)}$ can be extended as a linear function if and only if the four points plotted lie within a common plane for each one of these coordinate functions, meaning that there is a matrix representation of $H^{(1)} : \mathcal{S}^{(1)} \to \Delta^{(1)}$. Fig. 3-6 illustrates this fact for $G$ given in (3.70).

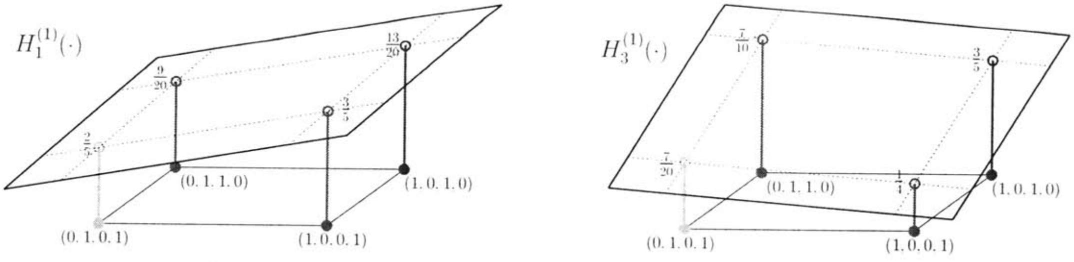
For general $M_*$-separability, $\Delta_*$ will be a $(p - 1)$-dimensional convex set whenever $\mathbb{1} \in M_*$. When the $j$th coordinate function of $H_*$ is defined at $p$ elements of $\mathcal{S}_*$, these $p$ plotted points representing the $j$th coordinate function will coincide in a plane. Provided that the $p$ plotted points of the function only satisfy a single linear constraint, the plane will be unique. When such a plane is unique and the underlying transition matrix $G$ is $M_*$-separable, the values of the $j$th coordinate function at the remaining $\eta - p$ elements of $\mathcal{S}_*$ will be predetermined by the plane, should $M_*$-separability be satisfied. This realization suggests that the constraints imposed by $M_*$-separability are captured by $p$, the number of linearly independent coordinate functions of $H_*$, multiplied by $\eta - p$, effectively the number of values of each coordinate function of $H_*$ that are constrained; this should be intuitively satisfying as it concurs with the result of Theorem 2.



(a) Defining $H^{(1)}(\pi^{(1)})$ by taking the appropriate convex combination of $H^{(1)}((0,1,0,1))$, $H^{(1)}((0,1,1,0))$, and $H^{(1)}((1,0,1,0))$.

(b) Defining $H^{(1)}(\pi^{(1)})$ by taking the appropriate convex combination of $H^{(1)}((0,1,0,1))$, $H^{(1)}((0,1,1,0))$, and $H^{(1)}((1,0,0,1))$ (inconsistent with the extension pictured in part (a) ).

Figure 3-5: Two ways of attempting a linear extension of $H^{(1)}$ at $\pi^{(1)}$. As illustrated in (a) and (b), there are two different ways to express the partial information vector $\pi^{(1)}$ as a convex combination of 3 extreme points in $\mathcal{S}^{(1)}$. $H^{(1)}(\pi^{(1)})$ is defined in each case by taking the appropriate convex combination of $H^{(1)}$ evaluated at the three extreme points, i.e., a linear extension. Because the underlying $G$ matrix does not exhibit $M^{(1)}$-separability, the two ways are inconsistent.



(a) $H_1^{(1)} : \mathcal{S}^{(1)} \to [0,1]$

(b) $H_3^{(1)} : \mathcal{S}^{(1)} \to [0,1]$.

Figure 3-6: Individual plots for the coordinates of the function $H^{(1)}(\cdot) = \left(H_1^{(1)}(\cdot),\ H_2^{(1)}(\cdot),\ H_3^{(1)}(\cdot),\ H_4^{(1)}(\cdot)\right)$.

## 3.4 The dimensions of $M^{(r)}$-separability and $\mathcal{M}^{(\overline{r})}$-separability

We now apply the results of Theorems 2 and 3 to the canonical examples of $M^{(r)}$-separability and $\mathcal{M}^{(\overline{r})}$-separability. As clear from the theorems, the tasks that remain involve determining the rank of $M^{(r)}$ and whether or not $\mathbb{1} \in \mathcal{R}(M^{(r)})$. Because $M^{(r)}$ can be expressed as a block Kronecker matrix, the approach explained in Section A.2 can be followed to derive its rank.

The relatively simple case of determining the rank of $M^{(1)}$ is first considered, and an equation for the dimension of $\mathcal{G}(M^{(1)})$ is derived as a function of $n$ and $\boldsymbol{m}$, i.e., the number of automata and each automaton's number of possible statuses. We compute the dimension of $M^{(1)}$-separable probabilistic models is compared to otherwise unconstrained Markovian models. Leveraging the same techniques used to determine the rank of $M^{(1)}$, we develop a general equation for the rank of $M^{(r)}$, and thus, the dimensions of $\mathcal{G}(M^{(r)})$ and $\mathcal{G}(\mathcal{M}^{(\overline{r})})$, as a function of $n$, $\boldsymbol{m}$ and $r$. At times, our discussion will involve some long and tedious derivations, although we feel there is value in following the derivations closely to be exposed to several interesting techniques. Those with less patience are encouraged to skip to Section 3.5 to see how $\mathcal{G}(M^{(r)})$ and $\mathcal{G}(\mathcal{M}^{(\overline{r})})$ vary in regards to $n$, $\boldsymbol{m}$, and $r$.

Intuitively, one can argue that the rank of $M^{(r)}$ should be roughly the number of independent pieces of information that the partial information vector $\mathbf{s}^{(r)} = \mathbf{s}_{\mathbf{x}}' M^{(r)}$ provides (we use the terms 'independent' and 'information' in a loosely defined sense). Recall that the partial information vector $\mathbf{s}^{(r)}$ provides the marginal PMFs for the statuses of any $r$-tuple of automata. There are $\binom{n}{r}$ subsets consisting of $r$ unique automata, and all information contained in $\mathbf{s}^{(r)}$ can be captured by these $\binom{n}{r}$ $r$th-order marginal PMFs. Yet there is overlapping information provided by any two $r$th-order marginal PMFs that have associated automata in common: both provide the marginal PMFs for any automata in common. To 'orthogonalize' this information into independent pieces, consider the following: $\mathbf{s}^{(r)}$ provides the univariate marginal PMFs for each automata. Assuming that all automata have the same number of possible statuses, then the univariate marginal PMFs provide $\binom{n}{1}(m-1)$ pieces of information—there are $\binom{n}{1}$ such marginal PMFs, and each provides $m-1$ pieces of information (as probabilities must sum to 1). Next, consider the bivariate marginal PMFs. There are $\binom{n}{2}$ bivariate marginal PMFs of interest, and considering that the univariate marginals PMFs are already known, each bivariate marginal PMF provides $(m-1)^2$ bits of information. This gives us a subtotal of $\binom{n}{1}(m-1) + \binom{n}{2}(m-1)^2$ pieces of information. By continuing this process, we would find that the partial information vector $\mathbf{s}^{(r)}$ contains

$$\sum_{r'=1}^{r} \binom{n}{r'} (m-1)^{r'} \tag{3.76}$$

bits of information, and consequently (3.76) would be the intuitive estimate for the rank of $M^{(r)}$. This estimate seems reasonable, particularly because $\mathbf{s}^{(n)}$ would provide complete information, i.e., $m^n - 1$ pieces of information, and we know by the binomial expansion theorem,

$$m^n - 1 = \sum_{r=1}^{n} \binom{n}{r} (m-1)^r \quad . \tag{3.77}$$

In fact, as proven based on the structure of $M^{(r)}$ in Theorem 5, such an intuitive estimate is off by 1, and only because we are trying to intuitively estimate the rank of $M^{(r)}$ by considering the partial information it provides when left multiplied by a probability vector, which is inherently constrained to sum to 1.

### 3.4.1 $M^{(1)}$-separability

We wish to illustrate the process of determining the dimension of a probabilistic model for the special case of $M^{(1)}$-separability. Recall that $M^{(1)}$ is the matrix that maps the network's state indicator vector to a sequence of status indicator vectors (and under expectations, maps a probability vector for the network state to a sequence of univariate marginal probability vectors). As has already been shown in (2.23), $M^{(1)}$ can be expressed as

$$M^{(1)} = \begin{bmatrix} M_1 \, M_2 \, \dots \, M_n \end{bmatrix} \quad , \tag{3.78}$$

where $M_j$ is the matrix mapping the state indicator vector to the status indicator vector for automaton $j$. As shown in (2.15), the mixed-product property of Kronecker products permits one to represent each $M_j$ as

$$M_j \triangleq \mathbb{1} \otimes \mathbb{1} \otimes \dots \otimes \underbrace{I}_{j\text{th term}} \otimes \dots \otimes \mathbb{1} \quad . \tag{3.79}$$

By (3.78) and (3.79), clearly $M^{(1)}$ is a block Kronecker matrix. We shall determine the rank of $M^{(1)}$ by following Appendix A.2 as a guide for determining the rank of a block Kronecker matrix. However, before proceeding to do so, we shall show that $\mathbb{1} \in \mathcal{R}(M^{(1)})$. Consider $M_j$, which is a subset of the columns of $M^{(1)}$, for any value of $j$ (3.78). Because each row of $M_j$ is a possible value (transposed) of the indicator vector for automaton $j$, each row must consist of all 0s except for a single 1. Consequently,

$$M_j \mathbb{1} = \mathbb{1} \quad , \tag{3.80}$$

and likewise, $\mathbb{1} \in \mathcal{R}(M^{(1)})$. This fact can also be shown algebraically. From (3.79), one can surmise that the $k$th column of $M_j$ is

$$\mathbb{1} \otimes \mathbb{1} \otimes \dots \otimes \underbrace{e_k}_{j\text{th term}} \otimes \dots \otimes \mathbb{1} \quad , \tag{3.81}$$

where $e_k$ is the $k$ standard unit vector of length $m_j$, and equivalently the $k$th column of an $m_j \times m_j$ identity matrix. By the distributive property (A.2), we can write

$$\begin{aligned} M_j \mathbb{1} &= \sum_{k=1}^{m_j} \mathbb{1} \otimes \mathbb{1} \otimes \dots \otimes e_k \otimes \dots \otimes \mathbb{1} \\ &= \mathbb{1} \otimes \mathbb{1} \otimes \dots \otimes \left( \sum_{k=1}^{m_j} e_k \right) \otimes \dots \otimes \mathbb{1} \\ &= \mathbb{1} \quad . \end{aligned} \tag{3.82}$$

We now return to our second objective, which is determining the rank of $M^{(1)}$. Although there are several approaches to determining the rank of $M^{(1)}$ (Appendix B in [12] derives the rank of $M^{(1)}$, denoted there as $B$, by induction), we pursue the approach outlined in Section A.2 for block Kronecker matrices that relies on two tools: the mixed-product property of Kronecker products (A.3) and Theorem 16b. A matrix $\widehat{M}^{(1)}$ will be constructed that shares the same range space as $M^{(1)}$, with all of its columns being either orthogonal (easily checked by the mixed-product property of Kronecker products) or equal. With such structure, the rank of $\widehat{M}^{(1)}$ can be determined by counting its unique columns.

$\widehat{M}^{(1)}$ will be a block Kronecker matrix,

$$\widehat{M}^{(1)} = \left[ \widehat{M}_1 \; \widehat{M}_2 \ldots \widehat{M}_n \right] \quad , \tag{3.83}$$

where

$$\widehat{M}_j \triangleq \mathbb{1} \otimes \mathbb{1} \otimes \ldots \otimes \underbrace{\Gamma^{(m_j)}}_{j\text{th term}} \otimes \ldots \otimes \mathbb{1} \quad , \tag{3.84}$$

and $\Gamma^{(m_j)}$ is an $m_j \times m_j$ orthogonal matrix with its first column being $\mathbb{1}$. For our purposes, we do not explicitly construct such a matrix $\Gamma^{(m_j)}$, only note that such a matrix exists.[7] In comparing the definitions of $M_j$ (3.79) and $\widehat{M}_j$ (3.84), we see that $I$, the $j$th term in the Kronecker product defining $M_j$, has been substituted for $\Gamma^{(m_j)}$, another square, orthogonal matrix that has $\mathbb{1}$ as its first column. As $\mathcal{R}(I) = \mathcal{R}(\Gamma^{(m_j)})$, by Theorem 16b, it follows that $\mathcal{R}(M_j) = \mathcal{R}(\widehat{M}_j)$, and furthermore, by Theorem 17, the columns of each $\widehat{M}_j$ must be orthogonal. Consider two columns in distinct blocks of $\widehat{M}^{(1)}$: column $k_1$ of $M_{j_1}$ and column $k_2$ of $M_{j_2}$, which will be denoted as $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively, for notational simplicity.

Evidently from (3.84),

$$\boldsymbol{x}_1 = \mathbb{1} \otimes \mathbb{1} \otimes \ \ldots \ \otimes \underbrace{\gamma_{k_1}^{(m_{j_1})}}_{j_1\text{th term}} \otimes \ldots \otimes \mathbb{1} \quad , \tag{3.85}$$

$$\boldsymbol{x}_2 = \mathbb{1} \otimes \mathbb{1} \otimes \ldots \otimes \underbrace{\gamma_{k_2}^{(m_{j_2})}}_{j_2\text{th term}} \otimes \ \ldots \ \otimes \mathbb{1} \tag{3.86}$$

where $\gamma_k^{(m_j)}$ is the $k$th column of $\Gamma^{(m_j)}$. If $k_1 = k_2 = 1$, both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are a column of all 1s. Otherwise, either $k_1 \neq 1$ or $k_2 \neq 1$, and without loss of generality we will assume that $k_1 \neq 1$. Provided that $k_1 \neq 1$, it will be assured that $\gamma_{k_1}' \mathbb{1} = 0$, and thus,

$$\begin{aligned}
\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle &= \boldsymbol{x}_1' \boldsymbol{x}_2 \\
&= \Big( \mathbb{1}' \otimes \mathbb{1}' \otimes \ \ldots \ \otimes \underbrace{\gamma_{k_1}^{(m_{j_1})'}}_{j_1\text{th term}} \otimes \ldots \otimes \mathbb{1}' \Big) \\
&\quad \cdot \Big( \mathbb{1} \otimes \mathbb{1} \otimes \ldots \otimes \underbrace{\gamma_{k_2}^{(m_{j_2})}}_{j_2\text{th term}} \otimes \ \ldots \ \otimes \mathbb{1} \Big) \\
&= (\mathbb{1}'\mathbb{1}) \otimes \ldots \otimes \underbrace{\big( \gamma_{k_1}^{(m_{j_1})'} \mathbb{1} \big)}_{j_1\text{th term}} \otimes \ldots \otimes \underbrace{\big( \gamma_{k_2}^{(m_{j_2})'} \mathbb{1} \big)}_{j_2\text{th term}} \otimes \ldots \otimes (\mathbb{1}'\mathbb{1}) \\
&= 0 \quad ,
\end{aligned} \tag{3.87}$$

[7] Should one wish to explicitly construct $\Gamma^{(m_j)}$, it is fairly straightforward to do. Naturally the first column is $\mathbb{1}$. The remaining columns must span $\mathcal{R}^{\perp}(\mathbb{1})$, the orthogonal complement to the all 1s vector. Determining a basis for $\mathcal{R}^{\perp}(\mathbb{1})$ is straightforward: consider the vectors $(1, -1, 0, \ldots 0)$, $(0, 1, -1, 0, \ldots 0)$, $\ldots$, $(0, \ldots, 0, 1, -1)$. Clearly such vectors are linearly independent and as all are orthogonal to $\mathbb{1}$, such vectors must form a basis for $\mathcal{R}^{\perp}(\mathbb{1})$. Although this is not an orthonormal basis, such a basis can be easily converted into an orthonormal basis via Gram-Schmidt.

Table 3.1: Dimension of $\mathcal{G}(M^{(1)})$ for varying $n$ ($m = 2$).

| | $\dim\left(\mathcal{G}(M^{(1)})\right)$ | $\dim(\mathcal{G})$ | $\frac{\dim\left(\mathcal{G}(M^{(1)})\right)}{\dim(\mathcal{G})}$ |
|---|---|---|---|
| $n = 2$ | 10 | 12 | 0.833 |
| $n = 3$ | 44 | 56 | 0.787 |
| $n = 4$ | 196 | 240 | 0.817 |
| $n = 5$ | 862 | 992 | 0.869 |
| $n = 10$ | 1037422 | 1047552 | .990 |
| $n = 100$ | $1.607 \times 10^{60}$ | $1.607 \times 10^{60}$ | 1.000 |

where the third equality is by the ubiquitous mixed-product property of Kronecker products (A.3). We conclude that the first column of each $M_j$ is the all 1s column, and otherwise, the columns of $\widehat{M}^{(1)}$ are orthogonal. As $\widehat{M}^{(1)}$ has a total of $\sum_{j=1}^{n} m_j$ columns,

$$\begin{aligned}\operatorname{rank}(M^{(1)}) &= \operatorname{rank}(\widehat{M}^{(1)}) \\ &= \sum_{j=1}^{n} m_j - n + 1 \quad .\end{aligned} \tag{3.88}$$

We now can conclude using (3.88) and Theorem 2 that

$$\begin{aligned}\dim\left(\mathcal{G}(M^{(1)})\right) &= \prod_j m_j \left(\prod_j m_j - 1\right) \\ &\quad - \left(\prod_j m_j - \sum_j m_j + n - 1\right)\left(\sum_j m_j - n\right) \quad .\end{aligned} \tag{3.89}$$

In the case that $m = m_j$ for all $j$, the expression for the dimension of $\mathcal{G}(M^{(1)})$ can be simplified to

$$\dim\left(\mathcal{G}(M^{(1)})\right) = m^n\left(m^n - 1\right) - \left(m^n - n(m-1) - 1\right)\left(n(m-1)\right) \quad . \tag{3.90}$$

We can see how the dimension of $\mathcal{G}(M^{(1)})$ evolves in relation to the dimension of the otherwise unconstrained set of row-stochastic transition matrices, denoted as $\mathcal{G}$. This relative growth is highlighted in Table 3.1 and Table 3.2 for the $m = 2$ and $m = 3$ cases, respectively. The reader may note some interesting asymptotic behavior, which will be discussed further in Section 3.5.

### 3.4.2 $M^{(r)}$-separability: extending $M^{(1)}$-separability

The techniques used to determine the rank of of $M^{(1)}$ can be generalized to determine the rank of $M^{(r)}$, for any value of $r < n$. To do so, we first need to show how $M^{(r)}$ can be represented. Recall that $M^{(r)}$ maps the network's state indicator vector to a sequence of $n^r$ $r$th-order joint status indicator vectors. Each joint status indicator vector $\mathbf{s}_j$ corresponds to

Table 3.2: Dimension of $\mathcal{G}(M^{(1)})$ for varying $n$ ($m = 3$).

| | $\dim\left(\mathcal{G}(M^{(1)})\right)$ | $\dim(\mathcal{G})$ | $\frac{\dim\left(\mathcal{G}(M^{(1)})\right)}{\dim(\mathcal{G})}$ |
|---|---|---|---|
| $n = 2$ | 56 | 72 | 0.778 |
| $n = 3$ | 582 | 702 | 0.829 |
| $n = 4$ | 5904 | 6480 | 0.911 |
| $n = 5$ | 56486 | 58806 | 0.961 |
| $n = 10$ | $3.486 \times 10^9$ | $3.487 \times 10^9$ | .9997 |
| $n = 100$ | $2.656 \times 10^{95}$ | $2.656 \times 10^{95}$ | 1.000 |

a different sequence of $r$ automata,[8] which is specified by the vector subscript $\boldsymbol{j}$. Naturally under expectations, $M^{(r)}$ maps the probability vector for the network state to the marginal probability vectors for the joint statuses of any $r$ automata.

Like $M^{(1)}$, we can represent $M^{(r)}$ in block form:

$$M^{(r)} = \left[ M_{\boldsymbol{j}^{(1)}} \; M_{\boldsymbol{j}^{(2)}} \; \ldots \; M_{\boldsymbol{j}^{(n^r)}} \right] \quad , \tag{3.91}$$

where each block $M_{\boldsymbol{j}}$ (a marginalizing matrix) maps the state indicator vector to the joint status indicator vector $\mathbf{s}_{\boldsymbol{j}}$. The sequence of vectors $\boldsymbol{j}^{(1)}$, $\boldsymbol{j}^{(2)}$, ..., $\boldsymbol{j}^{(n^r)}$ consists of all length $r$ vectors consisting of integers between 1 and $n$, in lexicographic order. For example, when $r = 3$, $\boldsymbol{j}^{(1)} = (1, 1, 1)$, $\boldsymbol{j}^{(2)} = (1, 1, 2)$, and so on.

Unlike the case when $r = 1$, $M^{(r)}$ for $r > 1$ is not a block Kronecker matrix. Prior to applying our techniques for determining the rank of block Kronecker matrices, some manipulation of the blocks is necessary, made possible by defining some new notation, the binary subset operator ('$\subset$') for vectors. Let the boolean function $\boldsymbol{k} \subset \boldsymbol{l}$ for vectors $\boldsymbol{k} \in \mathbb{R}^r$ and $\boldsymbol{l} \in \mathbb{R}^{\widetilde{r}}$ be defined as follows: for vector $\boldsymbol{k}$, define the unordered set $\mathcal{K}$ consisting of the values assumed by the elements of $\boldsymbol{k}$, and similarly define a set $\mathcal{L}$ for $\boldsymbol{l}$; return $\mathcal{K} \subset \mathcal{L}$. For example, suppose one wants to determine whether or not $(1, 4, 2, 4) \subset (4, 1, 2, 3)$. To answer this, we construct sets composed of the elements in each vector: $\{1, 2, 4\}$ and $\{1, 2, 3, 4\}$, respectively. Clearly, $\{1, 2, 4\} \subset \{1, 2, 3, 4\}$, and consequently it follows that $(1, 4, 2, 4) \subset (4, 1, 2, 3)$. Note that $\boldsymbol{k} \subset \boldsymbol{l}$ is well defined even when the vectors $\boldsymbol{k}$ and $\boldsymbol{l}$ have different lengths.

**Theorem 4.** *If $\boldsymbol{j} \subset \boldsymbol{k}$, then*

$$\mathcal{R}(M_{\boldsymbol{j}}) \subset \mathcal{R}(M_{\boldsymbol{k}}) \quad . \tag{3.92}$$

An immediate corollary is the following:

**Corollary 8.** *For $r \leq \widetilde{r}$,*

$$\mathcal{R}(M^{(r)}) \subset \mathcal{R}(M^{(\widetilde{r})}) \quad . \tag{3.93}$$

The corollary follows by noting that for any block $M_{\boldsymbol{j}}$ of $M^{(r)}$, there must exist a block $M_{\widetilde{\boldsymbol{j}}}$ in $M^{(\widetilde{r})}$ such that $\boldsymbol{j} \subset \widetilde{\boldsymbol{j}}$.

[8] In defining a sequence, the ordering of the elements *is* important and repeated elements are permitted.

*Proof of Theorem 4.* By assumption, we have $\boldsymbol{j} \subset \boldsymbol{k}$. Suppose that $\boldsymbol{j} \in \mathbb{R}^{r}$ and $\boldsymbol{k} \in \mathbb{R}^{\widetilde{r}}$. We can construct a matrix $M_{\boldsymbol{k}\to\boldsymbol{j}}$ that maps the joint status indicator vector for automata $\boldsymbol{k}$ to the joint status indicator vector for automata $\boldsymbol{j}$, i.e.,

$$\mathsf{s}_{\boldsymbol{k}}' M_{\boldsymbol{k}\to\boldsymbol{j}} = \mathsf{s}_{\boldsymbol{j}}' \quad . \tag{3.94}$$

The matrix $M_{\boldsymbol{k}\to\boldsymbol{j}}$ can be defined row by row. Suppose that $\mathsf{s}_{\boldsymbol{k}} = \boldsymbol{e}_l$, the $l$th standard unit vector. From this information that $\mathsf{s}_{\boldsymbol{k}} = \boldsymbol{e}_l$, we can determine the values of $\mathsf{s}_{k_1}$, $\mathsf{s}_{k_2}, \ldots, \mathsf{s}_{k_{\widetilde{r}}}$, where $\boldsymbol{k} = (k_1, k_2, \ldots, k_{\widetilde{r}})$. As $\boldsymbol{j} \subset \boldsymbol{k}$, we also know the values of $\mathsf{s}_{j_1}$, $\mathsf{s}_{j_2}, \ldots, \mathsf{s}_{j_r}$ and consequently the value of $\mathsf{s}_{\boldsymbol{j}}$, which we denote as $\boldsymbol{e}_{f(l)}$. Define the $l$th row of $M_{\boldsymbol{k}\to\boldsymbol{j}}$ as $\boldsymbol{e}_{f(l)}$, ensuring that $\boldsymbol{e}_l' M_{\boldsymbol{k}\to\boldsymbol{j}} = \boldsymbol{e}_{f(l)}'$. Note that for some values of $l$, it is not possible for $\mathsf{s}_{\boldsymbol{k}}$ to assume the value of $\boldsymbol{e}_l$, e.g., for $m_1 = 2$, $\mathsf{s}_1 \otimes \mathsf{s}_1$ cannot equal $\boldsymbol{e}_2$ or $\boldsymbol{e}_3$. When there are no possible combination of values for $\mathsf{s}_{k_1}, \ldots, \mathsf{s}_{k_{\widetilde{r}}}$ such that $\mathsf{s}_{\boldsymbol{k}} = \boldsymbol{e}_l$, define the $l$th row of $M_{\boldsymbol{k}\to\boldsymbol{j}}$ to be all 0s (technically how it is defined is irrelevant). One should now be convinced that when $M_{\boldsymbol{k}\to\boldsymbol{j}}$ is defined in this way for all $l = 1, 2, \ldots, \widetilde{k}$, it follows that $M_{\boldsymbol{k}\to\boldsymbol{j}}$ as constructed satisfies (3.94), and therefore,

$$\begin{aligned} \mathsf{s}_{\mathsf{x}}' M_{\boldsymbol{j}} = \mathsf{s}_{\boldsymbol{j}}' &= \mathsf{s}_{\boldsymbol{k}}' M_{\boldsymbol{k}\to\boldsymbol{j}} \\ &= \mathsf{s}_{\mathsf{x}}' M_{\boldsymbol{k}} M_{\boldsymbol{k}\to\boldsymbol{j}} \quad . \end{aligned} \tag{3.95}$$

As (3.95) holds for all possible values for $\mathsf{s}_{\mathsf{x}}$, which span $\mathbb{R}^{\eta}$, we conclude that $M_{\boldsymbol{j}} = M_{\boldsymbol{k}} M_{\boldsymbol{k}\to\boldsymbol{j}}$, and by linear algebra,

$$\mathcal{R}(M_{\boldsymbol{j}}) \subset \mathcal{R}(M_{\boldsymbol{k}}) \quad . \tag{3.96}$$

□

By lines of reasoning similar to those used to prove Corollary 8, we claim that for any block $M_{\boldsymbol{j}}$ in $M^{(r)}$ (3.91), there exists a vector $\boldsymbol{k}$ that corresponds to a strictly increasing sequence, with $\boldsymbol{j} \subset \boldsymbol{k}$ ($\boldsymbol{j}$ and $\boldsymbol{k}$ need not be distinct). Based on this observation, we define

$$\widetilde{M}^{(r)} \triangleq \begin{bmatrix} M_{\boldsymbol{k}_1} & M_{\boldsymbol{k}_2} & \ldots & M_{\boldsymbol{k}_{\binom{n}{r}}} \end{bmatrix} \tag{3.97}$$

where $\boldsymbol{k}_1$, $\boldsymbol{k}_2, \ldots, \boldsymbol{k}_{\binom{n}{r}}$ is the subsequence of $\boldsymbol{j}_1$, $\boldsymbol{j}_2, \ldots, \boldsymbol{j}_{n^r}$ in (3.91) consisting of all vectors that correspond to strictly increasing sequences. Evidently by our claim and Theorem 4,

$$\mathcal{R}(\widetilde{M}^{(r)}) = \mathcal{R}(M^{(r)}) \quad . \tag{3.98}$$

From this point forward, we mirror the steps taken to determine the rank of $M^{(1)}$, as it can be shown that $\widetilde{M}^{(r)}$ is a block Kronecker matrix. Effectively, this reduction from $M^{(r)}$ to $\widetilde{M}^{(r)}$ is the structural equivalent to the intuitive argument in the introduction to this section that only $\binom{n}{r}$ of the $r$th-order marginal PMFs produced by $M^{(r)}$ under left multiplication by a network state probability vector are necessary.

For each $\boldsymbol{k} = (k_1, k_2, \ldots, k_r)$ corresponding to a strictly increasing sequence, we can represent $M_{\boldsymbol{k}}$ as

$$M_{\boldsymbol{k}} = \mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes \underbrace{I}_{k_1\text{th term}} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes \underbrace{I}_{k_2\text{th term}} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes \underbrace{I}_{k_r\text{th term}} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1} \tag{3.99}$$

as the mixed product property of Kronecker products (A.3) ensures that $\mathsf{s}_{\boldsymbol{k}}' = \mathsf{s}_{\mathsf{x}}' M_{\boldsymbol{k}}$ for $M_{\boldsymbol{k}}$

defined as in (3.99).

For each block $M_{\boldsymbol{k}}$ in (3.97), we define a new matrix

$$\widehat{M_{\boldsymbol{k}}} \triangleq \mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes \underbrace{\Gamma^{(m_{k_1})}}_{k_1\text{th term}} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes \underbrace{\Gamma^{(m_{k_2})}}_{k_2\text{th term}} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes \underbrace{\Gamma^{(m_{k_r})}}_{k_r\text{th term}} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1} \quad , \tag{3.100}$$

where $\Gamma^{(m)}$, as in Section 3.4.1, is an $m \times m$ matrix with orthogonal columns and its first column as the all 1s column, $\mathbb{1}$. By Theorem 16, it follows that

$$\mathcal{R}(\widehat{M_{\boldsymbol{k}}}) = \mathcal{R}(M_{\boldsymbol{k}}) \tag{3.101}$$

for any block $M_{\boldsymbol{k}}$ in (3.97).

If we consider the matrix

$$\widehat{M}^{(r)} \triangleq \begin{bmatrix} \widehat{M}_{\boldsymbol{k}_1} & \widehat{M}_{\boldsymbol{k}_2} & \ldots & \widehat{M}_{\boldsymbol{k}_{\binom{n}{r}}} \end{bmatrix} \quad , \tag{3.102}$$

we can be assured by (3.98) and (3.101) that

$$\mathcal{R}(\widehat{M}^{(r)}) = \mathcal{R}(M^{(r)}) \quad , \tag{3.103}$$

and moreover, by construction, all columns of $\widehat{M}^{(r)}$ are either equal or orthogonal. Although the details are slightly different, the idea as to why pairs of columns must be either equal or orthogonal is the same as illustrated in (3.87) for the special case of $\widehat{M}^{(1)}$.

We can compute the rank of $\widehat{M}^{(r)}$ (and consequently by (3.103), the rank of $M^{(r)}$) by identifying the number of unique columns in $\widehat{M}^{(r)}$. By (3.100), each column of $\widehat{M}^{(r)}$ can be expressed as a sequence of Kronecker products of all 1s vectors interspersed with column vectors from the $\Gamma^{(m)}$ matrices, i.e.,

$$\mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes \underbrace{\gamma_{j_1}^{(m_{d_1})}}_{d_1\text{th term}} \otimes \mathbb{1} \ldots \mathbb{1} \otimes \underbrace{\gamma_{j_r}^{(m_{d_r})}}_{d_r\text{th term}} \otimes \mathbb{1} \ldots \otimes \mathbb{1} \tag{3.104}$$

for some positions $d_1, d_2, \ldots, d_r$ that identify the terms in the sequence of Kronecker products drawn from columns of $\Gamma^{(m_{d_1})}$, $\Gamma^{(m_{d_2})}, \ldots, \Gamma^{(m_{d_r})}$, respectively. Note that whenever the first column of $\Gamma^{(m_{d_i})}$ appears (i.e., $j_i = 1$) in (3.104), then we have an additional all 1s vector appearing in the chain of Kronecker products. By a basic counting argument, we determine

$$\begin{aligned} \operatorname{rank}(M^{(r)}) &= \operatorname{rank}(\widehat{M}^{(r)}) \\ &= \#\text{ of unique columns in } \widehat{M}^{(r)} \\ &= 1 + \sum_{r'=1}^{r} \sum_{1 \le d_1 < \ldots < d_{r'} \le n} \#\text{ ways for } d_1, \ldots,\ d_{r'} \text{ terms to be} \neq \mathbb{1} \\ &= 1 + \sum_{r'=1}^{r} \sum_{d_1=1}^{n-r'+1} \sum_{d_2>d_1}^{n-r'+2} \cdots \sum_{d_{r'}>d_{r'-1}}^{n} \prod_{i=1}^{r'} (m_{d_i} - 1) \quad . \end{aligned} \tag{3.105}$$

where the leading '1+' accounts for the column consisting of all 1s.

The expression in (3.105) can be simplified when $m_j = m$ for all $j$. This is summarized in the following theorem.

**Theorem 5.** *When $m_j = m$ for all $j$ (all automata have the same number of possible statuses), $M^{(r)}$ is a $m^n \times (nm)^r$ matrix such that*

$$\operatorname{rank}\left(M^{(r)}\right) = \sum_{r'=0}^{r} \binom{n}{r'} (m-1)^{r'} \quad , \tag{3.106}$$

*and consequently,*

$$\operatorname{rank}\left(\left[M^{(r)}\right]^{\perp}\right) = \sum_{r'=r+1}^{n} \binom{n}{r'} (m-1)^{r'} \quad . \tag{3.107}$$

The statement regarding the rank of $\left[M^{(r)}\right]^{\perp}$ in (3.107) follows from the binomial expansion theorem.

When there are automata with different numbers of possible statuses, we can conclude

$$\sum_{r'=0}^{r} \binom{n}{r'} (\underline{m}-1)^{r'} \leq \operatorname{rank}\left(M^{(r)}\right) \leq \sum_{r'=0}^{r} \binom{n}{r'} (\overline{m}-1)^{r'} \quad , \tag{3.108}$$

where $\overline{m} \triangleq \max_j m_j$, $\underline{m} \triangleq \min_j m_j$. As $\binom{n}{r} = \Theta(n^r)$ for fixed $r$ (B.6),

$$\operatorname{rank}\left(M^{(r)}\right) = \Theta(n^r) \quad \text{(as a function of } n) \quad , \tag{3.109}$$

whenever the number of possible statuses for each automata is bounded with $\underline{m} > 1$.

As derived in (3.105) and (3.106), we now have a way to compute the rank of $M^{(r)}$. Because $\mathbb{1} \in \mathcal{R}(M^{(1)})$, we know by Corollary 8 that $\mathbb{1} \in \mathcal{R}(M^{(r)})$ for all $r$. Thus we obtain the following corollary for the dimension of $\mathcal{G}(M^{(r)})$ by combining Theorem 2 with Theorem 5.

**Corollary 9.** *When $m_j = m$ for all $j$,*

$$\begin{aligned}
\dim\left(\mathcal{G}(M^{(r)})\right) &= \eta^2 - \operatorname{rank}(M^{(r)}) \cdot (\eta - \operatorname{rank}(M^{(r)})) - \operatorname{rank}(M^{(r)}) \\
&= \eta^2 - \left(\sum_{r'=0}^{r} \binom{n}{r'} (m-1)^{r'}\right) \left(\eta - \sum_{r'=0}^{r} \binom{n}{r'} (m-1)^{r'}\right) \\
&\quad - \sum_{r'=0}^{r} \binom{n}{r'} (m-1)^{r'} \\
&= \dim(\mathcal{G}) - \left(\sum_{r'=0}^{r} \binom{n}{r'} (m-1)^{r'} - 1\right) \left(\eta - \sum_{r'=0}^{r} \binom{n}{r'} (m-1)^{r'}\right)
\end{aligned} \tag{3.110}$$

### 3.4.3 $\mathcal{M}^{(\overline{r})}$-separability

In Theorem 5, the rank of $M^{(r)}$ is given in terms of $r$, $n$, and $m$. From the discussion leading up to Theorem 5, it should be evident that $\mathcal{M}^{(\overline{r})}$-separability is an example of regular separability. Consequently, we can apply Theorem 3 to determine the dimension of the matrices exhibiting $\mathcal{M}^{(\bar{r})}$-separability:

$$\begin{aligned}
\dim\Big(\mathcal{G}(\mathcal{M}^{(\bar{r})})\Big) &= \eta^2 - \eta - \Big(\mathrm{rank}(M^{(1)}) - 1\Big)\Big(\eta - \mathrm{rank}(M^{(1)})\Big)\Big) \\
&\quad - \sum_{r=2}^{\bar{r}} \Big(\mathrm{rank}\Big(M^{(r)}\Big) - \mathrm{rank}\Big(M^{(r-1)}\Big)\Big)\Big(\eta - \mathrm{rank}\Big(M^{(r)}\Big)\Big) \\
&= \eta^2 - \eta - n(m-1)\Big(\eta - 1 - n(m-1)\Big) \\
&\quad - \sum_{r=2}^{\bar{r}} \binom{n}{r}(m-1)^r\Big(\eta - \sum_{r'=0}^{r}\binom{n}{r'}(m-1)^{r'}\Big) \\
&= \eta^2 - 1 - \sum_{r=0}^{\bar{r}} \binom{n}{r}(m-1)^r\Big(\eta - \sum_{r'=0}^{r}\binom{n}{r'}(m-1)^{r'}\Big) \quad . \qquad (3.111)
\end{aligned}$$

The following corollary summarizes this result.

**Corollary 10.** *When $m_j = m$ for all $j$,*

$$\begin{aligned}
\dim\Big(\mathcal{G}(\mathcal{M}^{(\bar{r})})\Big) &= \eta^2 - 1 - \sum_{r=0}^{\bar{r}} \binom{n}{r}(m-1)^r\Big(\eta - \sum_{r'=0}^{r}\binom{n}{r'}(m-1)^{r'}\Big) \\
&= \dim(\mathcal{G}) - \sum_{r=1}^{\bar{r}} \binom{n}{r}(m-1)^r\Big(\eta - \sum_{r'=0}^{r}\binom{n}{r'}(m-1)^{r'}\Big) \\
&= \dim\Big(\mathcal{G}(M^{(\bar{r})})\Big) - \sum_{r=1}^{\bar{r}} \binom{n}{r}(m-1)^r\Big(\sum_{r'=r+1}^{\bar{r}}\binom{n}{r'}(m-1)^{r'}\Big) \quad . \qquad (3.112)
\end{aligned}$$

## 3.5 The evolving dimension of separability

In this section, we will focus on the evolving dimension of the set of matrices exhibiting $M^{(r)}$ and $\mathcal{M}^{(\bar{r})}$-separability. Our focus will be limited to cases where all automata have the same number of possible statuses, ($m_j = m$ for all $j$); consequently, the parameters to be varied are $r$ (or $\bar{r}$), $n$, and $m$. Up to this point, our notation for $M^{(r)}$-separability has hid the fact that our discussions are always in regards to particular values for $n$ and $m$. Because all three parameters $r$ (or $\bar{r}$), $n$, and $m$ will be varied as we analyze the evolving dimension of separability, we wish to incorporate $m$ and $n$ into our notation for $M^{(r)}$-separability. Hence in this section only, we write $M^{(r,n,m)}$-separability and $\mathcal{M}^{(\bar{r},n,m)}$-separability to indicate the particular values of $n$ and $m$.

### 3.5.1 Variation of $\dim\big(\mathcal{G}(M^{(r,n,m)})\big)$ with $r$, for fixed $n$ and $m$

In this section, we consider the variation in the dimension of the matrices exhibiting $M^{(r,n,m)}$-separability, as $r$ varies. In essence, we are investigating how the dimension changes when we assume that higher-order marginals can be propagated ($r$ increasing). We begin with a simple case when $m = 2$ (all automata have only two possible statuses), developing some intuition. We then proceed to the general $m$ case. As $n$ is fixed, our analysis involves varying $r$ between 1 and $n$.

### Special case: $m = 2$

Consider Corollary 9, and the analytical expression it provides for the dimension of $\mathcal{G}(M^{(r,n,m)})$:

$$\dim\left(\mathcal{G}(M^{(r,n,m)})\right) = \eta^2 - \left(\sum_{r'=0}^{r}\binom{n}{r'}(m-1)^{r'}\right)\left(\sum_{r'=r+1}^{n}\binom{n}{r'}(m-1)^{r'}\right) - \sum_{r'=0}^{r}\binom{n}{r'}(m-1)^{r'} \tag{3.113}$$

in the $m = 2$ case. The first term on the right hand side of (3.110) can be ignored, as it does not depend on $r$, and for the moment, let's also ignore the third term. Because the two factors of the middle term must sum to $\eta = m^n = 2^n$, the middle term is symmetric about $r = \frac{n-1}{2}$ as a function of $r$ (this is due to the binomial identity $\binom{n}{k} = \binom{n}{n-k}$). Moreover, when the leading '-' sign is considered, the middle term as a function of $r$ attains a unique minimum at $r = \frac{n-1}{2}$ when $n$ is odd, or at both $r = \frac{n}{2} - 1$ and $r = \frac{n}{2}$ when $n$ is even. The reintroduction of the final term does not appreciably change this characterization. Although it destroys the exact symmetry about $r = \frac{n-1}{2}$, there is still rough symmetry in a neighborhood around $r = \frac{n-1}{2}$ that expands as $n$ grows; in addition, when $n$ is even, the minimum at $r = \frac{n}{2}$ becomes unique (and thus in general for both even and odd $m$, $\dim\left(\mathcal{G}(M^{(r)})\right)$ attains its minimum at $r = \lfloor \frac{n}{2} \rfloor$).[9]

This may be a bit counterintuitive. One may be lead to believe that in going from an assumption that the univariate marginals propagate, $M^{(1)}$-separability, to one where the bivariate marginals propagate, $M^{(2)}$-separability, we would be assuming a 'richer' model. For our purposes, 'richness' corresponds to the number of free parameters of a probabilities model. This intuition would be supported by the growth in the size of the matrix propagating the partial information ($H^{(1)}$ would be a $nm \times nm$ matrix, while $H^{(2)}$ would be a $(nm)^2 \times (nm)^2$ matrix). Yet such intuition is countered by the fact that $M^{(1)}$-separability is a richer model than $M^{(2)}$-separability. This becomes obvious by comparing the size of the block of 0s in (3.3) that is necessitated by each of the two instances of separability. In fact, $M^{(r,n,m)}$-separability is a richer model than $M^{(r+1)}$-separability for all $r < \frac{n}{2}$ (when $m = 2$). By assuming that there is more information to propagate, one does not necessary obtain a model with more freedom. Rather, the dimension of $\mathcal{G}(M^{(r,n,m)})$ (and hence the richness of our model) is roughly symmetric about $r = \frac{n-1}{2}$. For example, when $m = 2$ and $n = 11$, there are roughly the same number of free parameters in assuming $M^{(3)}$-separability as $M^{(7)}$-separability (see Fig. 3-8). Examples illustrating some of this interesting behavior are illustrated in Figures 3-7, 3-8, and 3-9 for the cases of $n = 4$, $n = 11$, and $n = 30$, respectively. Note that when $r = 0$ or when $r = n$, we have effectively assumed nothing, and hence in these cases the dimension of $\mathcal{G}(M^{(r,n,m)})$ is the same as the dimension of $\mathcal{G}$.

### General case

When $m > 2$, we cannot simply use the binomial identity, $\binom{n}{k} = \binom{n}{n-k}$, to determine the value of $r$ for which $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ will be minimized. Let's look at this general case more rigorously, focusing first on determining the value of $r$ for which $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ is minimized (for fixed $n$ and $m$), which will be denoted as $r^*$. We must minimize (3.110) as

[9] This analysis will be made more rigorously in the general setting when we consider values $m$ greater than or equal to 2.

Figure 3-7: Variation of $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ as a function of $r$ ($m = 2$, $n = 4$).

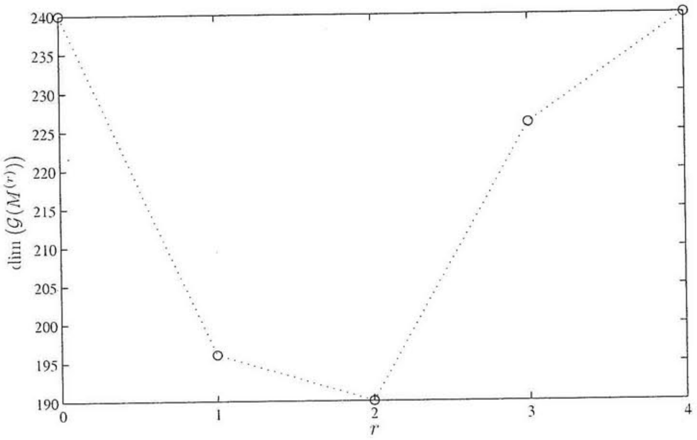


Figure 3-8: Variation of $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ as a function of $r$ ($m = 2$, $n = 11$).

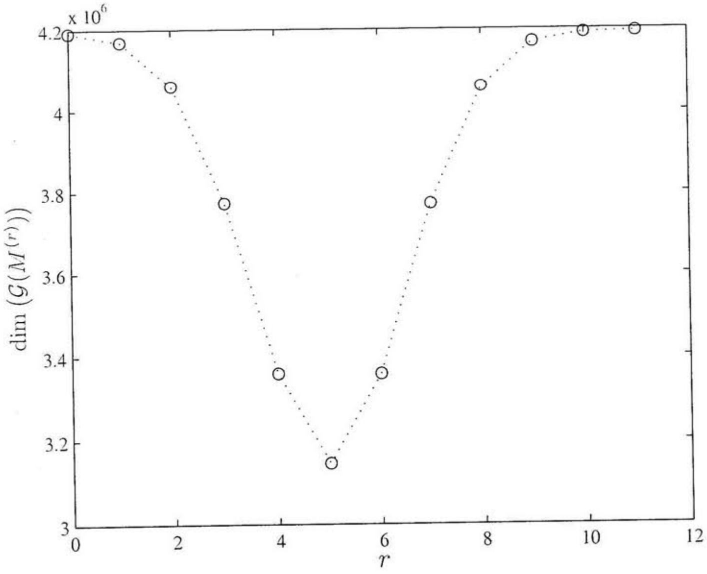
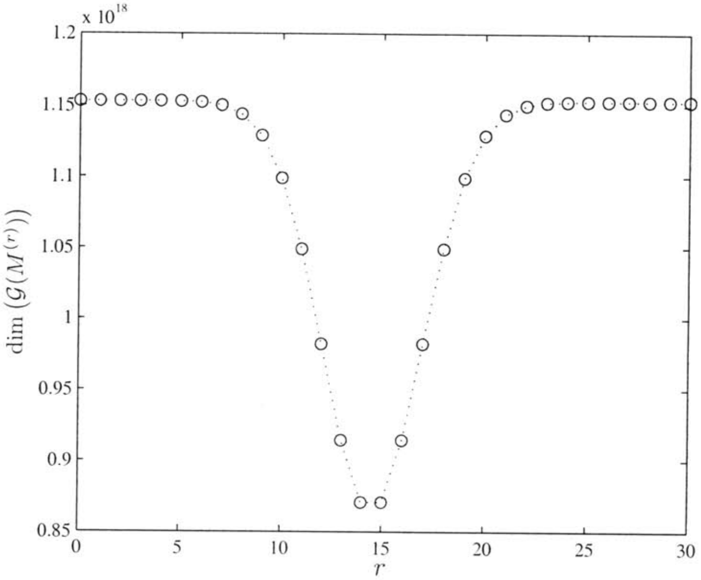


Figure 3-9: Variation of $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ as a function of $r$ ($m = 2$, $n = 30$).



a function of $r$, which can be best visualized as minimizing

$$\dim\left(\mathcal{G}(M^{(r,n,m)})\right) = \eta^2 - (\eta - p(r))p(r) - p(r) \quad , \tag{3.114}$$

where $p(r)$ is the rank of $M^{(r,n,m)}$ derived in Theorem 5 as

$$p(r) = \sum_{r'=0}^{r} \binom{n}{r'} (m-1)^{r'} \quad . \tag{3.115}$$

As (3.115) is only defined at integer $r$, it is convenient to extend $p(r)$ into a strictly increasing differentiable function over $[0, n]$ by interpolation, and then utilize calculus to find a minimum. We denote this extended function defined over $[0, n]$ as $\overline{p}(r)$,[10] and proceed by minimizing

$$\eta^2 - (\eta - \overline{p}(r))\overline{p}(r) - \overline{p}(r) \quad . \tag{3.116}$$

Any value of $r \in (0, n)$ for which (3.116) is minimized must satisfy

$$-(1+\eta)\frac{\partial}{\partial r}\overline{p}(r) + 2\overline{p}(r)\frac{\partial}{\partial r}\overline{p}(r) = 0 \quad . \tag{3.117}$$

As $\frac{\partial}{\partial r}\overline{p}(r) > 0$ by construction, it follows that there is a unique minimum, which we denote by $\overline{r}^*$, for which

$$\overline{p}(\overline{r}^*) = \frac{1}{2} + \frac{1}{2}\eta = \frac{1}{2} + \frac{1}{2}p(n) \quad . \tag{3.118}$$

Much as we did in the $m = 2$ case, we now wish to relate the integer $r$ that minimizes

[10] We do not state how the interpolation is performed, as such details are irrelevant, only insisting that the resulting extended function be strictly increasing and differentiable. The gamma function would suffice.

(3.116) ($r^*$), to something that is more tractable: the integer value of $r$ that minimizes only the middle term on the right hand side of (3.116). In pursuit of this objective, we attempt to find the unique integer $r_0$ at which

$$p(r_0 - 1) < \frac{1}{2}p(n) \leq p(r_0) \quad , \tag{3.119}$$

and first relate such a value to $\overline{r}^*$. Note that $r_0$ or $r_0 - 1$ must be the integer that minimizes the middle term of the right hand side of (3.114).

By definition, $r_0$ must satisfy (3.119), and therefore,

$$\begin{aligned} p(r_0 + 1) &= \binom{n}{r_0 + 1}(m-1)^{r_0+1} + p(r_0) \\ &> \frac{1}{2} + p(r_0) \\ &\geq \frac{1}{2} + \frac{1}{2}p(n) \\ &= \overline{p}(\overline{r}^*) \quad , \end{aligned} \tag{3.120}$$

and in addition,

$$\begin{aligned} p(r_0 - 1) &< \frac{1}{2}p(n) \\ &< \frac{1}{2} + \frac{1}{2}p(n) \\ &= \overline{p}(\overline{r}^*) \quad , \end{aligned} \tag{3.121}$$

where $\overline{r}^*$ is the unique minimum of the extended function $\overline{p}(r)$ characterized in (3.118).

By (3.120) and (3.121) and the fact that $\overline{p}(\cdot)$ is strictly increasing by construction, we conclude that

$$r_0 - 1 < \overline{r}^* < r_0 + 1 \quad . \tag{3.122}$$

Because the extended function of (3.116) has a unique minimum, the integer at which (3.114) is minimized (and thus the value of $r$ corresponding to the least rich instance of $M^{(r,n,m)}$-separability), must be $\lceil \overline{r}^* \rceil$ or $\lfloor \overline{r}^* \rfloor$. As $r_0$ is itself integer, it then follows from (3.122) that

$$r_0 - 1 \leq r^* \leq r_0 + 1 \quad . \tag{3.123}$$

Bounds on the value of $r$ for which $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ is minimized are given in (3.123) in terms of $r_0$. Although it may not seem immediately obvious, $r_0$ can be determined in terms of the parameters $m$ and $n$. The key is to relate $r_0$ to the median of a binomial random variable with parameters $n$ and $\rho = \frac{m-1}{m}$. As is well known, the mean of such a binomial random variable is $n\rho$; what may not be as well known is that the median of a binomial random variable always coincides roughly with the mean, and in particular is always either $\lfloor n\rho \rfloor$ or $\lceil n\rho \rceil$ [46]. Let's denote the median of a binomial random variable by $k_0$, and by definition of being the median, $k_0$ must satisfy

$$F(k_0 - 1) < \frac{1}{2} \leq F(k_0) \quad , \tag{3.124}$$

where $F(\cdot)$ is the cumulative distribution function for the binomial random variable, or equivalently,

$$\sum_{l=0}^{k_0-1} \binom{n}{l} \left(\frac{m-1}{m}\right)^l \left(\frac{1}{m}\right)^{n-l} < \frac{1}{2} \leq \sum_{l=0}^{k_0} \binom{n}{l} \left(\frac{m-1}{m}\right)^l \left(\frac{1}{m}\right)^{n-l} \quad . \tag{3.125}$$

If we multiply all terms in (3.125) by $m^n = \sum_{l=0}^{n} \binom{n}{l}(m-1)^l$, we obtain

$$\sum_{l=0}^{k_0-1} \binom{n}{l}(m-1)^l < \frac{1}{2}\sum_{l=0}^{n} \binom{n}{l}(m-1)^l \leq \sum_{l=0}^{k_0} \binom{n}{l}(m-1)^l \quad . \tag{3.126}$$

However, the alternative definition for $k_0$ given in (3.126) is precisely the definition of $r_0$ as the value of $r$ satisfying (3.119). Consequently, $r_0 = k_0$, and as $k_0$ equals $\lfloor n\frac{m-1}{m} \rfloor$ or $\lceil n\frac{m-1}{m} \rceil$ by [46], the following theorem can be deduced.

**Theorem 6.** *For fixed $n$ and $m$, the dimension of $\mathcal{G}(M^{(r,n,m)})$, or equivalently, the number of parameters necessary to specify uniquely any transition matrix of a $M^{(r,n,m)}$-separable model, decreases monotonically as $r$ increases, until attaining a minimum at $r^*$, after which it increases monotonically. The value $r^*$ at which $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ attains its minimum (as a function of $r$) can be bounded in terms of $m$ and $n$ as follows:*

$$\left\lfloor n\frac{m-1}{m} \right\rfloor - 1 \leq r^* \leq \left\lceil n\frac{m-1}{m} \right\rceil + 1 \quad . \tag{3.127}$$

Note that when $m = 2$, Theorem 6 is in agreement with our earlier conclusion that $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ attains its minimum at $\lfloor \frac{n}{2} \rfloor$.

For general $m$, we conclude that $M^{(r,n,m)}$-separability is a richer model than $M^{(r+1)}$-separability for all $r < \lfloor n\frac{m-1}{m} \rfloor - 1$. Again this is interesting, as it says that in the case of general $m$, by assuming that higher order marginals propagate rather than lower order marginals, we are assuming a model with less freedom, provided that such higher-order marginals involve a fraction of automata that is less than $\frac{m-1}{m}$. When $m > 2$, there does not seem to be a straightforward means to argue symmetry about the minimum $r_*$. Hence the only claim we can make in the general case of the behavior of $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ as $r$ varies (for fixed $m$ and $n$), is in regards to the value of $r$ at which $\mathcal{G}(M^{(r,n,m)})$ is of minimal dimension. When $n$ is large, however, there does still appear to be rough symmetry about the minimum; this fact will be discussed rigorously in Section 3.5.5. Figures 3-10, 3-11, and 3-12 illustrate the evolution of the dimension of $M^{(r,n,m)}$-separability as a function of $r$ for $m = 3$, and $n = 4$, $n = 12$ and $n = 40$, respectively. Fig. 3-13 does the same for $m = 5$ and $n = 30$.

### 3.5.2 Variation of $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ with $n$, for fixed $r$ and $m$

We now analyze the variation in the dimension of the set of matrices exhibiting $M^{(r,n,m)}$-separability, as $n$ grows. Effectively, we consider how the dimension of models with $r$th-order marginals propagating changes as we consider larger numbers of automata. As the dimension of such models will grow without bound as $n$ grows (exponentially, in fact, as will soon be clear), the real interest lies in comparing its growth to the otherwise unconstrained Markovian case.

Figure 3-10: Dimension of $\mathcal{G}\left(M^{(r,n,m)}\right)$, as a function of $r$ ($m = 3$, $n = 4$).

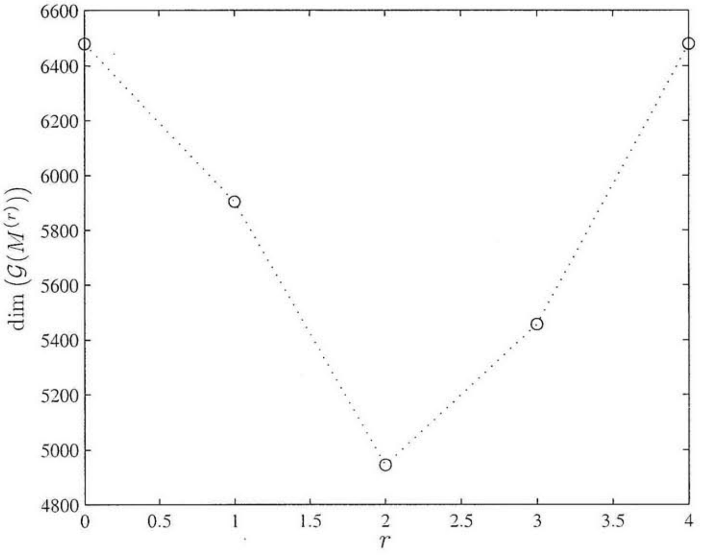


Figure 3-11: Dimension of $\mathcal{G}\left(M^{(r,n,m)}\right)$ as a function of $r$ ($m = 3$, $n = 12$).

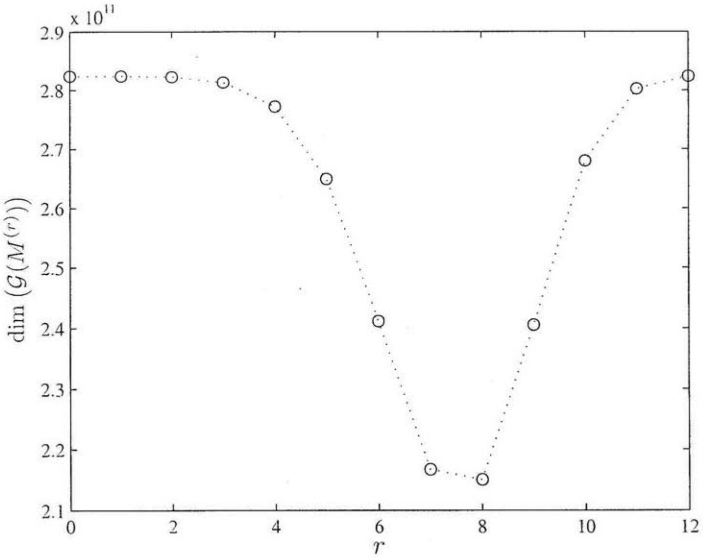


Figure 3-12: Dimension of $\mathcal{G}\left(M^{(r,n,m)}\right)$ as a function of $r$ ($m = 3$, $n = 40$).

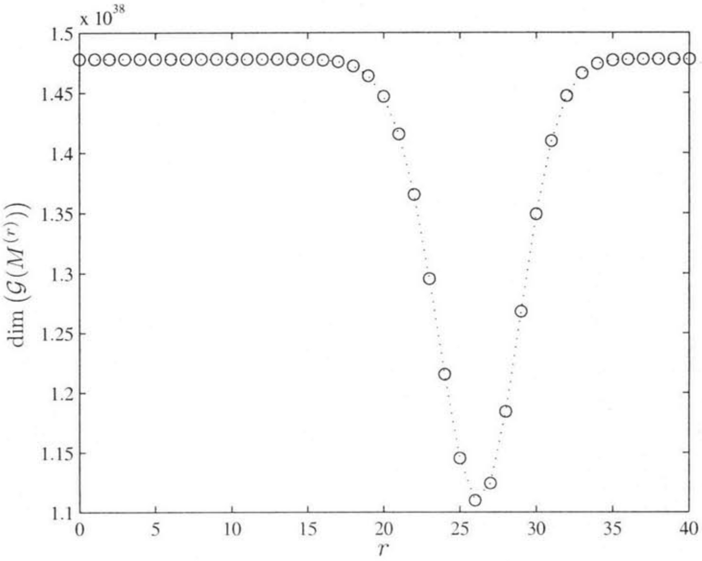
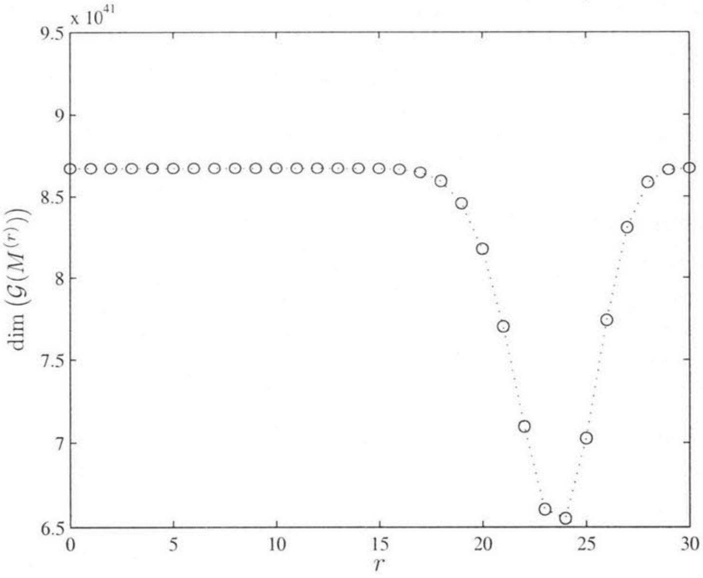


Figure 3-13: Dimension of $\mathcal{G}\left(M^{(r,n,m)}\right)$ as a function of $r$ ($m = 5$, $n = 30$).



In order to analyze the asymptotic behavior of $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ as $n$ increases, we need to develop asymptotic equivalences for some terms that appear in the expression for $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ given in Corollary 9. As shown in (B.6), $\binom{n}{k} = \Theta(n^k)$ when $k$ is fixed. As the rank of $M^{(r,n,m)}$ (given in Theorem 5) consists of a sum of $r$ such binomial terms, it is straightforward to argue that

$$\operatorname{rank}\Big(M^{(r,n,m)}\Big) = \Theta(n^r) \quad , \tag{3.128}$$

when $r$ and $m$ are fixed.

Applying this asymptotic result to Theorem 2, we derive an asymptotic version of Corollary 9:

$$\begin{aligned}\dim\left(\mathcal{G}(M^{(r,n,m)})\right) &= m^n(m^n-1) - \Big(\Theta(n^r)-1\Big)\Big(m^n - \Theta(n^r)\Big)\\ &= m^n(m^n-1) - \Theta(n^r)\Big(m^{n-O(1/n)}\Big)\\ &= m^n(m^n-1) - m^{n+\Theta(\log n)}\\ &= m^{2n-O(1/n)} \quad ,\end{aligned} \tag{3.129}$$

where we have used (B.17) and (B.15) to obtain the final expression. Evidently by (3.129), the dimension of $\mathcal{G}(M^{(r,n,m)})$ grows exponentially as a function of $n$. Of particular interest, however, is how the dimension of $\mathcal{G}(M^{(r,n,m)})$ changes relative to the dimension of the otherwise unconstrained Markovian case. In the past, we have denoted the set of otherwise unconstrained transition matrices as $\mathcal{G}$; however to emphasize $n$ and $m$ as parameters, we now denote the set of otherwise unconstrained $m^n \times m^n$ stochastic transition matrices as $\mathcal{G}^{(n,m)}$. Using the third equality of (3.129) to evaluate the difference between the dimensions of $\mathcal{G}^{(n,m)}$ and $\mathcal{G}(M^{(r,n,m)})$, we produce the following theorem.

**Theorem 7.** *The difference in the number of parameters necessary to uniquely specify an otherwise unrestricted Markovian network and one that satisfies $M^{(r,n,m)}$-separability grows exponentially with $n$, and in particular,*

$$\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(M^{(r,n,m)})\right) = m^{n+\Theta(\log n)} \quad . \tag{3.130}$$

*Nevertheless, the number of free parameters in both cases is asymptotically equal, i.e.,*

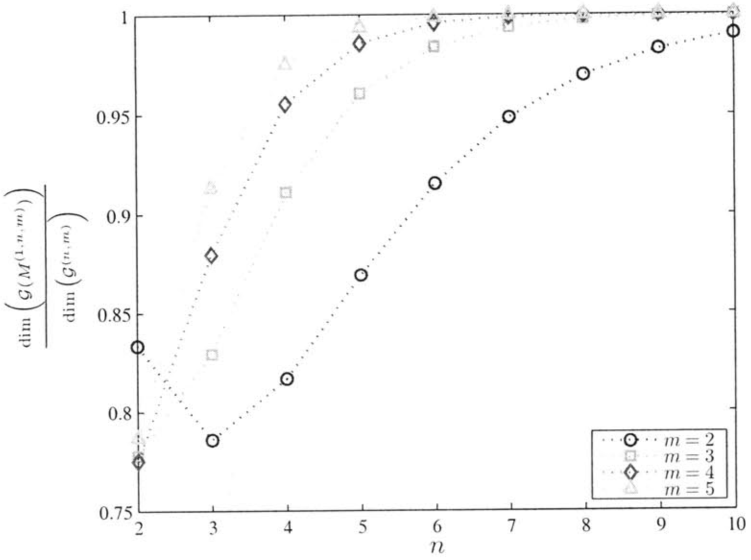
$$\dim\left(\mathcal{G}^{(n,m)}\right) \sim \dim\left(\mathcal{G}(M^{(r,n,m)})\right) \quad , \tag{3.131}$$

*and not only does the ratio of dimensions converge to 1, but it converges exponentially fast at an asymptotic rate of* $\ln m$.

The fact that the number of free parameters is equal asymptotically, whether or not $M^{(r,n,m)}$-separability is assumed, is clear by considering the relative difference between the dimensions of $\mathcal{G}^{(n,m)}$ and $\mathcal{G}(M^{(r,n,m)})$:

$$\begin{aligned}\frac{\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(M^{(r,n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} &= \frac{m^{n+\Theta(\log n)}}{m^n(m^n-1)}\\ &= m^{-(n-\Theta(\log n))}\\ &\longrightarrow 0 \quad \text{as } n \to \infty \quad .\end{aligned} \tag{3.132}$$

Figure 3-14: Fraction of free parameters for $M^{(r,n,m)}$-separability, relative to the unconstrained case, as a function of $n$ ($r = 1$).



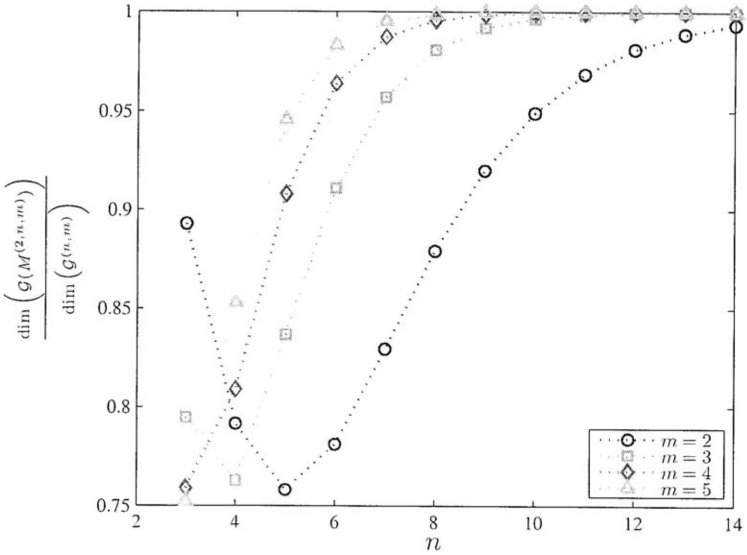
Note from (3.132) that not only does the relative difference converge to 0, but it converges exponentially fast at an asymptotic rate of

$$\lim_{n\to\infty} \frac{1}{n} \ln \frac{\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(M^{(r,n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} = \ln m \quad . \tag{3.133}$$

After the rank of $M^{(1,n,m)}$ was derived in Section 3.4.1, the evolution of the dimension of $\mathcal{G}(M^{(1,n,m)})$ was illustrated for various values of $n$ ($m$ fixed) in Tables 3.1 and 3.2. The tables suggested that the ratio of the dimensions of $\mathcal{G}(M^{(1,n,m)})$ and $\mathcal{G}^{(n,m)}$ converged rapidly to 1 as $n$ increased. In fact, as illustrated in (3.132), such convergence is exponential, for all values of $r$. Figures 3-14 and 3-15 illustrate this exponential convergence as a function of $n$ for several values of $m$ simultaneously, in the $r = 1$ and $r = 2$ cases, respectively. As corroborated by (3.132), the rate of exponential convergence is greater for larger values of $m$. Interestingly, for small values of $m$ in both the $r = 1$ and $r = 2$ cases, the ratio of the dimensions of $\mathcal{G}(M^{(r,n,m)})$ and $\mathcal{G}^{(n,m)}$ drops initially.

The rather interesting conclusion to make from this analysis is that as $n$ increases, networks exhibiting $M^{(r,n,m)}$-separability approach the generality of otherwise unconstrained networks, having at their disposal a fraction of the parameters of the unconstrained models that rapidly approaches 1. This is because the constraints of $M^{(r,n,m)}$-separability (consult Section 3.3 for a visualization) increasingly become negligible relative to the total number of free parameters. This generality in probabilistic models exhibiting $M^{(r,n,m)}$-separability is intriguing, considering that $M^{(r,n,m)}$-separable models offer substantial computational advantages, especially when $n$ is large. $M^{(r,n,m)}$-separable models have the ability to propagate partial information with computations of complexity polynomial in $n$, rather than exponential in $n$. When $n$ is large, the loss of generality in assuming $M^{(r,n,m)}$-separability

Figure 3-15: Fraction of free parameters for $M^{(r,n,m)}$-separability, relative to the unconstrained case, as a function of $n$ ($r = 2$).



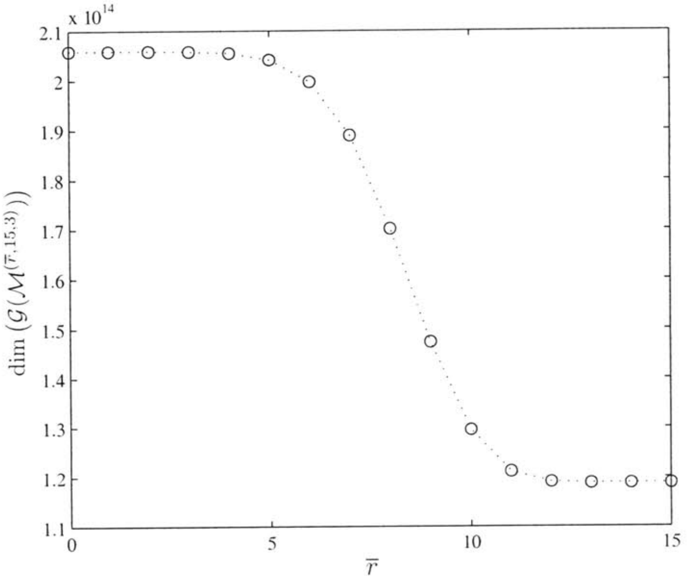
is relatively minimal, while the computational tractability that is gained is substantial. When one has the advantage of being able to engineer or design a probabilistic system, it becomes increasingly less restrictive, in a relative sense, to ensure $M^{(r,n,m)}$-separability when $n$ is large. But one should not take this too far. Should there exists an $M^{(r,n,m)}$-separable system that matches a high percentage of constraints for an arbitrary Markovian network of stochastic automata (Section 3.6 will explain this possibility further), one cannot conclude that the given network is $M^{(r,n,m)}$-separable. One must check the specific constraints that permit the propagation of partial information, the number of which is increasing exponentially with $n$.

### 3.5.3 Variation of $\dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right)$ with $\overline{r}$, for fixed $n$ and $m$

Consider the variation in the dimension of the set of transition matrices exhibiting $\mathcal{M}^{(\overline{r},n,m)}$-separability, as $\overline{r}$ varies between 1 and $n$. In spirit, this is similar to the variation discussed in Section 3.5.1, except as we increase $\overline{r}$, it is assumed that not only do the $\overline{r}$th-order marginals propagate, but the $r$th-order marginals, for all $r < \overline{r}$, still propagate, too. Because constraints accumulate as $\overline{r}$ grows, the dimension of $\mathcal{M}^{(\overline{r},n,m)}$-separability will monotonically decrease with increasing $\overline{r}$.

Figures 3-16 and 3-17 illustrate this decrease in dimension of the set of matrices exhibiting $\mathcal{M}^{(\overline{r},n,m)}$-separability, as a function of $\overline{r}$, for different values of $m$ and $n$. The figures suggest that the dimension is relatively constant until a transition region where it makes a sharp transition downward to roughly half its original level, before once again becoming relatively constant. The figures suggest that the transition region from relatively high dimension to relatively low dimension is complete at $\overline{r} \approx \frac{m-1}{m}n$, which one should recall from Section 3.5.1 corresponds to the value at which the dimension of $\mathcal{G}\left(M^{(r,n,m)}\right)$ is minimized

Figure 3-16: Dimension of $\mathcal{G}\Big(\mathcal{M}^{(\overline{r},n,m)}\Big)$ as a function of $\overline{r}$ ($m=3$, $n=15$).



as a function of $r$, for fixed $m$ and $n$. This argument will be made rigorous in Section 3.5.6 for large values of $n$.

### 3.5.4 Variation of $\dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right)$ with $n$, for fixed $\overline{r}$ and $m$

When $\overline{r}$ and $m$ are fixed, and $n$ grows, the dimension of $\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})$ will grow exponentially, which is straightforward to show using Corollary 10 and following techniques similar to those illustrated in Section 3.5.2:

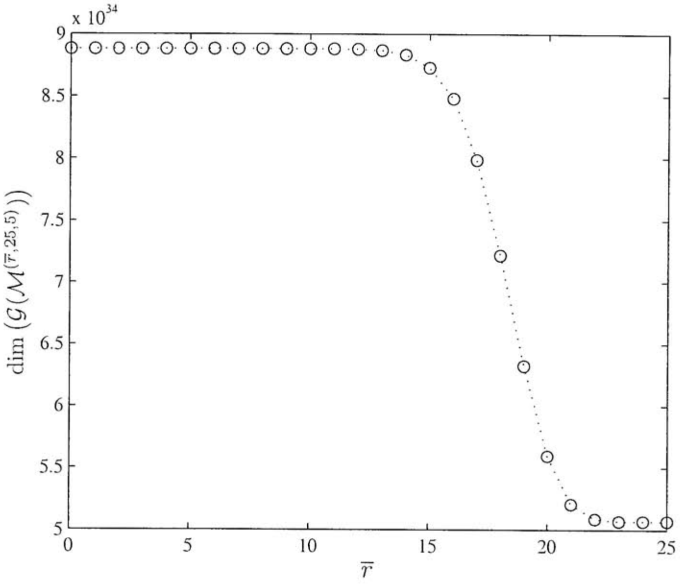
$$\begin{aligned}\dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right) &= m^{2n}-1-\sum_{r=0}^{\overline{r}}\binom{n}{r}(m-1)^r\left(m^n-\sum_{r'=0}^{r}\binom{n}{r'}(m-1)^{r'}\right)\\ &= m^{2n}-m^{n-O(1/n)}\sum_{r=0}^{\overline{r}}\binom{n}{r}(m-1)^r\\ &= m^{2n}-m^{n-O(1/n)}\Theta(n^{\overline{r}})\\ &= m^{2n-O(1/n)}\end{aligned} \tag{3.134}$$

where we have used (B.6), (B.14),(B.15), and (B.17).

Naturally, it is more illuminating to consider the number of free parameters under $\mathcal{M}^{(\overline{r},n,m)}$-separability in comparison to the otherwise unconstrained Markovian case. As

$$\dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right)\leq\dim\left(\mathcal{G}(M^{(\overline{r},n,m)})\right)\quad , \tag{3.135}$$

Figure 3-17: Dimension of $\mathcal{G}\left(\mathcal{M}^{(\overline{r},n,m)}\right)$ as a function of $\overline{r}$ ($m = 5$, $n = 25$).



it can be concluded from (3.130) that

$$\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right) \geq m^{n+\Theta(\log n)} \quad , \tag{3.136}$$

meaning that the absolute difference in the number of free parameters is growing exponentially. For the relative difference, we apply the third equality of (3.134) and find that

$$\begin{aligned} \frac{\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} &= \frac{\Theta(n^{\overline{r}})m^{n-O(1/n)}}{m^{2n-O(1/n)}} \\ &= m^{-(n-\Theta(\log n))} \\ &\longrightarrow 0 \quad \text{as } n \to \infty \quad . \end{aligned} \tag{3.137}$$

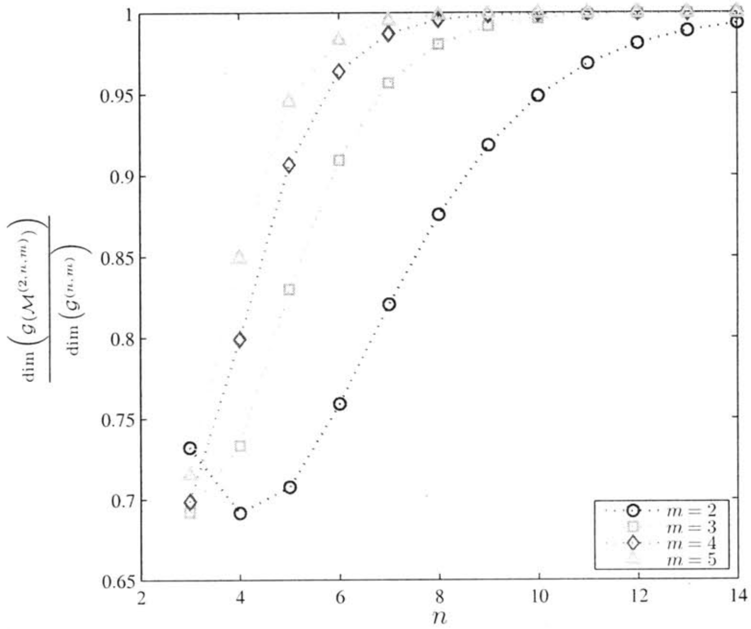
These results are consolidated into the following theorem resembling Theorem 7. In fact, Theorem 8 can be directly deduced from Theorem 7 upon realizing (3.135) in addition to

$$\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right) \leq \sum_{r=1}^{\overline{r}} \left(\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(M^{(r,n,m)})\right)\right) \quad . \tag{3.138}$$

**Theorem 8.** *The difference in the number of parameters necessary to uniquely specify an otherwise unrestricted Markovian network and one that satisfies $\mathcal{M}^{(\overline{r},n,m)}$-separability grows exponentially with $n$, and in particular,*

$$\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right) \geq m^{n+\Theta(\log n)} \quad . \tag{3.139}$$

Figure 3-18: Fraction of free parameters under $\mathcal{M}^{(\overline{r},n,m)}$-separability, relative to the unconstrained case, as a function of $n$ ($\overline{r} = 2$).



*Nevertheless, the number of free parameters in both cases is asymptotically equal, i.e.,*

$$\dim\left(\mathcal{G}^{(n,m)}\right) \sim \dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right) \quad , \tag{3.140}$$

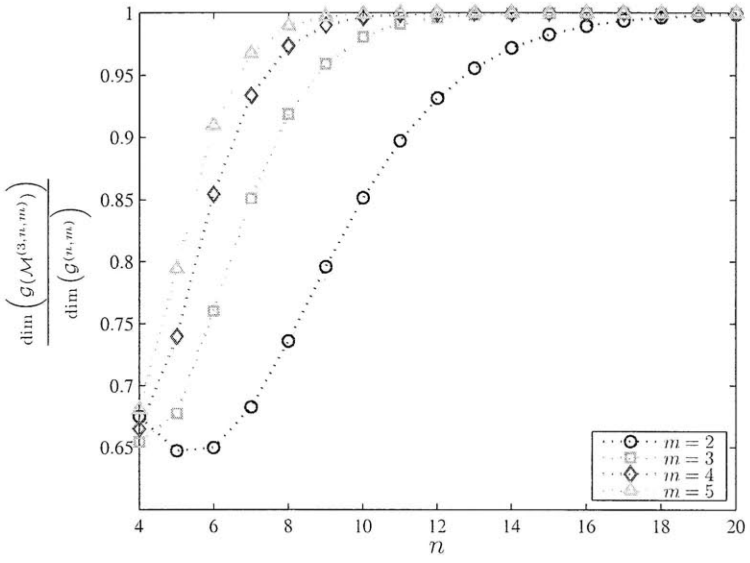
*and not only does the ratio of dimensions converge to 1, but it converges exponentially fast at an asymptotic rate of* $\ln m$.

Figures 3-18 and 3-19 illustrate the exponential convergence of the ratio of dimensions for several values of $m$ simultaneously, in the $\overline{r} = 2$ and $\overline{r} = 3$ cases, respectively. The figures affirm the increase in the rate of exponential convergence as $m$ increases, as claimed by Theorem 8. Qualitatively, the figures are similar to those illustrated in Figures 3-14 and 3-15.

### 3.5.5 Variation of $\dim\left(\mathcal{G}(M^{(r,n,m)})\right)$ with $n$ and $r = \lfloor \alpha n \rfloor$, for fixed $m$, $\alpha$

As was stated in Theorems 7 and 8, as $n$ increases, the relative fraction of free parameters under $M^{(r,n,m)}$-separability (or $\mathcal{M}^{(\overline{r},n,m)}$-separability) compared to the unrestricted case, approaches 1, regardless of the choice of $r$ (or $\overline{r}$). Suppose that $r$ increases with $n$, meaning that progressively higher-order marginals are assumed to propagate as $n$ increases. Let $r = \lfloor \alpha n \rfloor$ with $\alpha \in (0, 1)$. In order for the relative dimension of $\mathcal{G}(M^{(\lfloor \alpha n \rfloor,n,m)})$ to *not* approach 1, it is necessary that $\text{rank}(M^{(\lfloor \alpha n \rfloor,n,m)}) = \Theta(m^n)$, as only then would it be ensured that the block of 0s in (3.51) would have $\Theta(m^{2n})$ entries, on order with the number of entries of $G$. Knowing that $\text{rank}(M^{(\lfloor \alpha n \rfloor,n,m)}) = O(n^{\alpha n})$ does not definitely answer this question; hence, we require the exact equation for $\text{rank}(M^{(r,n,m)})$, as given in Corollary 9.

Figure 3-19: Fraction of free parameters under $\mathcal{M}^{(\overline{r},n,m)}$-separability, relative to the unconstrained case, as a function of $n$ ($\overline{r}=3$).



Using Corollary 9, we derive an equation for the relative difference in the dimensions of the set of transition matrices exhibiting $M^{(\lfloor\alpha n\rfloor,n,m)}$-separability and the set of transition matrices otherwise unconstrained:

$$\frac{\dim\left(\mathcal{G}^{(n,m)}\right)-\dim\left(\mathcal{G}(M^{(\lfloor\alpha n\rfloor,n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)}$$

$$=\frac{m^n(m^n-1)-m^n(m^n-1)+\left(\sum_{r=0}^{\lfloor\alpha n\rfloor}\left(\binom{n}{r}(m-1)^r\right)-1\right)\left(\sum_{r=\lfloor\alpha n\rfloor+1}^{n}\binom{n}{r}(m-1)^r\right)}{m^n(m^n-1)} \tag{3.141}$$

The key to understanding the asymptotic behavior of (3.141) as $n$ increases is to realize that the partial sums of the terms of a binomial expansion can be directly related to the cumulative distribution function of a binomial random variable; this technique was first previewed in Section 3.5.1. Upon making such a connection, standard results for random variables, e.g., Chernoff bounds [47] and the central limit theorem [48], can be applied to illuminate the asymptotic behavior of (3.141).

Proceeding, divide both the numerator and denominator of the right hand side of (3.141) by $m^{2n}$:

$$\frac{\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(M^{(\lfloor \alpha n \rfloor, n, m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} = \frac{\left(\sum_{r=0}^{\lfloor \alpha n \rfloor}\left(\binom{n}{r}\left(\frac{m-1}{m}\right)^r\left(\frac{1}{m}\right)^{n-r}\right) - \frac{1}{m^n}\right)\left(\sum_{r=\lfloor \alpha n \rfloor+1}^{n}\binom{n}{r}\left(\frac{m-1}{m}\right)^r\left(\frac{1}{m}\right)^{n-r}\right)}{\frac{m^n-1}{m^n}} , \tag{3.142}$$

and note that the partial sums can be related to a binomial random variable $\mathsf{z}_n$ with parameters $n$ and $\frac{m-1}{m}$:

$$\frac{\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(M^{(\lfloor \alpha n \rfloor, n, m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} = \frac{\left(\mathbb{P}\left(\mathsf{z}_n \leq \alpha n\right) - \frac{1}{m^n}\right)\left(\mathbb{P}\left(\mathsf{z}_n > \alpha n\right)\right)}{\frac{(m^n-1)}{m^n}} . \tag{3.143}$$

Because $\mathsf{z}_n$ is a sum of $n$ independent and identically distributed Bernoulli random variables with parameter $\frac{m-1}{m}$, by the central limit theorem, it follows that

$$\lim_{n\to\infty} \mathbb{P}\left(\mathsf{z}_n \leq \alpha n\right) = \frac{1}{2} , \tag{3.144}$$

for $\alpha = \frac{m-1}{m}$.

Otherwise, for $\alpha \neq \frac{m-1}{m}$, we can derive the following Chernoff bounds [47]:

$$\mathbb{P}\left(\mathsf{z}_n \leq \alpha n\right) \leq e^{D\left(\alpha \| \frac{m-1}{m}\right)n} \quad \text{for } \alpha < \frac{m-1}{m} \tag{3.145}$$

$$\mathbb{P}\left(\mathsf{z}_n \geq \alpha n\right) \leq e^{D\left(\alpha \| \frac{m-1}{m}\right)n} \quad \text{for } \alpha > \frac{m-1}{m} , \tag{3.146}$$

where $D\left(\alpha \| \frac{m-1}{m}\right)$ is the Kullback-Leibler divergence [49] between two Bernoulli random variables with parameters $\alpha$ and $\frac{m-1}{m}$, respectively.

By incorporating (3.144), (3.145), and (3.146) into (3.143), we obtain the following theorem.

**Theorem 9.**

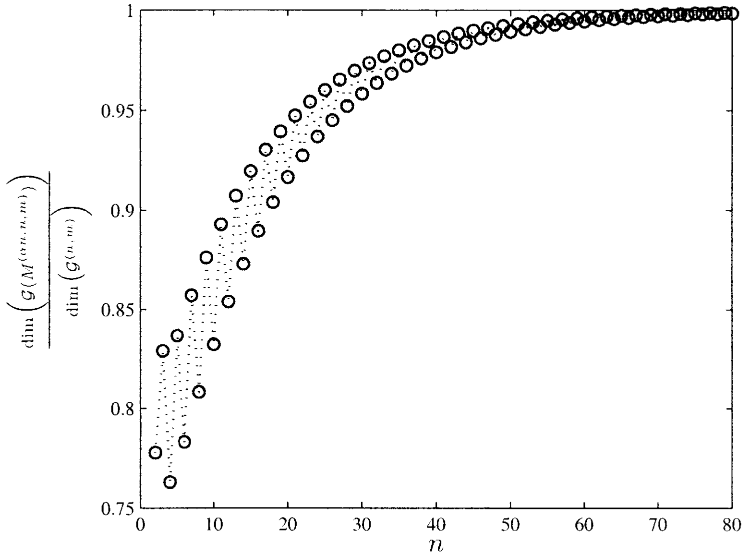
$$\lim_{n\to\infty} \frac{\dim\left(\mathcal{G}(M^{(\lfloor \alpha n \rfloor, n, m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} = \begin{cases} 1 & \textit{if } \alpha \neq \frac{m-1}{m} \\ \frac{3}{4} & \textit{if } \alpha = \frac{m-1}{m} \end{cases} . \tag{3.147}$$

*Moreover, when $\alpha \neq \frac{m-1}{m}$, the convergence of the ratio of the dimensions is exponential, at an asymptotic rate of at least $D\left(\alpha \| \frac{m-1}{m}\right)$.*

Although we do not demonstrate this, one can obtain exponential lower bounds on the probabilities of (3.145) and (3.146) to show that $D\left(\alpha \| \frac{m-1}{m}\right)$ is in fact the asymptotic rate of convergence when $\alpha \neq \frac{m-1}{m}$, not just an upper bound.

Intuitively, the reason why only $\alpha = \frac{m-1}{m}$ results in a fractional drop in relative dimension is that all mass of a binomial random variable with parameters $\frac{m-1}{m}$ and $n$ becomes concentrated about $\frac{m-1}{m}n + O(\sqrt{n})$ for large $n$. The central limit theorem in a rough sense states this fact, and Chernoff bounds definitively state this fact. Consequently, only when $\alpha = \frac{m-1}{m}$ is the rank of $M^{(\lfloor \alpha n \rfloor, n, m)}) = \Theta(m^n)$; visually, all blocks of (3.51) are roughly of

Figure 3-20: Fraction of free parameters for $M^{(\lfloor \alpha n \rfloor, n, m)}$-separability, relative to the unconstrained case, as a function of $n$ ($\alpha = \frac{1}{2} \neq \frac{m-1}{m}$, $m = 3$).



equal size, each having $\frac{1}{2}m^n + o(m^n)$ rows and columns. When $\alpha \neq \frac{m-1}{m}$, as $n$ increases, the block of 0s in (3.51) becomes negligible.

Figures 3-20, 3-21, and 3-22 illustrate such convergence of the ratio of dimensions as a function of $n$ for different values of $\alpha$, when $m = 3$. Because the rate of exponential convergence is governed by $D\left(\alpha \| \frac{m-1}{m}\right)$ and $D\left(\frac{1}{2}\|\frac{2}{3}\right) > D\left(\frac{3}{5}\|\frac{2}{3}\right)$, the convergence of the ratio of dimensions appears to be faster in Fig. 3-20 when $\alpha = \frac{1}{2}$, than in Fig. 3-21, when $\alpha = \frac{3}{5}$.

The takeaway from Theorem 9 is that even when $r = \lfloor \alpha n \rfloor$, the relative difference in the dimensions of the sets of transition matrices that are $M^{(r,n,m)}$)-separable, and those otherwise unconstrained, will be asymptotically equal, provided that $\alpha \neq \frac{m-1}{m}$. Such behavior was suggested in the progression from Fig. 3-11 ($n = 12$) to Fig. 3-12 ($n = 40$), where it is evident that the relative drop in dimension, becomes relatively more concentrated for larger values of $n$ about the minimum at $r_*$ (Theorem 6),[11]

$$\left\lfloor n\frac{m-1}{m} \right\rfloor - 1 \leq r^* \leq \left\lceil n\frac{m-1}{m} \right\rceil + 1 \quad . \tag{3.148}$$

Moreover, one can argue from (3.143) and the central limit theorem that for large $n$, there should be symmetry in the plot of the relative dimension of $\mathcal{G}\left(M^{(r,n,m)}\right)$ as a function of $r$ (for fixed $m$ and $n$) centered about $\frac{m-1}{m}n$, when viewed over intervals of $r$ on the order of $\sqrt{n}$. Figures 3-11 and 3-12 suggest such symmetry for relatively moderate values of $n$.

[11] Note that from our results in this section, we could have ascertained that for large $n$, $r_* = \frac{m-1}{m}n + o(n)$, a result clearly weaker than Theorem 6, which identifies $r_*$ as being in a fixed-length interval for any value of $n$.

Figure 3-21: Fraction of free parameters for $M^{(\lfloor \alpha n \rfloor, n, m)}$-separability, relative to the unconstrained case, as a function of $n$ ( $\alpha = \frac{3}{5} \neq \frac{m-1}{m}$, $m = 3$).

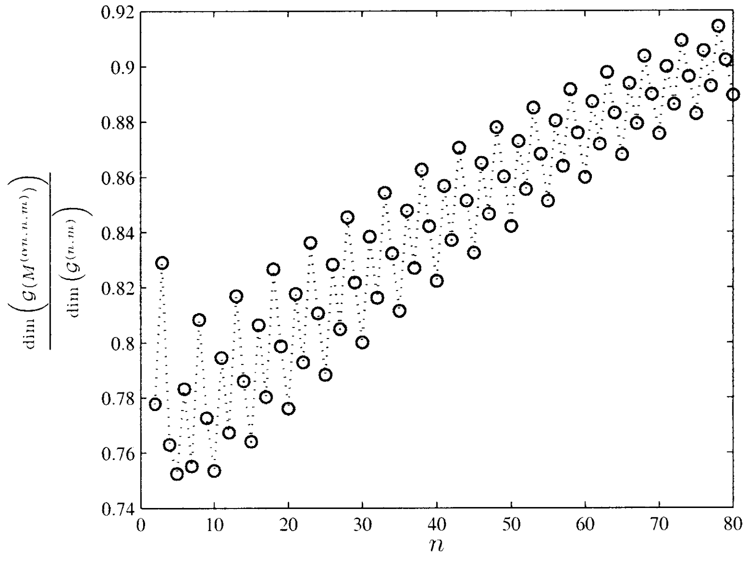
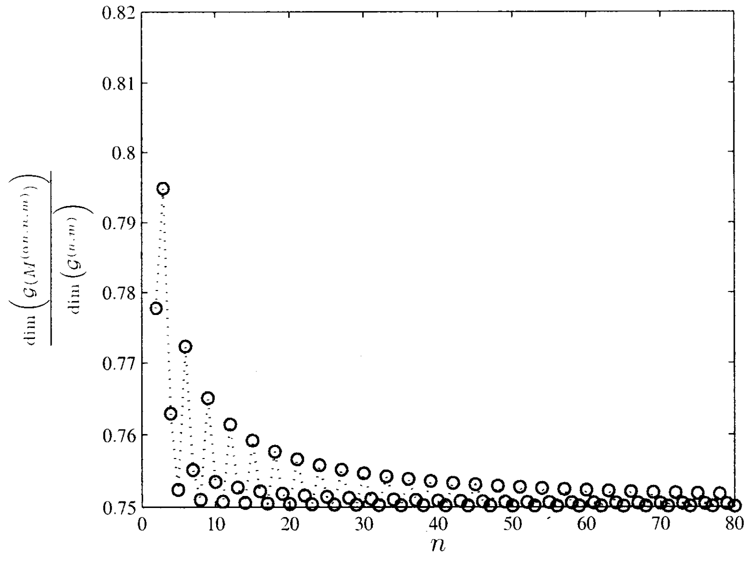


Figure 3-22: Fraction of free parameters for $M^{(\lfloor \alpha n \rfloor, n, m)}$-separability, relative to the unconstrained case, as a function of $n$ ($\alpha = \frac{2}{3} = \frac{m-1}{m}$, $m = 3$).



### 3.5.6 Variation of $\dim\left(\mathcal{G}(\mathcal{M}^{(\overline{r},n,m)})\right)$ with $n$ and $\overline{r} = \lfloor \alpha n \rfloor$, for fixed $m$, $\alpha$

This next section continues the story of last section where $r = \lfloor \alpha n \rfloor$ for $\mathcal{M}(\overline{r}, n, m))$-separability. A specific case of interest is when $\alpha = 1$, i.e., $\mathcal{M}^{(n)}$-separability. Hence, the results of this section will provide insights into the evolving dimension of models for which $r$th-order marginals can be propagated, for all values of $r$.

Before proceeding to the derivations, we would like to offer some visual intuition into the relative dimension of $\mathcal{M}^{(\overline{r},n,m)}$-separability when $\overline{r} = \lfloor \alpha n \rfloor$ and $\alpha \in (0, 1]$ ($m$ fixed). In Section 3.5.5, we showed that when $\alpha < \frac{m-1}{m}$, as $n$ grows large, the relative dimension of $\mathcal{G}(M^{(r,n,m)})$ compared to the unrestricted case approaches 1; equivalently, the relative number of entries in the block of 0s in (3.51) compared to the number of entries in the entire matrix converges to 0. Let's consider $\mathcal{M}^{(\overline{r},n,m)}$-separability from the perspective of first assuming $M^{(\overline{r},n,m)}$-separability, and then introducing the additional constraints to ensure $\mathcal{M}^{(\overline{r},n,m)}$-separability. By comparing the middle matrix on the right hand side of (3.51) with (3.68), it should be clear that the additional constraints to ensure $\mathcal{M}^{(\overline{r},n,m)}$-separability only affect $H^{(\overline{r})}$, requiring that it be block upper triangular. When $\alpha < \frac{m-1}{m}$, clearly $H^{(\lfloor \alpha n \rfloor)}$ will have a relative size approaching 0 as $n \to \infty$. Consequently, when $\alpha < \frac{m-1}{m}$, much like $\mathcal{G}(M^{(\lfloor \alpha n \rfloor,n,m)})$, the relative dimension of $\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n \rfloor,n,m)})$ approaches 1 as $n \to \infty$. On the other hand, when $\alpha > \frac{m-1}{m}$, the size of $H^{(\lfloor \alpha n \rfloor)}$ will approach 1 relative to $G$, and consequently the relative dimension of $\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n \rfloor,n,m)})$ will be roughly $\frac{1}{2}$, because of the imposed upper block diagonal structure on $H^{(\lfloor \alpha n \rfloor)}$ under $\mathcal{M}^{(\lfloor \alpha n \rfloor,n,m)}$-separability. By a similar argument, when $\alpha = \frac{m-1}{m}$, the relative dimension of $\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n \rfloor,n,m)})$ should be roughly $\frac{5}{8}$, as each block in (3.51) will be relatively the same size, with one block consisting of all 0s, and a second block being upper block triangular.

With this visual intuition, the following the arguments should be easier to follow. Evidently,

$$\begin{aligned}
&\frac{\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n \rfloor,n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} \\
&\quad = \frac{\dim\left(\mathcal{G}^{(n,m)}\right) - \dim\left(\mathcal{G}(M^{(\lfloor \alpha n \rfloor,n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} \\
&\qquad + \frac{\dim\left(\mathcal{G}(M^{(\lfloor \alpha n \rfloor,n,m)})\right) - \dim\left(\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n \rfloor,n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} ,
\end{aligned} \tag{3.149}$$

where the second equality follows by rearranging the terms. Using Corollary 10, we obtain an equation for the difference of the dimensions of $\mathcal{G}(M^{(\lfloor \alpha n \rfloor,n,m)})$ and $\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n \rfloor,n,m)})$ relative to the dimension of $\mathcal{G}^{(n,m)}$, that is, the second term on the right hand side of (3.149):

$$\begin{aligned}
&\frac{\dim\left(\mathcal{G}(M^{(\lfloor \alpha n\rfloor,n,m)})\right) - \dim\left(\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n\rfloor,n,m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} \\
&= \frac{\sum_{r=1}^{\lfloor \alpha n\rfloor}\binom{n}{r}(m-1)^r\left(\sum_{k=r+1}^{\lfloor \alpha n\rfloor}\binom{n}{k}(m-1)^k\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} \\
&= \frac{\frac{1}{2}\left(\sum_{r=1}^{\lfloor \alpha n\rfloor}\binom{n}{r}(m-1)^r\right)^2 - \frac{1}{2}\sum_{r=1}^{\lfloor \alpha n\rfloor}\left(\binom{n}{r}(m-1)^r\right)^2}{\dim\left(\mathcal{G}^{(n,m)}\right)} \\
&= \frac{\frac{1}{2}\left(\mathbb{P}\left(\mathsf{z}_n \le \alpha n\right)\right)^2 - \frac{1}{2}\sum_{r=1}^{\lfloor \alpha n\rfloor}\left(\mathbb{P}\left(\mathsf{z}_n = r\right)\right)^2}{\frac{1}{m^{2n}}\dim\left(\mathcal{G}^{(n,m)}\right)} \qquad (3.150)
\end{aligned}$$

where the second equality follows from the first by expanding out both expressions as a sum, and $\mathsf{z}_n$ is a binomial random variable with parameters $\frac{m-1}{m}$ and $n$.

Now consider the two terms in the numerator of the right hand side of (3.150), for each of the three cases, $\alpha < \frac{m-1}{m}$, $\alpha = \frac{m-1}{m}$, and $\alpha > \frac{m-1}{m}$. For the first term, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\mathsf{z}_n \le \alpha n\right)^2 = \begin{cases} 0 & \text{if } \alpha < \frac{m-1}{m} \\ \frac{1}{4} & \text{if } \alpha = \frac{m-1}{m} \\ 1 & \text{if } \alpha > \frac{m-1}{m} \end{cases} \qquad (3.151)$$

where the convergence to 0 and to 1 is exponential at an asymptotic rate of $2D\left(\alpha \| \frac{m-1}{m}\right)$, as evident from our Chernoff bounds (3.145) and (3.146). The convergence to $\frac{1}{4}$ when $\alpha = \frac{m-1}{m}$ is due to the central limit theorem.

For the second term, we note that when $\alpha < \frac{m-1}{m}$,

$$\begin{aligned}
\sum_{r=1}^{\lfloor \alpha n\rfloor}\mathbb{P}\left(\mathsf{z}_n = r\right)^2 &\le \sum_{r=1}^{\lfloor \alpha n\rfloor}\mathbb{P}\left(\mathsf{z}_n = r\right) \\
&= \mathbb{P}\left(\mathsf{z}_n \le \alpha n\right) \\
&\le e^{D\left(\alpha\|\frac{m-1}{m}\right)n} \quad , \qquad (3.152)
\end{aligned}$$

meaning that the second term converges to 0 exponentially at a rate of at least $D\left(\alpha\|\frac{m-1}{m}\right)$.

When $\alpha \ge \frac{m-1}{m}$, let's assume without loss of generality that $n$ is a whole number multiple of $m$. By (B.10), we would be ensured that the mode of $\mathsf{z}_n$ will be at $\frac{m-1}{m}n$, and consequently,

$$\begin{aligned}\sum_{r=1}^{\lfloor \alpha n\rfloor} \mathbb{P}\left(\mathsf{z}_n = r\right)^2 &\le \mathbb{P}\left(\mathsf{z}_n = \frac{m-1}{m}n\right) \sum_{r=1}^{\lfloor \alpha n\rfloor} \mathbb{P}\left(\mathsf{z}_n = r\right) \\ &\le \mathbb{P}\left(\mathsf{z}_n = \frac{m-1}{m}n\right) \\ &\le \frac{1}{\sqrt{2\pi n \frac{m-1}{m^2}}} \quad , \end{aligned} \tag{3.153}$$

where the final inequality is from (B.12).

The upper bound on the second term in the numerator of the right hand side of (3.150) in the case when $\alpha \ge \frac{m-1}{m}$ can be complemented with a lower bound:

$$\begin{aligned}\sum_{r=1}^{\lfloor \alpha n\rfloor} \mathbb{P}\left(\mathsf{z}_n = r\right)^2 &\ge \mathbb{P}\left(\mathsf{z}_n = \frac{m-1}{m}n\right)^2 \\ &\ge \frac{1}{2\pi n \frac{m-1}{m^2}} \quad . \end{aligned} \tag{3.154}$$

Evidently by (3.153) and (3.154), the second term in the numerator of the right hand side of (3.150) converges to 0 at a slow rate (not exponential) when $\alpha \ge \frac{m-1}{m}$: no faster than $\Theta\left(\frac{1}{n}\right)$, but at least as fast as $\Theta\left(\frac{1}{\sqrt{n}}\right)$.

Hence, combining (3.147), (3.149), (3.150), (3.151), (3.152), (3.153), and (3.154), we obtain the following theorem.

**Theorem 10.**

$$\lim_{n\to\infty} \frac{\dim\left(\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n\rfloor, n, m)})\right)}{\dim\left(\mathcal{G}^{(n,m)}\right)} = \begin{cases} 1 & \text{if } \alpha < \frac{m-1}{m} \\ \frac{5}{8} & \text{if } \alpha = \frac{m-1}{m} \\ \frac{1}{2} & \text{if } \alpha > \frac{m-1}{m} \end{cases} \quad . \tag{3.155}$$

*Moreover, when $\alpha < \frac{m-1}{m}$, the convergence of the ratio of the dimensions is exponential, with an asymptotic rate of $D\left(\alpha \| \frac{m-1}{m}\right)$. When $\alpha > \frac{m-1}{m}$, the convergence is at least as fast as $\Theta\left(\frac{1}{\sqrt{n}}\right)$, but no faster than $\Theta\left(\frac{1}{n}\right)$. When $\alpha = \frac{m-1}{m}$, the convergence is no faster than $\Theta\left(\frac{1}{n}\right)$.*

As evident from Theorem 10, $\mathcal{M}^{(n)}$-separability has an asymptotic relative dimension of 1/2.

Figures 3-23, 3-24, and 3-25 demonstrate the changes in the relative dimension of $\mathcal{G}(\mathcal{M}^{(\lfloor \alpha n\rfloor, n, m)})$ as $n$ increases, illustrating the three different possible behaviors. Note the apparent slow convergence in the $\alpha = \frac{m-1}{m}$ and $\alpha > \frac{m-1}{m}$ cases (Figures 3-24 and 3-25).

## 3.6 Interpreting the dimension of separability

This chapter has illustrated many interesting geometric and algebraic properties of separability in general, as well as specific properties exhibited by the canonical examples, $M^{(r)}$-separability and $\mathcal{M}^{(r)}$-separability. One of the more interesting results regarding separability is the evolution of the dimension of models of $M^{(r)}$-separability or $\mathcal{M}^{(r)}$-separability,

Figure 3-23: Fraction of free parameters for $\mathcal{M}^{(\lfloor \alpha n \rfloor, n, m)}$-separability, relative to the unconstrained case, as a function of $n$ ($\alpha = \frac{1}{2} < \frac{m-1}{m}$, $m = 3$).

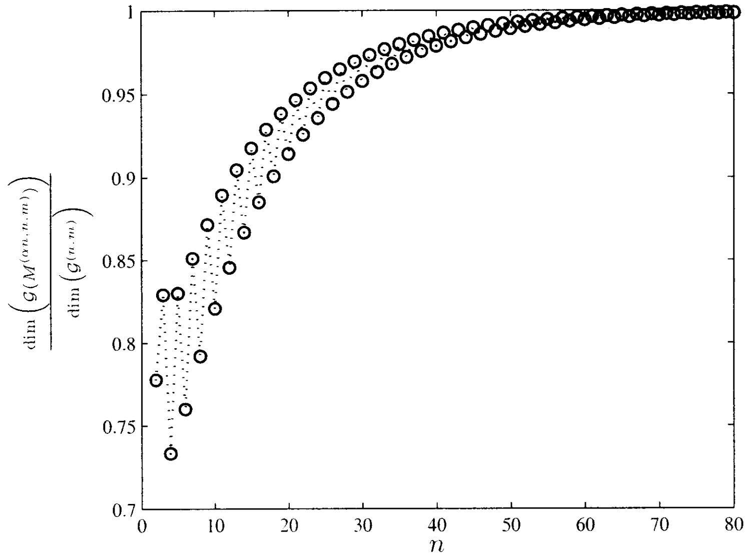


Figure 3-24: Fraction of free parameters for $\mathcal{M}^{(\lfloor \alpha n \rfloor, n, m)}$-separability, relative to the unconstrained case, as a function of $n$ ($\alpha = \frac{2}{3} = \frac{m-1}{m}$, $m = 3$).

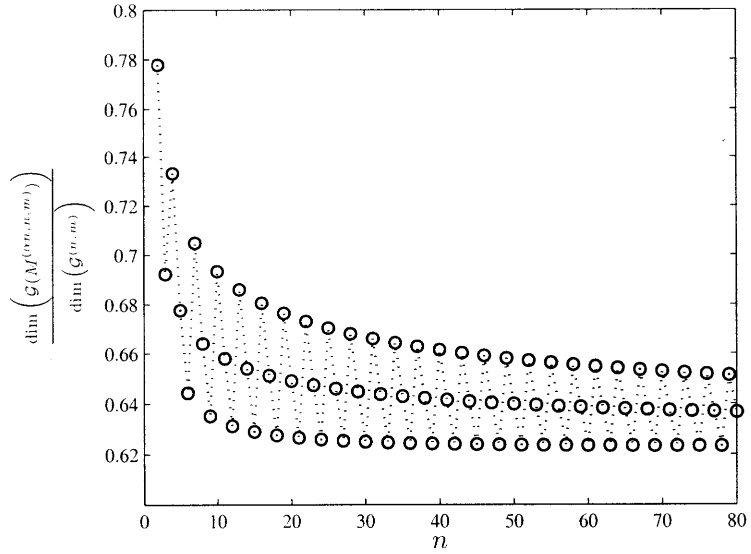
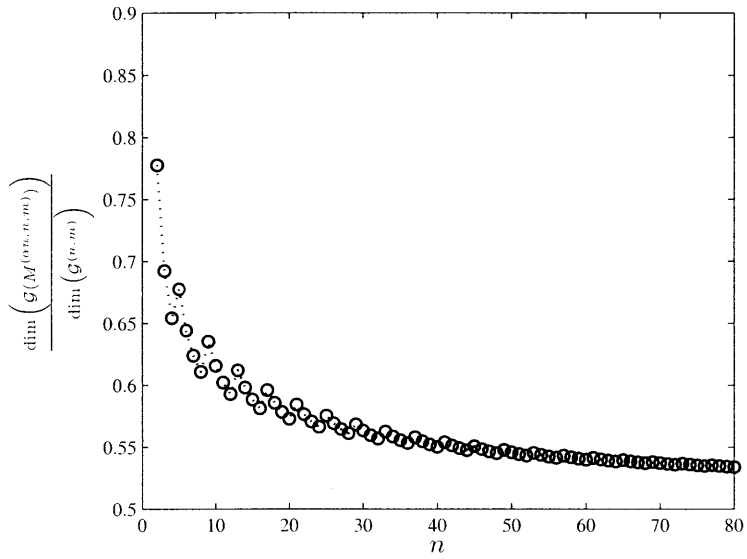


Figure 3-25: Fraction of free parameters for $\mathcal{M}^{(\lfloor \alpha n \rfloor, n, m)}$-separability, relative to the unconstrained case, as a function of $n$ ($\alpha = \frac{3}{4} > \frac{m-1}{m}$, $m = 3$).



both for fixed $r$ and $r$ increasing with $n$. As claimed in Theorems 7, 8, 9, and 10, such models (with some limitations on the growth rate for $\mathcal{M}^{(r)}$-separability) attain a relative fraction of 1 of the dimension of the unrestricted Markovian models, and do so at an exponential rate. The implication is that such higher relative dimension would be suggestive of more generality. This would be an advantage as far as being able to model and capture more general dynamics. On the other hand, one mistakenly may be led to believe that models which are not separable are in fact separable, because of the ability to fit a high percentage of the parameters of a general Markovian model with an $M^{(r)}$-separable model when $n$ is large. This section seeks to illustrate the potentially advantageous implications of $M^{(r)}$-separable models having high relative dimension, while simultaneously serving as a warning regarding the potential traps of high relative dimension.

If nonnegativity could be ignored, the implications of high relative dimension are clear. It would allow separable models to match a higher percentage of linear constraints. In this sense, they would be more general. For example, suppose that $\mathcal{G}(M_*)$ is of relative dimension $d$, $d \in (0,1)$, meaning that

$$\dim\left(\mathcal{G}(M_*)\right) = d \cdot \dim(\mathcal{G}) \quad . \tag{3.156}$$

Ignoring nonnegativity, this would imply that $\mathcal{G}(M_*)$ is an affine subspace of dimension $d \cdot \dim(\mathcal{G})$. On the other hand, suppose that one wishes to impose a set of $(d - \epsilon)\dim(\mathcal{G})$ independent linear constraints[12] on the transition matrix of a Markov chain, where $0 < \epsilon \ll 1$. We will refer to such constraints as *model-imposed constraints*. When $d \approx 1$, one can

[12] In order for the constraints on a row-stochastic matrix to be linearly independent, no more than $\eta - 1$ constraints can be imposed on each row.

think of such a set-up as imposing a number of independent linear constraints nearly equal to the number of entries in the model's transition matrix, e.g., specifying nearly all of the transition probabilities (except we are ignoring nonnegativity for the moment). Ignoring nonnegativity, the set of transition matrices satisfying the $(d-\epsilon)\dim(\mathcal{G})$ model-imposed constraints will be an affine subspace of dimension $(1-d+\epsilon)\dim(\mathcal{G})$. The intersection of the affine subspace of transition matrices satisfying the model-imposed constraints and the affine subspace of matrices satisfying $M_*$-separability, which we will refer to as the *intersection affine subspace*, will be of dimension $\epsilon \cdot \dim(\mathcal{G})$, provided that all such constraints are independent. There will exist an $M_*$-separable model satisfying all of the $(d-\epsilon)\dim(\mathcal{G})$ model-imposed constraints, provided that the intersection affine subspace of dimension $\epsilon \cdot \dim(\mathcal{G})$ intersects the nonnegative orthant.

Hence we have two concerns regarding the existence of $M_*$-separable models satisfying a set of $(d-\epsilon)\dim(\mathcal{G})$ model-imposed constraints. First, when one adds the model-imposed constraints to the linear constraints of $M_*$-separability, is the resulting set of linear constraints independent? Second, provided that all linear constraints are independent, will the intersection affine subspace of matrices intersect the nonnegative orthant? It is seemingly difficult to characterize general conditions to ensure that both concerns will be met. Nevertheless we can build some insight, discussing each concern separately. We will then follow up the discussion with some empirical results.

### 3.6.1 Ensuring consistent constraints

If all model-imposed constraints are chosen 'uniformly at random',[13] then almost surely all constraints will be linearly independent (simultaneously considering the model-imposed constraints and the constraints of $M_*$-separability). However, the given example of model-imposed constraints—specifying particular values for a fraction of the entries of the transition matrix—are far from 'uniformly at random' linear constraints. For example, if one wishes to constrain the $(i,j)$ entry of $G$ to be $\alpha_{ij}$, such a constraint can be expressed as

$$\boldsymbol{e}_i G \boldsymbol{e}_j = \alpha_{ij} \quad , \tag{3.157}$$

or equivalently,

$$(\boldsymbol{e}_j' \otimes \boldsymbol{e}_i')\mathrm{vec}(G) = \alpha_{ij} \quad . \tag{3.158}$$

Note that $\boldsymbol{e}_j' \otimes \boldsymbol{e}_i'$ has only one nonzero entry. The vector $\boldsymbol{e}_j' \otimes \boldsymbol{e}_i'$ is referred to an *orthogonal direction* (for a constraint $AG = b$, each row of $A$ is referred to as an orthogonal direction). We refer to any constraint involving only a few[14] of the entries of $G$ as a sparse constraint. The *entry-wise constraint* of the example is an extreme sparse constraint.

Recall the linear constraints of $M_*$-separability given in (3.48). The matrix on the left hand side of (3.48) is reproduced below:

$$\begin{bmatrix} M_*' \otimes [M_*^\perp]' \\ \mathbb{1}' \otimes I \end{bmatrix} \quad . \tag{3.159}$$

The rows of (3.159) are the orthogonal directions corresponding to the constraints of $M_*$-

[13] For a linear constraint $\boldsymbol{a}'\boldsymbol{x} = b$, we refer to $\boldsymbol{a}'$ as an *orthogonal direction*. Choosing linear constraints 'uniformly at random' corresponds to selecting the orthogonal directions uniformly over the surface of a hypersphere.

[14] What constitute 'a few' will change depending on the contexts.

separability. Inconsistency induced upon incorporating the model-imposed constraints is only possible if the row space of the orthogonal directions corresponding to the additional model-imposed constraints intersects with the row space of (3.159). When one is considering entry-wise constraints uniformly at random, such inconsistency is only possible if all of the nonzero entries corresponding to a vector in the row space of (3.159) are randomly chosen and given entry-wise constraints.

In order to assure consistency, we would like to show that the probability that entry-wise constraints will be imposed on all of the nonzero entries of some vector in the row space of (3.159) will vanish for large $n$. This may seem like a intractable task, as we would need to characterize all the possible patterns of 0 and nonzero entries for vectors in its row space. For example, if in the row space of (3.159) there exists a vector

$$\boldsymbol{v} = (*, *, 0, 0, \ldots, 0, *) \quad , \tag{3.160}$$

where '$*$' indicates a nonzero entry, then we could have inconsistency in our linear constraints if our model-imposed entry-constraints assigned values to all entries of $\vec{(G)}$ that correspond to the nonzero entries of $\boldsymbol{v}$.

Before proceeding, we feel it is helpful to define some terms. Let a vector indicating a particular pattern of 0s and nonzeros be referred to as a *sparse vector*, e.g., $\boldsymbol{v}$ in (3.160). A sparse vector's *count* is its number of nonzero entries, and denoted as $\|\boldsymbol{v}\|$. If we envision a sparse vector $\boldsymbol{v}$ as a pattern of 0s and 1s, its counts is equivalent to its 1-norm. We say that one sparse vector *covers* another sparse vector if its 0 entries are a subset of the other vector's 0 entries. We introduce the notion of covering as it substantially simplifies our task: when inconsistency is not induced for a particular sparse vector, it is not induced for all sparse vector that cover our particular sparse vector. Note that we can specify a sparse vector that corresponds to the entries of $G$ to which entry-wise constraints are imposed; we denote this (random) sparse vector as $\mathbf{c}$. This sparse vector $\mathbf{c}$ will have a count of $(d - \epsilon)dim(\mathcal{G})$.

We consider the idea of defining a subspace's *sparse-basis*. The sparse-basis is a set of sparse vectors in the subspace, such that for every vector in the subspace, it covers some sparse vector in the sparse-basis. The motivation behind a sparse-basis is that it offers a necessary and sufficient means of checking inconsistency. If we can define a sparse-basis for the row space of (3.159), there is inconsistency induced by the model-imposed constraints if and only if $\boldsymbol{c}$ covers a sparse vector in the sparse-basis. If we represent such a sparse-basis as $\mathcal{V}$, what we would want to show is that

$$\mathbb{P}\left( \cup_{\boldsymbol{v} \in \mathcal{V}} \{ \, \mathbf{c} \text{ covers } \boldsymbol{v} \, \} \right) \to 0 \tag{3.161}$$

as $n \to \infty$, meaning that the probability of inconsistency converges to 0 as $n$ grows large.

From this point forward, we will limit our focus to instances where $M_* = M^{(r)}$. Recall that we have illustrated the block Kronecker structure of a matrix with the same rank as $M^{(r)}$ (recall (3.97), (3.98), and (3.99)). We will base our subsequent arguments on this representation of $\mathcal{R}(M^{(r)})$.

Before we attempt to characterize all of the sparse vectors in the row space of (3.159) for $M_* = M^{(r)}$, we first make some observations regarding $M^{(r)}$ and $\left[M^{(r)}\right]^{\perp}$, whose columns determine the row space of (3.159). As $n$ becomes large, the number of columns in $\left[M^{(r)}\right]^{\perp}$ (exponential in $n$) dominates the number of columns in $M^{(r)}$ (polynomial in $n$). As $\left[M^{(r)}\right]^{\perp}$ is nearly a square matrix, it will have relatively sparse vectors in its column space. Consequently, we relax our problem and consider whether or not $\boldsymbol{c}$ covers any vector in the sparse-basis of the row space of

$$\left[ \left[ M^{(r)} \right]' \otimes I \right] \quad ,^{15} \tag{3.162}$$

which itself reduces to determining a sparse-basis for the column space of $M^{(r)}$. For each sparse vector in $M^{(r)}$, there are $\eta$ distinct sparse vectors in the row space of (3.162).

Each column of $M^{(r)}$ consists of 0s and 1s, and provides the marginal probability of $r$ random variables (automata) assuming particular values. We begin by considering the columns of $M^{(r)}$ as the sparse-basis for its column space, and then consider linear combinations of its columns to identify additional sparse vectors that should be in its sparse basis. Note that $M^{(r)}$ has $(mn)^r$ columns, each with a count of at least $m^{n-r}$ (this can be argued upon considering the Kronecker structure of $M^{(r)}$ deduced from (3.99)).

If there should exist sparse vectors in the column space of $M^{(r)}$ with a smaller count (fewer nonzero entries) than its own columns, this would suggest that by taking the proper linear combination of the entries of any partial information vector $\boldsymbol{\pi}^{(r)}$ (the sequence of $r$th-order marginals), once could obtain finer information about the joint distribution than that which is provided by the $r$th-order marginals (obviously *sometimes* finer information in the form of higher-order marginal distributions can be deduced from the $r$th-order marginals, such as when $\boldsymbol{\pi}^{(r)}$ is composed of indicator vectors, just not *always*). Although this argument is not rigorous, it suggests that all other sparse vectors in the column space of $M^{(r)}$ should have a count of at least $m^{n-r}$.

The hope is to define a sparse-basis for the column space of $M^{(r)}$ that contains a number of sparse vectors that is polynomial in $\eta = m^n$. The reasoning is as follows: if $f$ is the fraction of entries given entry-wise constraints, a particular entry is given an entry-wise constraint with probability $f$. Moreover, by repeated conditioning, we can derive the upper bound

$$\mathbb{P}\,(\, k \text{ specific entries are given entry-wise constraints}\,) \leq f^k \quad . \tag{3.163}$$

By applying the union bound to (3.161) and accepting the claim that $\|\boldsymbol{v}\| \leq m^{n-r}$ for all $\boldsymbol{v} \in \mathcal{V}$, we obtain

$$\begin{aligned} \mathbb{P}\,(\,\text{inconsistency}\,) &= \mathbb{P}\,(\,\cup_{\boldsymbol{v}\in\mathcal{V}}\{\,\mathbf{c} \text{ covers } \boldsymbol{v}\,\}\,) \\ &\leq \sum_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}\,(\,\mathbf{c} \text{ covers } \boldsymbol{v}\,) \\ &\leq |\mathcal{V}| f^{m^{n-r}} \quad . \end{aligned} \tag{3.164}$$

As long as the number of sparse vectors in the sparse-basis, i.e., $|\mathcal{V}|$, is polynomial in $\eta$, then (3.164) converges to 0 as $n$ increases. Although we have been unable to prove this, we conjecture that this is indeed the case.

### 3.6.2 Ensuring nonnegativity

The second concern, whether or not the $\epsilon \cdot \dim(\mathcal{G})$-dimensional affine subspace embodying both the model-imposed constraints and the linear constraints of $M_*$-sufficiency intersects

[15] We have not made a rigorous argument that we do indeed have a relaxation by considering (3.162) in place of (3.159). If one considers any sparse vector in the row space of (3.159), one should be convinced that it covers some sparse vector in the row space of (3.162).

the nonnegative orthant, is even less clear. As in the previous case when discussing the possibility of linear dependence when accumulating together all linear constraints, suppose that the affine subspace is' chosen uniformly at random.[16] In particular, let $\mathsf{S}_\epsilon \subset \mathbb{R}^{\eta\times\eta}$ be a subspace chosen uniformly at random of dimension $\epsilon \cdot \dim(\mathcal{G}) = \epsilon\eta(\eta-1)$. As discussed in Appendix C, $\mathsf{S}_\epsilon$ will almost surely intersect

$$\sum_{j=0}^{\epsilon\eta(\eta-1)} \binom{\eta^2}{j} \tag{3.165}$$

orthants, out of total of $2^{\eta^2}$ orthants. As there is no preference towards selecting particular orthants,

$$\begin{aligned} \mathbb{P}\left(\mathsf{S}_\epsilon \cap \mathbb{R}_+^{\eta\times\eta} \neq \varnothing\right) &= \frac{\sum_{j=0}^{\epsilon\eta(\eta-1)} \binom{\eta^2}{j}}{2^{\eta\times\eta}} \\ &= \sum_{j=0}^{\epsilon\eta(\eta-1)} \binom{\eta^2}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{\eta^2-j} \quad . \end{aligned} \tag{3.166}$$

Invoking results for the concentration of binomial probabilities, cf. (3.145) and (3.146), we conclude

$$\mathbb{P}\left(\mathsf{S}_\epsilon \cap \mathbb{R}_+^{\eta\times\eta} \neq \varnothing\right) \to \begin{cases} 1 & \text{if } \epsilon > 1/2 \\ 0 & \text{if } \epsilon < 1/2 \end{cases} \quad \text{as } \eta \to \infty \quad . \tag{3.167}$$

Hence, if we were to specify a fraction of $(d - \epsilon) < \frac{1}{2}$ of the entries of the transition matrix, and the intersection affine subspace was typical of random affine subspaces, then, as $n \to \infty$ with probability 1 there will exist a nonnegative transition matrix exhibiting $M_*$-separability and all of the model-imposed constraints.

However, the intersection affine subspace is likely not well characterized as a typical random affine subspace, at least in the cases of $M^{(r)}$-separability. In these cases, the two affine subspaces whose intersection is of interest exhibit several interesting properties. Recall that the affine subspace of matrices exhibiting $M^{(r)}$-separability includes the transition matrix with all of its entries equaling $\frac{1}{\eta}$, which is a point along the central ray of the nonnegative orthant in $\mathbb{R}^{\eta\times\eta}$. Moreover, the model-based constraints are always chosen so that there exist nonnegative matrices satisfying such constraints. Evidently both subspaces intersect the convex set of stochastic transition matrices, with one subspace intersecting its centroid. This suggests that it is rather possible that the intersection of these subspaces will be close to these two points in the set of stochastic transition matrices that each subspace is known to include, and thus more likely that the intersection passes through the nonnegative orthant. Naturally, the orientations of the two affine subspaces relative to one

[16] Choosing a *subspace* uniformly at random is similar to choosing linear constraints uniformly at random. The dimension of the subspace is fixed, and then its basis vectors are chosen uniformly over a hypersphere, i.e., with equal preference to all directions. Thus a random subspace is defined via a random matrix representing its basis vectors. An *affine subspace* chosen uniformly at random can be obtained from a subspace chosen uniformly at random by then shifting the subspace by an amount drawn from a rotationally symmetric probability density function (such as a multivariate Gaussian composed of IID Gaussian random variables). The idea is that there is no preference to any directions.

another determines how far away such an intersection may occur.

To motivate this idea—that the intersection of two affine subspaces known to intersect will often be close to a region where the affine subspaces are known to be close to one another—consider the simple example in $\mathbb{R}^2$, where two 1-dimensional affine subspaces (lines) are known to pass through points A and B, respectively, which are at a distance of $r$ from one another. If it is assumed that the angle corresponding to each line is chosen independently and uniformly at random, then by a simple geometric exercise, one can show that with probability $\frac{1}{2}$, the intersection will occur at a point that is at a distance of $r$ or less from point A. Naturally, this idea can be generalized to random affine subspaces of higher dimension, however, it appears as though the mathematics become substantially more complicated in higher dimensions.

### 3.6.3 Empirical results

Having discussed these issues at a theoretical level, we feel it is imperative to explore them computationally. To this end, we explore the following scenario. We randomly generate an $m^n \times m^n$ stochastic transition matrix; we will discuss shortly two sampling methods that were considered. Next, we select a fraction $f$ of the entries of the matrix uniformly at random. The values of these entries of the randomly chosen matrix constitute the model-based constraints. Then, we check to see if there exists a matrix satisfying the model-based constraints that is $M^{(r)}$-separable. Computationally this is checking the feasibility of a linear program, which can be evaluated for relatively small values of $n$. We note whether infeasibility is due to inconsistent linear constraints (which would almost surely be the case whenever $f$ exceeds the relative dimension of $M^{(r)}$-separability), or it is due to the constraints of nonnegativity. Naturally, we wish to see the behavior of the empirical probabilities for there existing an $M^{(r)}$-separable model satisfying the model-based constraints as $f$ and $n$ are varied (but only up to a certain threshold for which solving the linear program is tractable).

#### Randomly sampling stochastic matrices

We consider two different approaches to generating a random stochastic transition matrix. In both cases, each row of the randomly generated stochastic transition matrix is obtained independently, meaning that the problem reduces to generating random probability vectors. One approach is to sample the probability vectors along a hypersphere that can be inscribed in the simplex. Note that the maximum radius of a hypersphere that can be inscribed in the simplex for an alphabet of size $\eta$ is $\frac{1}{\sqrt{\eta(\eta-1)}}$.[17] The motivation for this approach is that the distance of a randomly generated matrix to the centroid of the polytope of transition matrices, the stochastic transition matrix with constant entries that is known to be $M^{(r)}$-separable, will be constant. This would ensure that the two affine subspaces (the matrices exhibiting $M^{(r)}$-separability, and the matrices satisfying the model-imposed constraints)

[17] One may be misled to believe that the maximum radius that can be inscribed in the simplex is $\frac{1}{\eta}$: the centroid probability vector $(1/\eta, 1/\eta, \dots, 1/\eta)$ can be disturbed by a vector of length $1/\eta$ with only a single nonzero entry and upon adding such a disturbance to the centroid probability vector, the resulting disturbed vector will have an element that equals 0. Although this is true, the only concern are disturbances within the affine subspace that contains the simplex, and the hence the disturbances of interest must have their entries sum to 0 (for the disturbed vector to lie in the affine subspace). The result as stated is easily derived by geometry or calculus; one finds that the intuitive disturbance vector of minimal length is $(1/\eta, -1/(\eta(\eta-1)), \dots, -1/(\eta(\eta-1)))$.

would pass through points of a known distance apart, akin to the 2-dimensional example discussed earlier. This random sampling can be accomplished by generating a vector of independent and identically distributed samples from a Gaussian distribution (which will have no preference for particular directions), transforming it so that it sums to 0,[18] scaling such a vector, and then adding it to the constant probability vector to generate a new probability vector that is a known distance from the constant probability vector. When $n$ was varied, we elected to impose a constant average squared deviation on each entry of the transition matrices as $n$ increased. This was after considering two other approaches that could be interpreted as being biased to larger values of $n$. One of these abandoned approaches was to sample probability vectors along the largest inscribed hypersphere for a given value of $n$. By following such an approach, as $n$ increased, the distance (as calculated in increasingly higher dimensions) of the randomly sampled matrices to the centroid would be decreasing as $\eta^{-1/2} = m^{-n/2}$. In a second approach that was also eventually abandoned, we considered maintaining a hypersphere of constant radius for each probability vector (row of the randomly generated matrix) in increasingly higher dimensions as $n$ increased. Although in this approach the distance of the randomly sampled matrices to the centroid was increasing as $\sqrt{\eta} = m^{n/2}$, we felt that such an approach could still be biasing our results to appear more favorable for large $n$, as such larger distances as $n$ increased were achieved via matrices with $\eta^2$ entries, and thus, the average squared deviation of a given entry in a matrix was decreasing as $\eta^{-1} = m^{-n}$ as $n$ increased. Hence we converged on the approach of maintaining constant average squared deviation in the entries of the randomly generated matrices as $n$ increased.

Alternatively we consider generating probability vectors uniformly at random over the simplex. This can be accomplished by drawing independent exponential random variables with parameter $\lambda = 1$ and then scaling [51]. This approach has the disadvantage of not maintaining constant distances of the randomly generated matrices to the centroid matrix. On the other hand, it has the advantage of sampling from all probability vectors, and not being limited to those along the hypersphere. Note that as the length of the probability vector, $\eta$, becomes large, the radius of the largest inscribed hypersphere, $\frac{1}{\sqrt{\eta(\eta-1)}}$, converges to 0, while the distance from the centroid to the furthest probability vector approaches 1. Furthermore, the ratio of the hypervolume of the largest inscribed hypersphere to the hypervolume of the simplex approaches 0 [52]. Thus, when sampling over an inscribed hypersphere, as the length of the probability vector increases, the distance from the centroid of the sampled probability vectors decreases, and the proportion of probability vectors of such a distance or less converges to 0. Now we proceed to some results.

Figures 3-26 and 3-27 offer a comparison of the two sampling methods. In both figures, only the $n = 5$, $m = 2$ case of $M^{(1)}$-separability is considered. The $x$-axis variable in the figures is $f$, the fraction of entries that are constrained. As computed in Table 3.1, the relative dimension of $M^{(1)}$-separability, for $n = 5$ and $m = 2$ is 0.87. We know that for

[18] Note that we require an $\eta \times 1$ random vector that has no preference for any particular directions within the subspace of vectors that sum to 0. To generate such a vector, we generate an $(\eta - 1) \times 1$ within the subspace's coordinate system that has no preference for individual directions, i.e., an $(\eta - 1) \times 1$ vector of IID Gaussian random variables. Next, the vector is augmented with a 0 to be represented in a coordinate system in $\eta$-dimensions (with the last coordinate being the normal to the subspace). Note that such a vector can still be thought of as a degenerate multivariate normal consisting of independent random variables. When the multivariate normal is transformed to Cartesian coordinates, because of the properties of linear transformations of multivariate normal random variables [50], the transformed random vector in Cartesian coordinates that sums to 0 is still multivariate Gaussian.

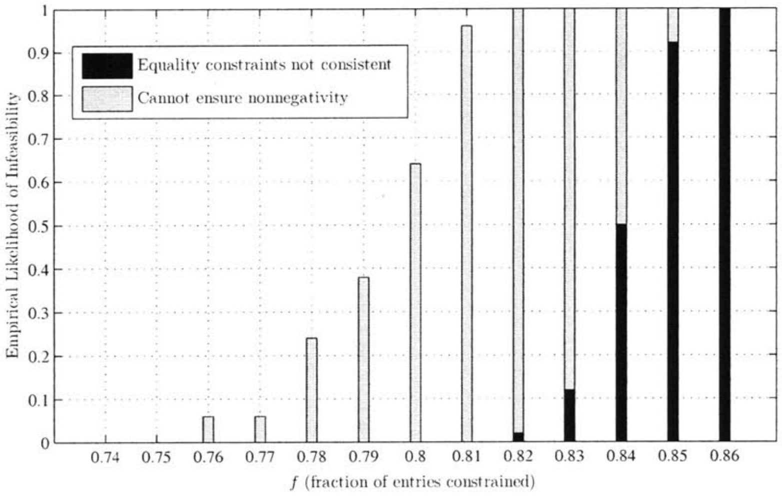


Figure 3-26: Empirical infeasibility of matching with an $M^{(1)}$-separable model a fraction $f$ of the entries of a random stochastic matrix whose rows are sampled over the largest hypersphere inscribed in the simplex (50 trials for each value of $f$, $n$=5, $m = 2$).

$f > 0.87$, with probability 1 we will have infeasibility due to inconsistency. For $f < 0.87$, we expected that when randomly generating matrices with rows on the largest inscribed hypersphere, i.e., a fixed distance to the centroid (a matrix known to exhibit $M^{(1)}$-separability), that we would be able to match constraints with higher values of $f$ than in the case when matrices were uniformly sampled. Indeed this hypothesis was corroborated empirically. In Fig. 3-26, when the sampling is done at a fixed distance, for both $f = 0.74$ and $f = 0.75$, in 100% of the trials (50/50) the random entry-wise constraints could be satisfied with an $M^{(1)}$-separable model. As $f$ increases towards 0.87, the probability of infeasibility empirically grows, initially dominated by infeasibility due to nonnegativity, and eventually infeasibility is wholly accounted for by inconsistency of the constraints. In Fig. 3-27, when the sampling is performed uniformly, infeasibility due to an inability to guarantee nonnegativity is small at $f = 0.56$, but at $f = 0.64$, all 20 trials resulted in infeasibility due to nonnegativity. Evidently without any guarantees that the affine subspace corresponding to the model-imposed constraints would be some minimum distance to a matrix satisfying $M^{(1)}$-separability, a nonnegative solution is less likely to be assured.

What is clear is that our intersection affine subspace is, as expected, more likely to intersect the nonnegative orthant than a random affine subspace of the same relative dimension: in the case when matrices are sampled uniformly, with $f = 0.56$, which would correspond to the intersection affine subspace having a relative dimension of 0.33 (and a relative dimension of 0.32 relative $\mathbb{R}^{32\times 32}$) with high empirical probability it intersects the nonnegative orthant (meaning a feasible solution exists). If our intersection affine subspace was a random affine subspace, it would require a relative dimension of 0.5 relative to $\mathbb{R}^{32\times 32}$ to intersect the nonnegative orthant with high probability.

Figures 3-28 and 3-29 illustrate how the empirical probability of infeasibility changes for increasing values of $n$, with $f$ fixed, $m = 2$, and $r = 1$, for each of our two sampling methods. In both cases, there were 20 trials for each value of $n$. As given in Table 3.1 and analyzed in Section 3.5, the relative dimension of $M^{(1)}$-separability is increasing towards 1 as $n$ increases.

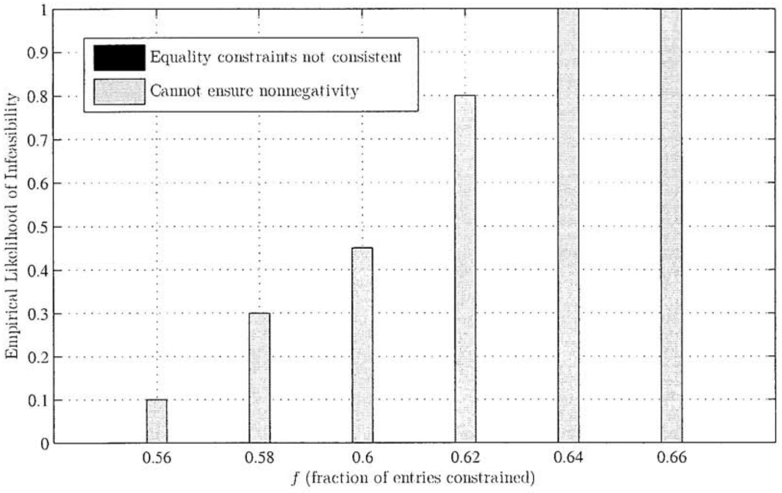


Figure 3-27: Empirical infeasibility of matching with an $M^{(1)}$-separable model a fraction $f$ of the entries of a random stochastic matrix whose rows are sampled uniformly over the simplex (20 trials for each value of $f$, $n$=5, $m = 2$).

One would expect that as the relative dimension increases, matching a fixed percentage of constraints ($f$) becomes more likely as $n$ increases. Empirical results corroborate this hypothesis. As shown in Fig. 3-28, for $f = 0.8$, infeasibility due to consistency in the linear constraints occurs with high empirical probability for $n = 3$ and $n = 4$ (the relative dimensions in these cases are $d = 0.79$ and $d = 0.82$). Rather surprisingly, for $n = 3$ it *was* possible to achieve feasibility in 3/20 trials. Note that because it was not possible to constrain exactly 0.8 of the entries, for $n = 3$, we constrained a number entries equal to the largest integer yielding a faction less than 0.8, which happened to be exactly equal the dimension of the set of matrices exhibiting $M^{(1)}$-separability. For $n = 5$ (relative dimension 0.87) and $n = 6$ (relative dimension 0.92), 20/20 trials were feasible when matching a fraction of 0.8 of the entries of a random matrix. Note that for each trial, we generate a random stochastic matrix with each row independently and uniformly sampled from a hypersphere inscribed in the simplex. As $n$ increases, we maintain a constant average square error for each entry (with the average over the entries in a given row). We select this average entry error to be such that for $n = 6$, we sample each row of our random matrix from from the largest hypersphere that can be inscribed in the simplex. Fig. 3-29 tells a similar story, except we illustrate the empirical likelihood of infeasibility when trying to match a fraction of $f = 0.63$ random entry-wise constraints. For the particular $f$ illustrated, empirically we have feasibility with high probability when $n = 6$; for $n < 6$ feasibility is unlikely.

These results suggest that the increased relative dimension of $M^{(r)}$-separability as $n$ increases, at least in the $n = 1$ case, does permit one to match an increasingly large percentage of entry-wise constraints. In this relative sense, $M^{(r)}$-separability becomes more general as $n$ increases.

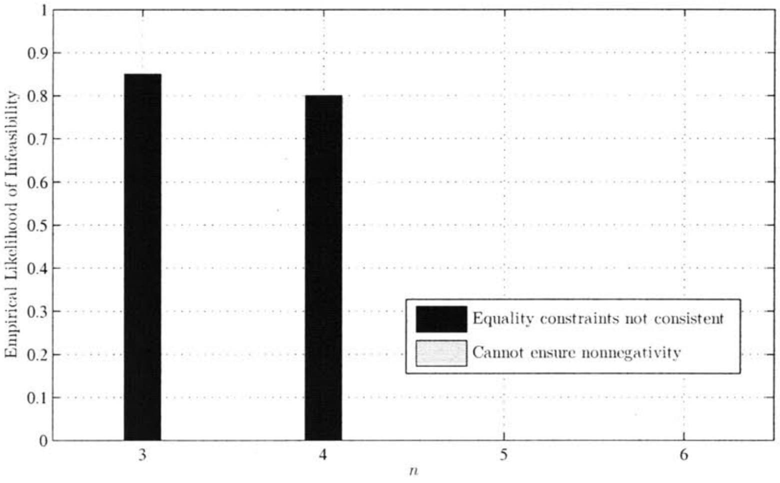


Figure 3-28: Empirical infeasibility as $n$ varies of matching with an $M^{(1)}$-separable model a fraction $f = 0.8$ of the entries of a random stochastic matrix whose rows are sampled over a hypersphere inscribed in the simplex (20 trials for each value of $f$; $m = 2$).

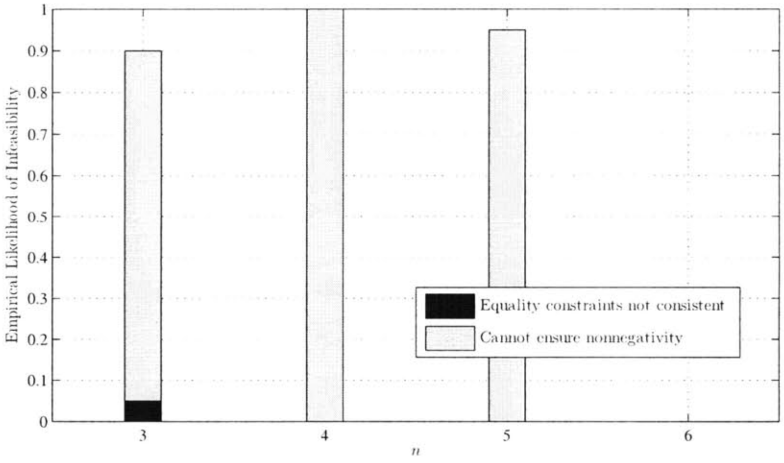


Figure 3-29: Empirical infeasibility as $n$ varies of matching with an $M^{(1)}$-separable model a fraction $f = 0.63$ of the entries of a random stochastic matrix whose rows are sampled uniformly over the simplex (20 trials for each value of $f$; $m = 2$).

## 3.7 Conclusions

This chapter completes a thorough analysis of the properties of separable models. It illustrated how the notion of separability offers an alternate perspective to invariance in linear systems. We showed the algebraic relationships between the eigenvalues and eigenvectors of the matrix that propagates partial information, $H_*$, and the underlying Markov chain's transition matrix $G$. In Section 3.3, we offered an intuitive visualization of the constraints imposed by separability. The subsequent sections were focused on determining exact equations for the rank of $M^{(r)}$, to then analyze the evolving dimension (and generality, in a sense) of $M^{(r)}$- and $\mathcal{M}^{(r)}$-separability. In Section 3.5, we illustrated many interesting results about the evolving dimension of separability for varying $r$ and $n$. The final section illustrated what high relative dimension could offer for our canonical examples of separability, which could prove useful in an engineering or design application.

# Chapter 4

# The Generalized Influence Model

We have introduced separability and illustrated its many algebraic and geometric properties. The constraints that must be satisfied to ensure separability are clear. For the canonical examples of separability, the evolution of such constraints as the model parameters vary—$n$, the number of random variables/automata, and $r$, the order of the propagating marginals—has been characterized and the potential generality of modeling with separability has been illustrated. What is missing are meaningful parametric models exhibiting separability. Recall the block triangular form first appearing in (3.3), and note that it essentially offers a parametrization:

$$G \xrightarrow{\text{lin. trans.}} \begin{bmatrix} H_* & * \\ 0 & * \end{bmatrix} \quad . \tag{4.1}$$

Although we have given meaning to $H_*$, the other block matrices composing such a parametrization have been ignored. There is not much to meaning to extract from such a parametrization, and more importantly, it will not be tractable. Under any form of separability, there will be a constant fraction of the number of free parameters as in the unrestricted case, meaning the number of parameters will be exponential in $n$. This fact follows upon realizing that the block of 0s in (4.1) cannot dominate the $\eta \times \eta$ matrix; the sum of its rows plus columns must equal $\eta$.

The hope would be to specify a tractable parametrization that exhibits separability. Such parametric models would allow one to generate synthetic data for separable models—something that is not possible by propagating partial information via $H_*$. In order to be tractable, such parametric models must introduce additional assumptions.

For the case of $M^{(1)}$-separability, which propagates the univariate marginals, one's first thought may be to define a parametric model that updates as follows: after propagating a sequence of indicator vectors $\mathbf{s}^{(1)}$ via the $H^{(1)}$ matrix, one receives a sequence of univariate PMFs, and to generate a new sample, i.e., a new sequence of indicator vectors, one independently realizes each of these PMFs. Indeed, what results is an $M^{(1)}$-separable model that is fully parameterized by $H^{(1)}$, i.e., a representation that requires $\Theta(n^2)$ storage. By assuming independent updates for each random variable, we obtained a tractable representation. This model is precisely the influence model (IM) introduced in [12, 13]. Although the IM may illustrate many interesting behaviors (see the cited references for several examples), it has its limitations. There can be no coordination among the random variables in how they update. This chapter introduces the generalized influence model (GIM), which introduces conditionally dependent updates to the parametric structure of the influence model. Like the IM, the GIM is $\mathcal{M}^{(n)}$-separable. It is a scalable parameterization; a GIM may defined in as few as $\Theta(n^2)$ parameters; however more complex GIMs may require up to the number of free parameters under $\mathcal{M}^{(n)}$-separability, $\Theta(c^n)$.

## 4.1 Definition

The GIM will be introduced as satisfying a particular form for the probabilistic update of its network state, paralleling the approach used for the IM in [12, 13]. We shall later show that the GIM can be defined in several other equivalent ways.

A GIM is a time-homogeneous, jointly Markovian network of stochastic automata that updates from time $t-1$ to $t$ as follows:

> At time $t$, automaton $j$ (for all $j = 1, 2, \dots n$) probabilistically chooses an automaton (possibly itself) as its *influencer*, denoted by $\mathsf{i}_j[t]$ (but we shall simply write $\mathsf{i}_j$ when the time $t$ is understood). Automaton $j$ then uses the influencer's previous status indicator vector $\mathbf{s}_{\mathsf{i}_j}[t-1]$ to update itself via a probabilistically chosen *selector matrix* $\mathsf{A}_j[t]$ (but we again simply write $\mathsf{A}_j$ when the time $t$ is understood):
>
> $$\mathbf{s}_j[t]' = \mathbf{s}_{\mathsf{i}_j}[t-1]'\mathsf{A}_j \, . \tag{4.2}$$
>
> The choices of $\mathsf{i}_j$ and $\mathsf{A}_j$ by automaton $j$ are made independently of all past updates, but *not necessarily independently* of the current choices of the other automata (and this is the key distinction from the IM).

The random $m_j \times m_{\mathsf{i}_j}$ selector matrix $\mathsf{A}_j$ is a row-stochastic matrix consisting of only 0's and 1's—which assures that the current status indicator vector for automaton $j$ using (4.2) is indeed again an indicator vector. The random variables $\{\mathsf{i}_j\}$ and random matrices $\{\mathsf{A}_j\}$ are governed by a time-invariant joint probability distribution that characterizes the GIM and fixes its joint transition matrix $G$. Note that when we write $\mathbf{s}_{\mathsf{i}_j}$, we do not mean the indicator vector for the random variable $\mathsf{i}_j$; rather we mean the indicator random variable for $\mathsf{x}_{\mathsf{i}_j}$. It is as though we are employing the shorthand notation $\mathbf{s}_i$ for $\mathsf{i}_j = i$. As most indicator vector notation in this chapter is of the shorthand form, subsequent instances should be clear.

### 4.1.1 $M^{(1)}$-separability from the perspective of Pfeffer

Because of the GIM's particular update structure, it exhibits $\mathcal{M}^{(n)}$-separability. This will be illustrated in Section 4.3. By being $\mathcal{M}^{(n)}$-separable, any GIM must be $M^{(1)}$-separable, and as noted in Section 2.2.1, $M^{(1)}$-separability can be equivalently characterized as local instances of Pfeffer's separability (without any assumptions of a Bayesian network). To describe $M^{(1)}$-separability from the perspective of Pfeffer's separability, we must consider the equation for the probabilistic update of each random variable (automaton) of the network:

$$\mathbb{E}\left[\mathbf{s}_j[t]' \,|\, \mathbf{s}_{\mathsf{x}}[t-1]\right] = \mathbf{s}_{\mathsf{x}}[t-1]'GM_j \, , \tag{4.3}$$

which is obtained by marginalizing the Markov equation (2.5) , i.e., multiplying by $M_j$, to obtain the conditional PMF for the update of automaton $j$.

The conditional PMF in (4.3) is considered separable according to Pfeffer [30] when it can be written as a sum of structured terms, each involving only a single status indicator vector:

$$\mathbb{E}\left[\,\mathsf{s}_j[t]' \,|\, \mathsf{s}_{\mathsf{x}}[t-1]\,\right] = \sum_{k=1}^{n} d_{jk}\left(\mathsf{s}_k[t-1]'A_{jk}\right) \ , \tag{4.4}$$

for some nonnegative scalars $\{d_{jk}\}$, $\sum_k d_{jk} = 1$, and row-stochastic matrices $\{A_{jk}\}$. Note that the state indicator vector on the right of (4.3) has been replaced in the separable form of (4.4) by its constituent status indicator vectors. As will be demonstrated in Section 4.3.1, the $d_{jk}$ and $A_{jk}$ of (4.4) have meaning related to the probabilistic GIM update given in (4.2): $d_{jk}$ is the probability that automaton $j$ selects automaton $k$ as its influencer, and $A_{jk}$ is the conditional expected value of $\mathsf{A}_j$, given that automaton $j$ selects automaton $k$ as its influencer. The *influence matrix* $D$ is defined as the $n \times n$ matrix whose $(j,k)$ entry is $d_{jk}$ and whose transpose is the weighted adjacency matrix of an *influence network graph*, $\Gamma(D')$. Each node in the influence network graph corresponds to an automaton, and for each node $j$ in $\Gamma(D')$ (which is associated with automaton $j$), the weight of an arc terminating at the node is the probability that this automaton $j$ will.pick as influencer the automaton at which the arc originates.

The individual separability conditions in (4.4) for each automaton $j$ can be assembled into the matrix equation

$$\mathbb{E}\left[\mathsf{s}^{(1)}[t]' \,|\, \mathsf{s}_{\mathsf{x}}[t-1]\right] = \mathsf{s}^{(1)}[t-1]'H^{(1)} \ , \tag{4.5}$$

where the matrix $H$ can be expressed in block form in terms of the $d_{jk}$ and $A_{jk}$ as follows:

$$H^{(1)} = \begin{bmatrix} d_{11}A_{11} & d_{21}A_{21} & \dots & d_{n1}A_{n1} \\ d_{12}A_{12} & d_{22}A_{22} & \dots & d_{n2}A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1n}A_{1n} & d_{2n}A_{2n} & \dots & d_{nn}A_{nn} \end{bmatrix} . \tag{4.6}$$

Note that the sparsity structure of the influence network graph impacts the sparsity of the matrix $H^{(1)}$. By taking an expectation over $\mathsf{s}_{\mathsf{x}}[t-1]$ and invoking Markovianity, we obtain our definition of $M^{(1)}$-separability (2.28).

### 4.1.2 GIM relationship to IM

One can think of an IM as a GIM with the following additional constraints at each time $t$:

- the influencers $\mathsf{i}_1$, $\mathsf{i}_2, \dots \mathsf{i}_n$ are mutually independent;

- the selector matrices $\mathsf{A}_1$, $\mathsf{A}_2 \dots \mathsf{A}_n$ are conditionally independent, given the influencers.

These constraints ensure that the updated statuses of all automata are conditionally independent, given the current network state, i.e., the automata independently update. Such a conditional independence property holds if and only if the conditional probability vector for the updated network state, given the current network state, has the following Kronecker factorable form:

$$\begin{aligned}\mathbb{E}\left[\, \mathsf{s}_{\mathsf{x}}[t]' \,|\, \mathsf{s}_{\mathsf{x}}[t-1]\,\right] &= \mathbb{E}\left[\bigotimes_{j=1}^{n} \mathsf{s}_j[t]' \,|\, \mathsf{s}_{\mathsf{x}}[t-1]\right] \\ &= \bigotimes_{j=1}^{n} \mathbb{E}\left[\, \mathsf{s}_j[t]' \,|\, \mathsf{s}_{\mathsf{x}}[t-1]\,\right] \ , \qquad (4.7)\end{aligned}$$

where the first equality follows from the definition of the state indicator vector as given in Section 2.1.3, and the second equality from the conditional independence assumption.

In (2.5) we used the joint transition matrix $G$ to represent this conditional PMF for the updated network state given its current state. Consequently, under the conditional independence assumption,

$$\bigotimes_{j=1}^{n} \mathbb{E}\left[\, \mathsf{s}_j[t]' \,|\, \mathsf{s}_{\mathsf{x}}[t-1]\,\right] = \mathsf{s}_{\mathsf{x}}[t-1]'G \ . \qquad (4.8)$$

The term on the right side of (4.8) is simply the row of $G$ selected by the indicator vector $\mathsf{s}_{\mathsf{x}}[t-1]$, so (4.8) shows that each row of $G$ in this case must have the Kronecker factorable form of the term on the left. More specifically, combining (4.8) with (4.3) shows that the entries of $G$ for an IM satisfy polynomial equalities. This conditional independence property satisfied by any IM thus imposes a significant restriction on the form of the joint transition matrix $G$. The GIM relaxes this conditional independence property of the IM, thereby permitting more complex behavior.

## 4.2 Some GIM examples

We now present three examples of a GIM. The first illustrates the richer behavior possible by modeling the weather in four cities as a GIM instead of an IM. The second example shows how a collection of gamblers betting at the same roulette table under some general conditions can be modeled as a GIM. This example is an instance of a special class of GIMs that we refer to a *GIMs of coupled Markov chains.* A third example is shuffling a deck of cards, an instance of a Markov chain over permutations, which we graciously borrow from [9]. This special class of GIMs is referred to as *permutation GIMs.* Markov chains over permutations have been investigated at length in both [9] and [10], from the perspective of representation theory for the symmetric group. More on this connection will be developed in Section 4.5. Two additional examples will be presented in Section 4.6 after introducing some further properties of the GIM.

### 4.2.1 Weather in four cities

We return to the example of Chapter 1 of the weather in four cities. Over each 24 hour period, four cities in relatively close proximity experience weather interactions. Suppose (for the purposes of our example) that each day one of possibly hundreds of weather patterns randomly comes into being, independently of all previous weather patterns. The weather pattern specifies for each city an upwind city (possibly itself) as its influencer, and specifies the city's future weather as a function of the influencer city's current weather. Fig. 4-1

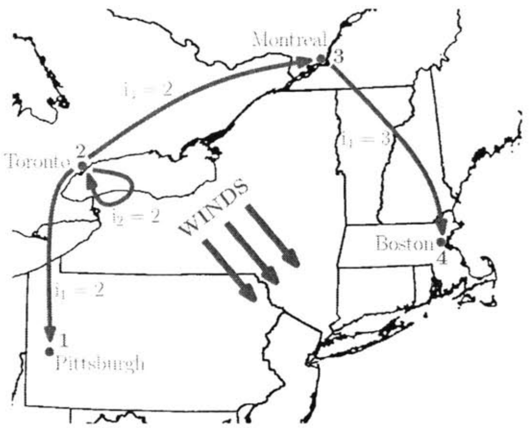


Figure 4-1: Example of a weather pattern and cities' influencers.

illustrates a possible weather pattern, and its corresponding pattern of influencers, with the arrows pointing from the influencers.

This weather example illustrates the interesting global behavior that is possible in a GIM but not in an IM. If modeled as an IM, the influencers could not be chosen in a coordinated way—they would have to be chosen independently for each city. Because the GIM permits automata to update themselves in a dependent fashion, cities may choose their influencers in a coordinated way, and may also collectively warm or cool in relation to their influencers' weather.

The weather dynamics in this example illustrate a necessary element of any GIM: that new information received at an automaton from an influencer (the influencer's current weather) overrides any other current information (e.g., the city's own current weather). As each automaton can only choose a single influencer in any realization of the stochastic process's update, a GIM prohibits combining information from multiple sources when an automaton updates. The GIM only exhibits a combining of information in an averaged or probabilistic sense, not realization by realization.

The example also highlights another necessary aspect of a GIM: the presence of an independent and identically distributed (IID) master process, the sequence of weather patterns. This process defines the pattern of influencers as well as the evolution of each automaton's status, given its influencer's current status. A realization of this IID process along with an initial network state defines a realization of a GIM.

### 4.2.2 Roulette table

We use a GIM now to model the winnings (or losses) of several gamblers at the same roulette table, under a particular set of restrictions. In this example, each automaton corresponds to a player, and each automaton's status is the total value of that gambler's chips. The set of possible statuses for each player is assumed finite. The GIM iterates upon each spin of the roulette wheel.

Our model assumes that before each spin of the wheel, the gamblers agree upon a collective but randomly chosen gambling policy. The chosen policy is independent and identically distributed at each spin, and independent of all past outcomes of the roulette wheel. The gambling policy determines each gambler's bets as a function of his or her chip total.

For example, Mike, Mark and Jon are all gambling together at the same roulette table. Under one gambling policy, they all bet on black, and each bets half of his chip total, not exceeding \$100. Under another gambling policy, Mike bets \$10 on on black (or his chip total of it is less than \$10), Mark bets half of his chips on 00, and Jon plays only a \$1 on $4, 5, 7, 8$. Other joint behaviors correspond to other gambling policies, each of which occurs with some known probability. It is the chosen gambling policies together with the spins of the roulette wheel that constitute the IID master process of this GIM. An outcome of this process at a given time determines the influencers and selector matrices of an iteration.

An interesting aspect of this example is that each gambler always chooses itself as its influencer, and consequently the time evolution of a single gambler's chip total is itself a time-homogeneous Markov chain. However, these Markov chains are coupled—two gamblers who almost always place similar bets will typically win and lose in tandem. A GIM in which each automaton always chooses itself as its influencer is a natural structure for modeling a set of coupled, time-homogeneous Markov chains that transition as a function of a common IID process. As noted in the introduction of this section, the class of GIMs where the automata always choose themselves as influencers is referred to as a GIM of coupled Markov chains.

Besides being Markovian, each automaton's updated status given its current status is always conditionally independent of all other automata's current and past statuses. This additional property ensures that we have $M_j$-separability for all automaton $j$. Recall from Section 2.2.4 that a network satisfying such instances of separability is referred to as $\mathcal{M}_n$-separable.

### 4.2.3 Deck of cards

A third example of a GIM is a standard deck of 52 cards being repeatedly shuffled. The state of the deck is the particular permutation of the cards in the deck. However, in our context of networks of stochastic automata, we wish to represent the state of the deck of cards as a set of random variables. We associate a random variable with each position in the deck, i.e., a random variable is associated with the first position, a second random variable with the second position, and so on. The random variable (automaton) indicates the current card in the respective position. If the top card in the deck is the seven of diamonds, then the first automaton exhibits the status corresponding to the seven of diamonds. We can think of repeatedly shuffling a deck of cards as a discrete-time stochastic process on these 52 random variables, each assuming values in an alphabet of size 52.

Each shuffle randomly changes the state of the deck, and as long as the act of shuffling each time is an IID process, we have a Markov chain. This Markov chain is also a GIM with special structure. The realization of the IID process at a given time specifies a particular shuffle. This shuffle determines an influencer for each random variable. If position $j$ chooses position $i$ as its influencer, and position $i$ at time $t-1$ indicates the queen of spades, then the updated status of position $j$ at time $t$ indicates the particular card in position $i$ at time $t-1$, i.e., the queen of spades. The set of influencers is always a permutation. Note that for this example, the $\mathsf{A}_j$ matrices are identities, as each position receives (copies) the card of its influencer. In Section 4.5, the algebraic structure of permutation GIMs is developed, and its connection to representation theory for the symmetric group is explained.

## 4.3 Illustrating the GIM's $\mathcal{M}^{(n)}$-separability

The tractability of the GIM is twofold: it exhibits $\mathcal{M}^{(n)}$-separability, while potentially admitting a tractable representation. Here we demonstrate its separability; we will discuss in Section 4.4 the potential complexity of GIM representations.

By nature of being $\mathcal{M}^{(n)}$-separable, the $\binom{n}{r}$ $r$th-order marginal PMFs can be propagated linearly, for each $r$, $1 \leq r < n$. We illustrate this fact for general $r$ in what follows.

### 4.3.1 Linear propagation of $\boldsymbol{\pi}^{(r)}[t]$

Consider any block of the current 2nd-order state array vector, which we will denote by $\mathsf{s}_{\boldsymbol{j}}[t]$, with $\boldsymbol{j} = (j_1, j_2, \ldots, j_r)$. By (4.2) and the mixed-product property of Kronecker products (A.3), we can express this current 2nd-order status indicator vector as

$$\begin{aligned}
\mathsf{s}_{\boldsymbol{j}}[t]' &= \mathsf{s}_{j_1}[t]' \otimes \mathsf{s}_{j_2}[t]' \otimes \ldots \otimes \mathsf{s}_{j_r}[t]' \\
&= \left(\mathsf{s}_{\mathsf{i}_{j_1}}[t-1]'\mathsf{A}_{j_1}\right) \otimes \left(\mathsf{s}_{\mathsf{i}_{j_2}}[t-1]'\mathsf{A}_{j_2}\right) \otimes \ldots \otimes \left(\mathsf{s}_{\mathsf{i}_{j_r}}[t-1]'\mathsf{A}_{j_r}\right) \\
&= \left(\mathsf{s}_{\mathsf{i}_{j_1}}[t-1]' \otimes \mathsf{s}_{\mathsf{i}_{j_2}}[t-1]' \otimes \ldots \otimes \mathsf{s}_{\mathsf{i}_{j_r}}[t-1]'\right) \left(\mathsf{A}_{j_1} \otimes \mathsf{A}_{j_2} \otimes \ldots \otimes \mathsf{A}_{j_r}\right) \\
&= \mathsf{s}_{\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}}[t-1]'\mathsf{A}_{\boldsymbol{j}} \; ,
\end{aligned} \tag{4.9}$$

where we have further extended our vector subscript notation by writing

$$\boldsymbol{\mathsf{i}}_{\boldsymbol{j}} \triangleq \boldsymbol{\mathsf{i}}_{(j_1, j_2, \ldots, j_r)} = (\mathsf{i}_{j_1}, \mathsf{i}_{j_2}, \ldots, \mathsf{i}_{j_r}) \; , \tag{4.10}$$

$$\mathsf{A}_{\boldsymbol{j}} \triangleq \mathsf{A}_{(j_1, j_2, \ldots, j_r)} = \mathsf{A}_{j_1} \otimes \mathsf{A}_{j_2} \otimes \ldots \otimes \mathsf{A}_{j_r} \; . \tag{4.11}$$

Note that $\mathsf{A}_{\boldsymbol{j}}$ is a random selector matrix, as it is a Kronecker product of random selector matrices, and that $\mathsf{s}_{\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}}[t-1]$ is a block of the $r$th-order state array vector at time $t-1$. We now observe that (4.9) has the same form as the random update equation (4.2), with the scalar subscripts $j$ and $\mathsf{i}_j$ replaced by the vector subscripts $\boldsymbol{j}$ and $\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}$.

Conditioned both on the previous network state, represented by $\mathsf{s}^{(r)}[t-1]$, and on having automata $\boldsymbol{j}$ choosing automata $\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}$ as their respective influencers, we take the expected value of (4.9) to obtain

$$\begin{aligned}
\mathbb{E}\left[\mathsf{s}_{\boldsymbol{j}}[t]' \,\middle|\, \mathsf{s}^{(r)}[t-1], \boldsymbol{\mathsf{i}}_{\boldsymbol{j}}\right] &= \mathbb{E}\left[\mathsf{s}_{\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}}[t-1]'\mathsf{A}_{\boldsymbol{j}} \,\middle|\, \mathsf{s}^{(r)}[t-1], \boldsymbol{\mathsf{i}}_{\boldsymbol{j}}\right] \\
&= \mathsf{s}_{\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}}[t-1]'\mathbb{E}\left[\mathsf{A}_{\boldsymbol{j}} \,\middle|\, \mathsf{s}^{(r)}[t-1], \boldsymbol{\mathsf{i}}_{\boldsymbol{j}}\right] \\
&= \mathsf{s}_{\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}}[t-1]'\mathbb{E}\left[\mathsf{A}_{\boldsymbol{j}} \mid \boldsymbol{\mathsf{i}}_{\boldsymbol{j}}\right] \\
&= \mathsf{s}_{\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}}[t-1]'A_{\boldsymbol{j}\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}} \; ,
\end{aligned} \tag{4.12}$$

where $A_{\boldsymbol{j}\boldsymbol{\mathsf{i}}_{\boldsymbol{j}}}$ denotes the row-stochastic matrix $\mathbb{E}\left[\mathsf{A}_{\boldsymbol{j}} \mid \boldsymbol{\mathsf{i}}_{\boldsymbol{j}}\right]$. Note that we have not expressed the conditional expectation of $\mathsf{s}_{\boldsymbol{j}}[t]$ as $\boldsymbol{\pi}_{\boldsymbol{j}}[t]$, so as to maintain clarity in the conditioning.

By taking an expectation with respect to $\mathbf{i}_j$ in (4.12), we find

$$\begin{aligned}\mathbb{E}\left[\mathbf{s}_j[t]' \,\middle|\, \mathbf{s}^{(r)}[t-1]\right] &= \sum_{\boldsymbol{i}} \mathbf{s}_{\boldsymbol{i}}[t-1]' A_{j\boldsymbol{i}} \cdot \mathbb{P}\left(\mathbf{i}_j = \boldsymbol{i} \,\middle|\, \mathbf{s}^{(r)}[t-1]\right) \\ &= \sum_{\boldsymbol{i}} \mathbf{s}_{\boldsymbol{i}}[t-1]' A_{j\boldsymbol{i}} \mathbb{P}\left(\mathbf{i}_j = \boldsymbol{i}\right) \\ &= \sum_{\boldsymbol{i}} \mathbf{s}_{\boldsymbol{i}}[t-1]' A_{j\boldsymbol{i}} d_{j\boldsymbol{i}} \,. \end{aligned} \tag{4.13}$$

where $d_{j\boldsymbol{i}}$ denotes $\mathbb{P}\left(\mathbf{i}_j = \boldsymbol{i}\right)$. Note that as $\mathbf{s}_j[t]$ is an $r$th-order status indicator vector, it is a block of the current $r$th-order state array vector. Because (4.13) holds for all $\boldsymbol{j}$, the expected value of the $r$th-order state array vector conditioned on the previous network state can be expressed as follows:

$$\mathbb{E}\left[\mathbf{s}^{(r)}[t]' \,\middle|\, \mathbf{s}^{(r)}[t-1]\right] = \mathbf{s}^{(r)}[t-1]' H^{(r)} \,, \tag{4.14}$$

for all times $t$, with

$$H^{(r)} = \begin{bmatrix} d_{\boldsymbol{k}_1\boldsymbol{k}_1} A_{\boldsymbol{k}_1\boldsymbol{k}_1} & \cdots & d_{\boldsymbol{k}_{n^r}\boldsymbol{k}_1} A_{\boldsymbol{k}_{n^r}\boldsymbol{k}_1} \\ d_{\boldsymbol{k}_1\boldsymbol{k}_2} A_{\boldsymbol{k}_1\boldsymbol{k}_2} & \cdots & d_{\boldsymbol{k}_{n^r}\boldsymbol{k}_2} A_{\boldsymbol{k}_{n^r}\boldsymbol{k}_2} \\ \vdots & \ddots & \vdots \\ d_{\boldsymbol{k}_1\boldsymbol{k}_{n^r}} A_{\boldsymbol{k}_1\boldsymbol{k}_{n^r}} & \cdots & d_{\boldsymbol{k}_{n^r}\boldsymbol{k}_{n^r}} A_{\boldsymbol{k}_{n^r}\boldsymbol{k}_{n^r}} \end{bmatrix} \,, \tag{4.15}$$

where the $\{\boldsymbol{k}_l\}$ are ordered lexicographically, so $\boldsymbol{k}_1 = (1, 1, \ldots, 1)$, $\boldsymbol{k}_2 = (1, 1, \ldots, 1, 2)$, etc.

From (4.14) one can show, as a consequence of joint Markovianity of the automata that

$$\mathbb{E}\left[\mathbf{s}^{(r)}[t]' \,\middle|\, \mathbf{s}^{(r)}[\tau]\right] = \mathbf{s}^{(r)}[\tau]' \left[H^{(r)}\right]^{t-\tau} \,, \tag{4.16}$$

for all integer times $t > \tau$. By taking a final expectation with respect to $\mathbf{s}^{(r)}[\tau]$, we have

$$\boldsymbol{\pi}^{(r)}[t]' = \boldsymbol{\pi}^{(r)}[\tau]' \left[H^{(r)}\right]^{t-\tau} \,, \tag{4.17}$$

which holds for any initial condition $\boldsymbol{\pi}_{\mathbf{x}}[\tau]$.

In (4.14), (4.16), and (4.17), we have demonstrated $M^{(r)}$-separability according to Corollary 1, Corollary 5, and Corollary 4, respectively. By exhibiting $\mathcal{M}^{(n)}$-separability, the GIM offers a scalable framework that allows one to analyze the stochastic network by choosing a level $r$ that is sufficient for one's needs.

## 4.4 Geometric perspective on IM, GIM, and $\mathcal{M}^{(n)}$-separability

When the IM was originally introduced [12, 13], it was defined as following a particular parametric update. Similarly, the GIM is defined as updating itself in a parametric way. These *parametric* probabilistic models stand in contrast to $M_*$-separability and $\mathcal{M}$-separability, which are probabilistic models that satisfy some particular constraints (e.g., some marginal information can be propagated). As would be of immediate interest to one with an algebraic perspective [53], and especially to those interested in algebraic statistical models [54, 55], we would like to understand the relationships between our parametric probabilistic models and our *constraint-based* probabilistic models. As Markovianity of the network state is always assumed, our algebraic statistical models are fully-specified by the transition matrix for the network state (technically there is also an initial probability vector, but we will ignore it as it is irrelevant). The primary tool that we will use to analyze these relationships is geometry, and in particular, the geometry of the sets of transition matrices that compose these classes of probabilistic models. As the case in Chapter 3, we will always fix our model size when developing the relationships among the parametric and constraint-based probabilistic models, and by model size we mean a fixed network configuration, i.e., a fixed number of $n$ automata, with the $j$th automaton having a fixed number of $m_j$ possible statuses, $1 \leq j \leq n$.

### 4.4.1 $\mathcal{G}$: the otherwise unconstrained transition matrices

We begin with a simple case that will motivate our understanding of the more complex models: the otherwise unconstrained set of $\eta \times \eta$ transition matrices. The other classes of algebraic statistical models that we will consider will be associated with subsets of transition matrices contained within this set of otherwise unconstrained transition matrices. Transition matrices must be stochastic, that is, each row must be nonnegative and sum to 1. Hence, the set of otherwise unconstrained $\eta \times \eta$ transition matrices is the intersection of the the nonnegative orthant, $\mathbb{R}_{+}^{\eta\times\eta}$, with the affine subspace of appropriately-sized matrices with all rows summing to 1, i.e., the affine subspace of matrices $G$ satisfying

$$G\mathbb{1} = \mathbb{1} \quad . \tag{4.18}$$

The affine subspace of $\eta\times\eta$ matrices satisfying (4.18) is an $\eta(\eta-1)$-dimensional set. An affine subspace's intersection with the nonnegative orthant may not necessarily be of the same dimension as the affine subspace, but for our case this will be assured, because the affine subspace of matrices satisfying (4.18) intersects the interior of the nonnegative orthant. For example, the positive matrix with constant entries $\frac{1}{\eta}\mathbb{1}\mathbb{1}'$ is in this intersection.

This intersection is a bounded convex set, characterized by the intersection of an affine subspace with a polyhedral cone, $\mathbb{R}_{+}^{\eta\times\eta}$, and thus, must be finitely-generated. To identify the matrices that are the extreme points of the set of transition matrices (if such extreme points are not immediately obvious), note that each row of a transition matrix is a probability vector of length $\eta$. A probability vector can be thought of as the expected value of the random indicator vector for the underlying random variable, where the indicator vector is a vector of all 0s with a single 1 whose random position indicates the particular value in its alphabet being assumed by the random variable in a realization. Thus, the set of probability vectors of length $\eta$ is a $(\eta-1)$-dimensional convex set that is generated by the $\eta$ row vectors consisting of all 0s except for a single 1. Note that because this convex set's dimension $(\eta-1)$ and its number of extreme points $(\eta)$ differ by 1, any probability vector in this convex set has a unique expansion in terms of extreme points.

If each row of a stochastic matrix can be generated by the appropriately sized indicator vectors, it should be clear that the $\eta \times \eta$ stochastic matrices can be generated by the matrices that are composed of $1 \times \eta$ indicator vectors as their rows, i.e., $\eta \times \eta$ selector matrices. There are a total of $\eta^{\eta}$ selector matrices that serve as the extreme points of this $\eta(\eta-1)$-dimensional set of stochastic transition matrices. Note that a particular expansion of a transition matrix as a convex combination of selector matrices can be thought of as a specification of a joint PMF on $\eta$ random variables, each with alphabet size $\eta$ (think of each row of a given selector matrix as corresponding to an indicator vector for a different random variable, with the weight associated with a given selector matrix in the expansion as the probability of the random variables assuming the combination of values specified by the selector matrix). Provided that a transition matrix has at least two rows with entries in the open unit interval, its expansion in terms of selector matrices is not unique. Or from a probabilistic perspective, when one is provided a sequence of $\eta$ marginal PMFs with at least two being nondeterministic, the joint PMF that can generate such marginal PMFs is not unique. An obvious expansion of a stochastic matrix in terms of the extreme points (the selector matrices) can always be obtained by selecting the coefficients of the expansion to correspond to what would be independent random variables. Such a convex combination would involve all $\eta^\eta$ extreme points for a matrix with positive entries. However, compact convex expansions exist, as by Carathéodory's theorem, any stochastic matrix can be expressed as a convex combination of no more than $\eta(\eta-1)+1$ selector matrices [45], i.e., the dimension of the convex set plus 1.

Although this example of characterizing the otherwise unconstrained transition matrices was somewhat pedantic, the concepts developed offer much intuition into the less familiar cases that we will henceforth consider.

### 4.4.2 $\mathcal{G}(\mathcal{M}^{(n)})$: the transition matrices exhibiting $\mathcal{M}^{(n)}$-separability

In Chapter 3, the set of transition matrices exhibiting $\mathcal{M}^{(n)}$-separability, $\mathcal{G}(\mathcal{M}^{(n)})$, was thoroughly characterized and analyzed. The linear constraints satisfied by $\mathcal{G}(\mathcal{M}^{(n)})$ include (4.18), in addition to

$$\left[\left[M^{(r)}\right]^{\perp}\right]' G M^{(r)} = 0 \, , \tag{4.19}$$

for all $1 \leq r < n$. Like the unconstrained stochastic matrices, $\mathcal{G}(\mathcal{M}^{(n)})$ is an intersection of an affine subspace with $\mathbb{R}_+^{\eta \times \eta}$. The only difference is that the affine subspace is of lower dimension because of the additional linear constraints that must be satisfied.

### 4.4.3 $\mathcal{G}_{GIM}$: the GIM transition matrices

We demonstrated in Section 4.3 that any GIM exhibits $\mathcal{M}^{(n)}$-separability. One may then wonder if $\mathcal{M}^{(n)}$-separability for a network implies that it must be a GIM. This is not true, it can be shown that the set of GIM joint transition matrices, which we denote by $\mathcal{G}_{GIM}$, is a strict subset of the convex set of row-stochastic matrices exhibiting $\mathcal{M}^{(n)}$-separability. It has additional structure that we subsequently develop.

The influencers $\{\mathsf{i}_j\}$ and status update matrices $\{\mathsf{A}_j\}$ of the GIM update equation (4.2) determine the update of the network state. For every possible combination of values for these random quantities, there is a corresponding random selector matrix that describes how the state indicator vector updates. In other words, for a given automata configuration, there exists a function $f$ mapping the random influencers and status update matrices to some random selector matrix $\mathsf{G}$,

$$\mathsf{G} = f\left(\mathsf{i}_1, \mathsf{i}_2, \ldots \mathsf{i}_n, \mathsf{A}_1, \ldots \mathsf{A}_n\right) \, , \tag{4.20}$$

such that

$$\mathsf{s}_\mathsf{x}[t]' = \mathsf{s}_\mathsf{x}[t-1]'\mathsf{G} \, . \tag{4.21}$$

The GIM's special update structure defines the range of the function $f$ in (4.20). This range, a set of selector matrices from which $\mathsf{G}$ assumes values, is determined by the associated automata configuration, and is a strict subset of all selector matrices of the appropriate size, as evident from the linear constraints (4.19) satisfied by any matrix in $\mathcal{G}(\mathcal{M}^{(n)})$. A particular GIM, which is characterized by a joint PMF over the influencers $\{\mathsf{i}_j\}$ and status update matrices $\{\mathsf{A}_j\}$, can be equivalently characterized by its PMF for $\mathsf{G}$—implicitly defined by (4.20)—which assigns a probability to each selector matrix in the range of $f$. A GIM's joint transition matrix is accordingly the expected value of $\mathsf{G}$.

Any joint transition matrix $G$ that can be expressed as a convex combination of selector matrices in the range of $f$ in (4.20) corresponds to a GIM. In particular, each selector matrix in the range of $f$ itself corresponds to a GIM. Therefore, the set of all transition matrices of GIMs sharing a particular automata configuration is indeed convex, with its extreme points being the selector matrices in the range of $f$. The set of GIM transition matrices is generated by a subset of the extreme points of $\mathcal{G}$, all of which must also be extreme points of $\mathcal{G}(\mathcal{M}^{(n)})$.

We now revisit a question that was previously raised. Does the set of GIM joint transition matrices coincide with the set of joint transition matrices exhibiting $\mathcal{M}^{(n)}$-separability? In other words, does $\mathcal{G}(\mathcal{M}^{(n)})$ have extreme points in addition to those defining the set of GIM transition matrices? Although both sets are convex, the answer is no. The set of GIM joint transition matrices with respect to a particular automata configuration can be defined by its extreme points, all of which are selector matrices. On the other hand, $\mathcal{G}(\mathcal{M}^{(n)})$ has additional extreme points. It can be shown, however, that all of the extreme points of $\mathcal{G}(\mathcal{M}^{(n)})$ that are selector matrices are also extreme points of $\mathcal{G}_{GIM}$ (this will be easily argued once we introduce the set of IM transition matrices). Thus, $\mathcal{G}_{GIM}$ can be envisioned as formed from the extreme points of $\mathcal{G}(\mathcal{M}^{(n)})$ as follows: consider the strict subset of its extreme points that are selector matrices and form its convex hull.

### 4.4.4 $\mathcal{G}_{IM}$: the IM transition matrices

The transition matrix of an IM is characterized by (i) being row-stochastic, (ii) satisfying the linear constraints imposed by $M^{(1)}$-separability (the $r = 1$ version of (4.19) ), and (iii) satisfying the polynomial constraints imposed by (4.8) and (4.3), a consequence of the independent updates of the automata. Thus the set of all transition matrices of IMs, denoted as $\mathcal{G}_{IM}$, is a real algebraic variety intersected with the nonnegative orthant. This means that the IM is an algebraic statistical model [54, 55]. Yet more can be said regarding the structure of $\mathcal{G}_{IM}$, and in particular, its connection to $\mathcal{G}_{GIM}$. Recall that the extreme points of $\mathcal{G}_{GIM}$ are selector matrices, and thus correspond to deterministic GIMs. The automata of a deterministic GIM are (trivially) updated independently, and thus any deterministic GIM is also an IM with the same automata configuration (recall Section 4.1.2). Consequently, the convex hull of $\mathcal{G}_{IM}$ is $\mathcal{G}_{GIM}$. By nature of the automata updating independently in an IM, clearly $\mathcal{G}_{IM}$ is a strict subset of $\mathcal{G}_{GIM}$.

We already claimed without proof that all extreme points of $\mathcal{G}(\mathcal{M}^{(n)})$ that are selector matrices are also extreme points of $\mathcal{G}_{GIM}$. It follows that such extreme points must also be in $\mathcal{G}_{IM}$. This can be argued as follows: consider any selector matrix $G \in \mathcal{G}(\mathcal{M}^{(n)})$. Obviously, $G$ is an extreme point of $\mathcal{G}(\mathcal{M}^{(n)})$. As a consequence of $G$'s membership in $\mathcal{G}(\mathcal{M}^{(n)})$, it must be $M^{(1)}$-separable. Furthermore, its automata update independently, by nature of being a selector matrix. It follows that $G$ must be an IM, and thus in both $\mathcal{G}_{IM}$ and $\mathcal{G}_{GIM}$.

Fig. 4-2 attempts to evoke the relationships of the four sets of interest ($\mathcal{G}$, $\mathcal{G}(\mathcal{M}^{(n)})$,

$\mathcal{G}_{GIM}$, and $\mathcal{G}_{IM}$), using a three-dimensional picture.

### Example

A jointly Markovian network of stochastic automata with the simplest non-degenerate automata configuration (two automata, each with two statuses) has a $4 \times 4$ transition matrix. The otherwise unconstrained $4 \times 4$ row-stochastic matrices form a 12-dimensional convex set. The linear constraints imposed by $\mathcal{M}^{(n)}$-separability are given by (4.19) (as $n = 2$, only the $r = 1$ version applies):

$$\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}' \begin{bmatrix} g_{11} & g_{12} & g_{13} & g_{14} \\ g_{21} & g_{22} & g_{23} & g_{24} \\ g_{31} & g_{32} & g_{33} & g_{34} \\ g_{41} & g_{42} & g_{43} & g_{44} \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} = 0 \,, \tag{4.22}$$

or equivalently,

$$\begin{aligned} g_{11} + g_{12} + g_{41} + g_{42} - (g_{21} + g_{22} + g_{31} + g_{32}) &= 0 \\ g_{13} + g_{14} + g_{43} + g_{44} - (g_{23} + g_{24} + g_{33} + g_{34}) &= 0 \\ g_{11} + g_{13} + g_{41} + g_{43} - (g_{21} + g_{23} + g_{31} + g_{33}) &= 0 \\ g_{12} + g_{14} + g_{42} + g_{44} - (g_{22} + g_{24} + g_{32} + g_{34}) &= 0 \end{aligned} \tag{4.23}$$

The second and fourth equations in (4.23) follow from the first and third equations, in conjunction with the linear constraints on the entries of the row-stochastic matrix $G$, namely that each of its rows sum to 1 (4.18). As $\mathcal{M}^{(n)}$-separability imposes only two additional linear constraints, $\mathcal{G}(\mathcal{M}^{(n)})$ forms a 10-dimensional convex set.

Among the extreme points of this 10-dimensional convex set are the selector matrices that are the extreme points of $\mathcal{G}_{GIM}$. Simply scanning the 256 possible selector matrices of this $4 \times 4$ case for those that satisfy the two independent linear constraints in (4.23) yields the 36 selector matrices whose convex hull defines $\mathcal{G}_{GIM}$. In this example, the dimension of $\mathcal{G}_{GIM}$ equals the dimension of $\mathcal{G}(\mathcal{M}^{(n)})$. In fact, we have empirically found that the dimensions of these two sets are equal for all network configurations satisfying $n \leq 5$ with $m_j \leq 5$, but no general result has been established. Yet the two sets do not coincide. The $4 \times 4$ joint transition matrix

$$G = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \tag{4.24}$$

corresponds to a network of stochastic automata that is not a GIM but $\mathcal{M}^{(n)}$-separability, thereby demonstrating that $\mathcal{G}_{GIM}$ is a *strict* subset of $\mathcal{G}(\mathcal{M}^{(n)})$, which is true for other automata configurations as well. In addition, for a network with the transition matrix $G$ as given in (4.24), each automaton will be individually Markovian, with degenerate transition matrices

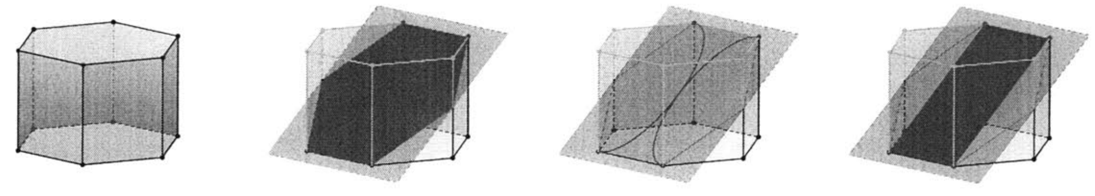
$$\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \,, \tag{4.25}$$

and furthermore, the network will exhibit $\mathcal{M}_n$-separability (for $n = 2$). Evidently, this example also shows that $\mathcal{M}_n$-separability does not imply of GIM of coupled Markov chains.

For an IM with the same automata configuration, the additional polynomial constraints imposed by having independent updates of the automata (combining (4.8) and (4.3) ) require that

$$\begin{aligned} g_{k1} &= (g_{k1}+g_{k2})(g_{k1}+g_{k3}) \\ g_{k2} &= (g_{k1}+g_{k2})(g_{k2}+g_{k4}) \\ g_{k3} &= (g_{k1}+g_{k3})(g_{k3}+g_{k4}) \\ g_{k4} &= (g_{k2}+g_{k4})(g_{k3}+g_{k4})\,, \end{aligned} \tag{4.26}$$

for all $k \in \{1, 2, 3, 4\}$. Thus, the IM transition matrices consist of the joint transition matrices satisfying $\mathcal{M}^{(n)}$-separability along with the constraints embodied in (4.26).



(a) $\mathcal{G}$ is the convex hull of all selector matrices (black dots).

(b) $\mathcal{G}(\mathcal{M}^{(n)})$ is the intersection of a subspace (plane) with $\mathcal{G}$.

(c) Red curves schematically $\mathcal{G}_{IM}$. This set is an algebraic variety in $\mathbb{R}_{+}^{\eta\times\eta}$.

(d) $\mathcal{G}_{GIM}$ is the convex hull of $\mathcal{G}_{IM}$ or equivalently, the convex hull of the selector matrices in $\mathcal{G}(\mathcal{M}^{(n)})$ (red dots).

Figure 4-2: Schematic illustration of the geometric relationships of the IM and GIM transition matrices for a given automata configuration.

### 4.4.5 Products of transition matrices

Because the joint transition matrices of GIMs with a common automata configuration form a convex set whose extreme points are selector matrices, one can show that such a set of transition matrices is closed under matrix multiplication, provided that its extreme points are closed under matrix multiplication. We will use this style of argument to show that the set of all GIM transition matrices sharing a common automata configuration is closed under matrix multiplication.

Consider any two deterministic GIMs sharing the same automata configuration, with respective transition matrices $G^{(1)}$ and $G^{(2)}$ (note that a GIM is deterministic when, and only when, its joint transition matrix $G$ is a selector matrix). We can distinguish between the influencers and status update matrices of the two GIMs using superscripts (i.e., $\mathsf{i}_j^{(1)}$ and $\mathsf{A}_j^{(1)}$; $\mathsf{i}_j^{(2)}$ and $\mathsf{A}_j^{(2)}$, respectively). Because both GIMs are deterministic, we can represent the influencers and status update matrices as functions, that is,

$$\begin{aligned} \mathsf{i}_j^{(k)} &= i^{(k)}(j) \\ \mathsf{A}_j^{(k)} &= A^{(k)}(j)\,. \end{aligned} \tag{4.27}$$

The transition matrix $G^{(1)}G^{(2)}$ is a selector matrix, and can be shown to be the transition matrix of a deterministic GIM whose influencers $\mathsf{i}_j^{(1-2)}$ and status update matrices $\mathsf{A}_j^{(1-2)}$

are deterministic and can be expressed in terms of the functions for the influencers and status update matrices of the two constituent GIMs of (4.27) as follows:

$$\begin{aligned} \mathsf{i}_j^{(1-2)} &= i^{(1)}\left(i^{(2)}(j)\right) \\ \mathsf{A}_j^{(1-2)} &= A^{(1)}\left(i^{(2)}(j)\right) A^{(2)}(j) \ . \end{aligned} \tag{4.28}$$

Therefore, the transition matrices of deterministic GIMs are closed under matrix multiplication, and consequently the set of transition matrices of all GIMs with a common automata configuration is itself closed under matrix multiplication (forming a multiplicative semi-group). An open question is whether or not the set of transition matrices of GIMs is the closure under multiplication of the set of IM transition matrices with a common automata configuration. Under additional assumptions, a subset of the GIM transition matrices satisfy the additional structure of a multiplicative group, i.e., there always exists an inverse and identity. This will be discussed in the subsequent section.

## 4.5 The GIM and the symmetric group: their connection

Recall the example of shuffling the deck of cards from Section 4.2. As explained within the example, shuffling a deck of cards can be conceived of as a GIM. However, we feel it is valuable to digress briefly and explain this same model from the perspective of symmetric group theory. We will introduce this alternate perspective for permutation GIMs without requiring any prior knowledge of representation theory for the symmetric group. This perspective bridges a connection between representation theory of the symmetric group and the propagation of partial information of GIMs.

Before shuffling a deck of cards, the deck has a particular state. Each of the $n$ unique cards has a particular position. If the cards are identified by the integers 1 to $n$, a state of the deck could be specified by noting that card 1 is in position 12, card 2 in position 19, card 3 in position 22, and so forth, with the positions corresponding to the ordering of the cards if the cards were flipped over one-by-one from the top of a face-down deck. Note that the state of the deck can be recognized as a permutation on $n$ integers.

Upon each shuffle, the order of the cards is (possibly) changed, but the new state remains a permutation. The interesting fact of card shuffling is that without looking at the cards face up, one can describe the particular shuffle itself as a permutation on the integers 1 to $n$, e.g., the top card becomes the 22nd card, the 2nd card becomes the 19th card, etc. Therefore, a shuffle defines its own permutation on the cards that, when applied to the state of the deck (itself a permutation), defines a new state (a new permutation). This process is governed by group theory for the symmetric group of order $n$, the set of permutations on the integers 1 to $n$ [7]. By definition, a group consists of a set of elements and a binary operation that allows one to take two elements and define a third element, with an identity element and inverse always existing. The shuffling of a card deck illustrates the symmetric group's operation, which is *composition*: the permutation associated with the shuffle is *composed* with the permutation of the initial card ordering to define a new permutation (a new card ordering). Repeated shuffling of a card deck can be envisioned as repeatedly composing permutations.

Group representation theory [6, 7], and in particular, representation theory for the symmetric group [11, 9], is concerned with identifying the elements of the symmetric group as invertible matrices, that is, defining a function that maps each permutation (an element in the symmetric group) to a square invertible matrix, such that the group operation (composition of permutations) can be represented as multiplication of the corresponding matrices. Such a mapping is called a *homomorphism*, as it is a structure-preserving mapping from one group, in our case the symmetric group, to a second group, in our case the *general linear group*, a multiplicative group of matrices. By representing a group as matrices with matrix multiplication becoming the group operation, linear algebra can offer many insights into the group. This concept may seem rather abstract; however, we will show that the propagation of partial information under $\mathcal{M}^{(n)}$-separability for permutation GIMs is, in effect, a representation for the symmetric group. This will be explained shortly. But first we must discuss the two representations of permutation GIMs.

### 4.5.1 The two representations of permutation GIMs

In the language of automata and networks, Section 4.2 suggests that one conceive of each automaton as corresponding to a different position in the deck, with the automaton indicating the unique card in the particular position, i.e., $\mathbf{s}_i$ indicates the card in the $i$th position. We refer to this representation as the *positions representation.* For Markov chains on permutations, there is a rather important square matrix composed of these status indicator vectors as its columns:

$$\mathsf{S}[t] \triangleq \begin{bmatrix}\mathbf{s}_1[t] & \mathbf{s}_2[t] & \dots & \mathbf{s}_n[t]\end{bmatrix} \quad . \tag{4.29}$$

When the network state corresponds to a permutation, $\mathsf{S}$ is a permutation matrix. Consequently, there is always an alternate perspective for such models, where the *rows* of $\mathsf{S}$ are thought of as the status indicator vectors. In this alternate perspective for shuffling, the automata correspond to unique cards that indicate their particular positions; denote automaton $i$'s status indicator vector for this alternate representation as $\overline{\mathbf{s}_i}$. We refer to this representation as the *cards representation.* Evidently,

$$\begin{bmatrix}\mathbf{s}_1 & \mathbf{s}_2 & \dots \mathbf{s}_n\end{bmatrix} = \mathsf{S} = \begin{bmatrix}\overline{\mathbf{s}}_1' \\ \vdots \\ \overline{\mathbf{s}}_n'\end{bmatrix} \quad , \tag{4.30}$$

and the state array vector for each of the representations is obtained by either column-vectorization or row-vectorization of $\mathsf{S}$:

$$\begin{aligned}\mathbf{s}^{(1)} &= \mathrm{vec}(\mathsf{S}) \\ \overline{\mathbf{s}}^{(1)} &= \mathrm{rvec}(\mathsf{S}) \quad ,\end{aligned} \tag{4.31}$$

where $\mathbf{s}^{(1)}$ is the 1st-order state array vector for the positions representation, and $\overline{\mathbf{s}}^{(1)}$ is the 1st-order state array vector for the cards representation.

When the automata are associated with positions, as described in Section 4.2, the influencers chosen by the automata form a permutation. Put in another way, the influencers at time $t$ can be represented via a random permutation matrix $\mathsf{D}[t]$, which we refer to as the *random influence matrix.* The $j$th row of $\mathsf{D}[t]$ equals $\boldsymbol{e}_i$ when $\mathsf{i}_j = i$, that is, the $j$th row of $D[t]$ is an indicator vector for the random variable $\mathsf{i}_j$ (at time $t$), the influencer of automaton $j$. By time-homogeneity, the expected value of $\mathsf{D}[t]$ is the influence matrix introduced in Section 4.1.1. As described in Section 4.2, for the positions representation, the random selector matrices $\mathsf{A}_j$ are always identities.

For the positions representation, we can express updates of the status indicator vectors using only the random influence matrices. In particular,

$$\begin{bmatrix} \mathsf{s}_1[\tau] & \mathsf{s}_2[\tau] & \ldots & \mathsf{s}_n[\tau] \end{bmatrix} \cdot D[\tau+1]' \cdot D[\tau+2]' \cdot \ldots \cdot D[t]' = \begin{bmatrix} \mathsf{s}_1[t] & \mathsf{s}_2[t] & \ldots & \mathsf{s}_n[t] \end{bmatrix} \quad . \tag{4.32}$$

On the other hand, for the dynamics of the cards representation to be captured as a GIM, each automaton must always choose itself as an influencer, i.e., we have a GIM of coupled Markov chains. The random selector matrix $\mathsf{A}_j$ describes how the positions of the cards are rearranged. In particular, the $k$th row of $\mathsf{A}_j$ indicates the new position of the card that was formerly in the $k$th position before current shuffle. As this description of the rearrangement of the cards is the same for each card $j$, at each time $t$ all random selector matrices $\mathsf{A}_j[t]$ must be equal; hence all random selector matrices can be collectively denoted as $\mathsf{A}[t]$. By this cards representation, we have an alternate description of permutation GIMs: as GIMs of coupled Markov chains with the special property that at each time $t$, all $\mathsf{A}_j[t]$ are equal.

Analogously to what we demonstrated in (4.32) for the positions representation, we can express the updates of the status indicator vectors for the cards representation using only the random selector matrices $\mathsf{A}[t]$:

$$\begin{bmatrix} \bar{\mathsf{s}}_1[\tau]' \\ \bar{\mathsf{s}}_2[\tau]' \\ \vdots \\ \bar{\mathsf{s}}_n[\tau]' \end{bmatrix} \cdot \mathsf{A}[\tau+1] \cdot \mathsf{A}[\tau+2] \cdot \ldots \cdot \mathsf{A}[t] = \begin{bmatrix} \bar{\mathsf{s}}_1[t]' \\ \bar{\mathsf{s}}_2[t]' \\ \vdots \\ \bar{\mathsf{s}}_n[t]' \end{bmatrix} \quad . \tag{4.33}$$

Note that (4.33) simultaneously captures each status vector update equation, (4.2) for all $j$, as the automata always choose themselves as influencers.

### 4.5.2 The 1st-order permutation representation of the symmetric group

Both (4.32) and (4.33) are the same update equation. This becomes obvious by recalling (4.30) in addition to recognizing that $\mathsf{D}[t]'$ of the positions representation equals $\mathsf{A}[t]$ of the cards representation. In representation theory of the symmetric group, this equation, (4.32) or equivalently (4.33), is referred to as the *1st-order permutation representation* [10]. We have represented a succession of composition of permutations as a product of invertible matrices—which is the idea of representation theory for groups.

As evident in both (4.32) and (4.33), the 1st-order permutation representation begins with a permutation matrix $\mathsf{S}[\tau]$ that captures the initial arrangement of the deck at time $\tau$. Then a succession of random $n \times n$ permutation matrices describe the random rearrangements of the cards for each shuffle. The product of these matrices is a permutation matrix $\mathsf{S}[t]$ that describes the state of the deck at time $t$ after $(t-\tau)$ shuffles.

For example, if $n = 3$, we may have the realized product of permutation matrices

$$\underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{\substack{\text{state of deck} \\ \text{at time } \tau}} \cdot \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}}_{\text{1st shuffle}} \cdot \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{\text{2nd shuffle}} = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\substack{\text{state of deck} \\ \text{at time } \tau+2}} \quad , \tag{4.34}$$

which would indicate an initial state at time $\tau$ followed by two shuffles. Initially, card 3 is in the top position, card 2 in the middle and card 1 on the bottom. Upon the first shuffle, represented by the second matrix, the cards in the 2nd and 3rd positions swap. On the second shuffle, the cards in the 1st and 3rd positions swap. What results is a state of a deck represented by the matrix on the right hand side of (4.34): card 1 is in the middle of the deck, card 2 at the top, and card 3 at the bottom.

### 4.5.3 Propagating partial information

The 1st-order state array vector for the cards representation can be obtained from $\mathsf{S}$ by row-vectorization (4.31). Using this fact, we can use the 1st-order permutation representation of the symmetric group (4.33) to derive our update equations for the 1st-order state array vector, namely,

$$\begin{aligned}
\operatorname{rvec}\left(\begin{bmatrix} \bar{\mathsf{s}}_1[t]' \\ \vdots \\ \bar{\mathsf{s}}_n[t]' \end{bmatrix}\right) &= \operatorname{rvec}\left(\begin{bmatrix} \bar{\mathsf{s}}_1[\tau]' \\ \vdots \\ \bar{\mathsf{s}}_n[\tau]' \end{bmatrix} \mathsf{A}[\tau+1]\mathsf{A}[\tau+2]\ldots\mathsf{A}[t]\right) \\
\bar{\mathsf{s}}^{(1)}[t]' &= \operatorname{rvec}\left(I^{t-\tau}\begin{bmatrix} \bar{\mathsf{s}}_1[\tau]' \\ \vdots \\ \bar{\mathsf{s}}_n[\tau]' \end{bmatrix} \cdot \mathsf{A}[\tau+1]\mathsf{A}[\tau+2]\ldots\mathsf{A}[t]\right) \\
&= \bar{\mathsf{s}}^{(1)}[\tau]'\,(I\otimes\mathsf{A}[\tau+1])\,(I\otimes\mathsf{A}[\tau+2])\ldots(I\otimes\mathsf{A}[t]) \quad . \qquad (4.35)
\end{aligned}$$

We can then take an expectation of $\bar{\mathsf{s}}^{(1)}[t]$ conditioned on $\bar{\mathsf{s}}^{(1)}[\tau]$ to obtain our familiar update equation,

$$\mathbb{E}\left[\bar{\mathsf{s}}^{(1)}[t]' \,\|\, \mathsf{s}^{(1)}[\tau]\right] = \bar{\mathsf{s}}^{(1)}[\tau]'\,(I\otimes\mathbb{E}[\mathsf{A}])^{t-\tau} \quad , \qquad (4.36)$$

where we have followed the same probabilistic arguments detailed in Section 4.3.1 as well as used the distributive property of Kronecker products (A.2) to show that $\mathbb{E}[I\otimes\mathsf{A}] = I\otimes\mathbb{E}[\mathsf{A}]$.

From (4.36), it is evident that for the cards representation,

$$\overline{H}^{(1)} = I\otimes\mathbb{E}[\mathsf{A}] \quad . \qquad (4.37)$$

Evidently, the 1st-order permutation representation of the symmetric group is simply the realization representation of $M^{(1)}$-separability for permutation models, but in matrix form as opposed to vector form! Note that both representations require the same computation, as $H^{(1)}$ is a block diagonal matrix.

Using the same arguments and a little more Kronecker algebra, we can show that the matrix that propagates the partial information in the form of univariate marginals for the positions representation must be

$$H^{(1)} = \mathbb{E}\left[\mathsf{D}'\right]\otimes I \quad , \qquad (4.38)$$

where $\mathbb{E}[\mathsf{A}] = \mathbb{E}[\mathsf{D}']$.

Reassuringly, $\overline{H}^{(1)}$ and $H^{(1)}$ are similar, i.e., they share the same eigenvalues, as evident by the properties of eigenvalues of Kronecker products as described in Section A.1, or more simply, by realizing from (4.30) that $\bar{\mathsf{s}}^{(1)}$ is a permutation of $\mathsf{s}^{(1)}$.

### 4.5.4 Additional representations of the symmetric group

The parametric form of the GIM automatically permits one to recognize additional representations of the symmetric group. For example, we have

$$\begin{bmatrix} \mathsf{s}_1[\tau]' \otimes \mathsf{s}_1[\tau]' \\ \mathsf{s}_1[\tau]' \otimes \mathsf{s}_2[\tau]' \\ \vdots \\ \mathsf{s}_n[\tau]' \otimes \mathsf{s}_n[\tau]' \end{bmatrix} \cdot (\mathsf{A}[\tau+1] \otimes \mathsf{A}[\tau+1]) \cdot (\mathsf{A}[\tau+2] \otimes \mathsf{A}[\tau+2]) \cdot \ldots \cdot (\mathsf{A}[t] \otimes \mathsf{A}[t])$$

$$= \begin{bmatrix} \mathsf{s}_1[t]' \otimes \mathsf{s}_1[t]' \\ \mathsf{s}_1[t]' \otimes \mathsf{s}_2[t]' \\ \vdots \\ \mathsf{s}_n[t]' \otimes \mathsf{s}_n[t]' \end{bmatrix} , \qquad (4.39)$$

which is a somewhat more redundant version of the *2nd-order permutation representation* of the symmetric group [10]. Under expectations and row-vectorization, this redundant 2nd-order permutation representation becomes the $M^{(2)}$-separability equation.

The representations that we have mentioned and their extensions derived from the $M^{(r)}$-separability equations are considered reducible representations for the symmetric group. We refer the reader to [11, 9, 10] for a discussion of *irreducible* representations for the symmetric group. The irreducible representations for the symmetric group are obtained by decomposing the reducible representations via a process that is more extensive, but similar in spirit to how redundant partial information vectors like $\boldsymbol{\pi}^{(r)}$ are represented in more compact forms by using marginalization matrices of full rank.

### 4.5.5 Analyzing the eigenvalues of $G$

Using $\overline{H}^{(1)}$ (or equivalently, $H^{(1)}$) to analyze the eigenvalues of $G$, the underlying transition matrix, is tricky on two fronts: $H^{(1)}$ has irrelevant eigenvalues[1], and secondly, the underlying transition matrix $G$ is not irreducible. Let's first consider the latter of these two issues. There is a recurrent class for the states corresponding to permutations, as well as several other recurrent classes for states that are not permutations (it can be easily argued that there are no transient classes). Our interest lies only in the eigenvalues of $G$ that correspond to the recurrent class of states that are permutations—and not the eigenvalues corresponding to the other recurrent classes. However, one can argue that the convergence to steady-state for the recurrent class consisting of states that are permutations must be as slow as, if not slower, than the convergence to steady-state for any of the other recurrent classes (provided that we have a steady-state, which is ensured when the shuffling technique gives rise to an aperiodic recurrent class on permutation states; this is always assumed). One way to show this is that the dynamics in any of the other recurrent classes can be thought of as the output of a hidden Markov model (HMM) [56] whose underlying Markov chain is on the recurrent class of permutation states. The idea is that several of the unique cards are indistinguishable as the outputs of the HMM that captures the dynamics in any of the other recurrent classes. We can then claim that $\overline{H}^{(1)}$'s second largest relevant eigenvalue by modulus must be a lower bound on the modulus of the eigenvalue of $G$ that determines

[1] Relevant eigenvalues are eigenvalues that can be excited by partial information vectors $\boldsymbol{\pi}^{(1)}$, which lie in a particular subspace as $M^{(1)}$ fails to have full column rank.

the speed of convergence to steady-state.

As stated in Section A.1, the eigenvalues of the Kronecker product of two matrices are the pairwise products of the constituent matrices' eigenvalues. The particular form of $\overline{H}^{(1)}$ as given in (4.37) indicates a rather special fact regarding the eigenvalues of $\overline{H}^{(1)}$: they are the eigenvalues of $\mathbb{E}[\mathsf{A}]$. Because of the freedom offered by having a Kronecker product with the identity matrix, moreover, one can argue that the second largest eigenvalue by modulus in $\mathbb{E}[\mathsf{A}]$ will be a relevant eigenvalue in $\overline{H}^{(1)}$. Hence, $\mathbb{E}[\mathsf{A}]$ can be used to derive lower bounds on the eigenvalues of $G$ that determine the speed of convergence to steady-state. As the cards representation allow us to conceive of card shuffling as a GIM of coupled Markov chains, we have $M_i$-separability for each automaton $i$ under the cards representation. By (4.33), $\mathbb{E}[\mathsf{A}]$ is the matrix that propagates the partial information $\boldsymbol{\pi}_i$ under $M_i$-separability. Thus, rather than using the $n^2 \times n^2$ matrix $\overline{H}^{(1)}$ to obtain lower bounds on the speed of convergence for permutation GIMs, we can obtain an equivalent lower bound by using the $n \times n$ matrix $\mathbb{E}[\mathsf{A}]$.

Lastly, we note that by following the approach in Section 4.3.1 and making use of the mixed-product property of Kronecker product (A.3), one can show that the underlying transition matrix $G$ for the cards representation can also be expressed in terms of $\mathsf{A}$:

$$G = \mathbb{E}\left[\mathsf{A}^{\otimes n}\right] \quad . \tag{4.40}$$

However, (4.40) does not seem to be particularly useful for determining the eigenvalues of $G$, as in general, $\mathbb{E}[\mathsf{A}]^{\otimes r} \neq \mathbb{E}[\mathsf{A}^{\otimes r}]$.

We close this discussion on the connections between the GIM and Markov chains that preserve permutations by noting that one may consider other multiplicative groups consisting of a larger subset of $\mathcal{G}_{GIM}$ than the subset preserving permutations. For example, if we consider the GIMs with the restriction that each random influencer matrix $\mathsf{D}$ and random selector matrices $\mathsf{A}_1, \ldots, \mathsf{A}_n$ are all permutation matrices, we have a multiplicative group (as an inverse always exists). In fact, one can show that the GIMs with this description constitute the largest subset of $\mathcal{G}_{GIM}$ that is a multiplicative group—as permutation matrices are the only selector matrices that are invertible.

## 4.6 Additional GIM examples

The first example in this section illustrates how under specific conditions, a sampled IM can be analyzed as several separate GIMs of smaller order. The second example establishes that a different IM extension presented in [57] is itself a GIM.

### 4.6.1 IM with periodic influence network graph

The dynamics of an IM are partially captured by the influence network graph $\Gamma(D')$ and its corresponding influence matrix $D = [d_{jk}]$, whose $(j, k)$ entry is the probability of automaton $j$ choosing automaton $k$ as its influencer (see Section 4.1.1 for a discussion of the $d_{jk}$).

Whenever the influence network graph $\Gamma(D')$ is bipartite (i.e., has period 2, see [2]), there must exist a partition of the automata into two sets, $S_1$ and $S_2$, such that automata in $S_1$ always choose automata in $S_2$ as their influencers, and vice-versa. For example, an IM modeling the spread of a virus has an influence matrix $D$ with such a periodic structure. One subset of automata corresponds to people, and the other subset of automata corresponds

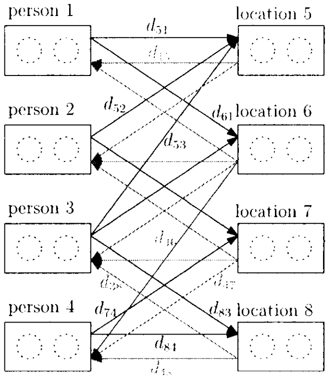


Figure 4-3: Periodic $\Gamma(D')$ (with statuses shown).

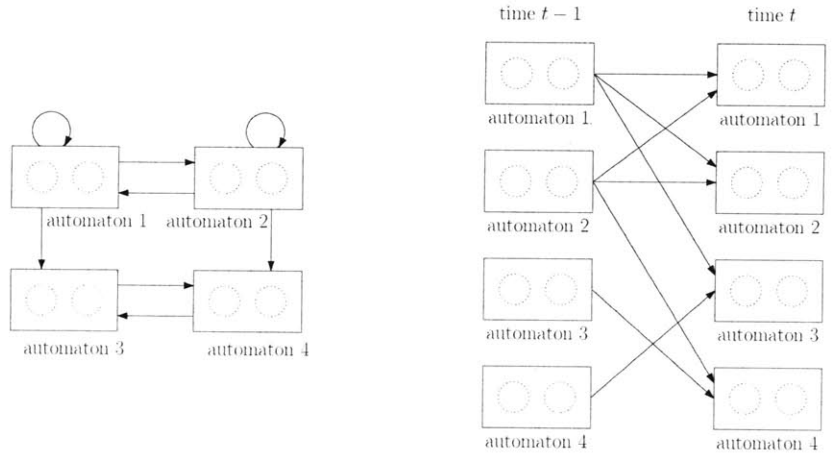
to locations where they might mingle, or to intermediary disease hosts (vectors) [58], [59]. Each automaton exhibits one of two statuses, either susceptible or infected (with respect to the virus). The periodic structure of $D$ demands that for the spread of infection, an infected person first must infect a location, with a susceptible person then being infected by the infected location. An example of such a periodic $D$ is illustrated in the bipartite influence network graph of Fig. 4-3.

Whenever the influence network graph $\Gamma(D')$ of an IM is periodic with period $\tau$, we can define a *sampled model* with joint transition matrix $G^{\tau}$, such that the dynamics of the sampled model mirrors that of our original IM (with its joint transition matrix $G$) sampled every $\tau$ iterations. Because the transition matrices of GIMs are closed under matrix multiplication, $G^{\tau}$ is the joint transition matrix of a GIM, *but not necessarily of an IM.* In the sampled model, the automata within a periodic class (a partition of automata induced by the periodicity of the influence network graph) are always influenced only by other automata within the class. Therefore, each class of automata constitutes its own lower-order GIM. Because our original unsampled model is an IM and consequently assumes conditional independence of automata updates, each of these lower-order GIMs updates independently.

In our example of a spreading virus, the sampled model with transition matrix $G^2$ consists of two independently-updating GIMs, one for the locations and another for the people. For any IM with an influence network graph $\Gamma(D')$ that is periodic with period $\tau$, the corresponding sampled model can be analyzed as $\tau$ independently-updating GIMs, each of smaller order than the original IM.

### 4.6.2 IM extension allowing same-time influencers

An extension of the IM introduced in [57] has the same flavor as the influence model, except that same-time influencers are permitted. This extension can be visualized graphically. Consider the influence network graph $\Gamma(D')$: an arc from node (automaton) $j$ to node (automaton) $k$ exists when automaton $j$ can choose automaton $k$ as its influencer for its update. For an IM, because its automata update independently, a Bayesian network [41] encoding these conditional independence relationships can be derived from the influence network graph as follows. Two instances of each automaton are drawn, one at time $t-1$ and another at time $t$. If $d_{jk} > 0$, that is, if the arc from $k$ to $j$ exists in the influence network graph $\Gamma(D')$, then an arc from automaton $k$ at time $t-1$ to automaton $j$ at time $t$ is drawn in the Bayesian network. Contrary to the custom for Bayesian networks, the arcs in our Bayesian network are labeled with weights inherited from the corresponding arcs in the influence network graph. As will soon be evident, when an IM is given, such arc weights are informative.. This new graph is referred to as the *IM-update Bayesian network* for time $t$. An example of an IM's influence network graph $\Gamma(D')$ and its corresponding IM-update Bayesian network is illustrated in Fig. 4-4. For an IM, the influence network graph is a more compact means to represent the conditional independence relationships encoded by the IM-update Bayesian network.



(a) Influence network graph $\Gamma(D')$.

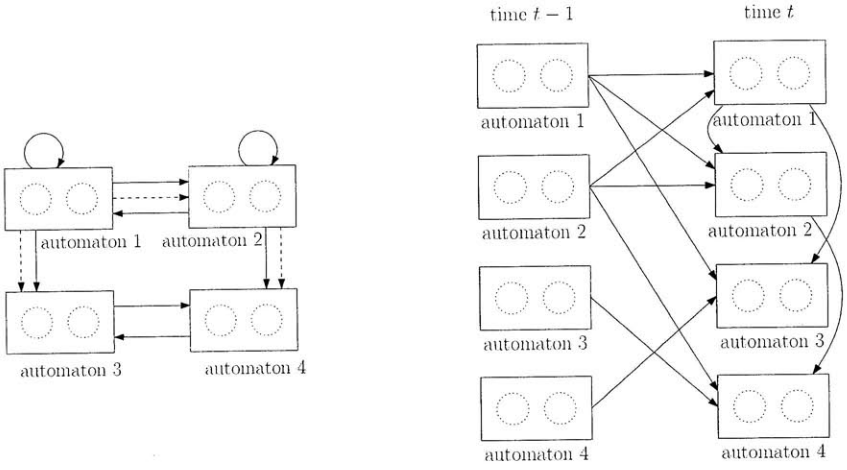
(b) IM-update Bayesian network for time $t$.

Figure 4-4: Comparing $\Gamma(D')$ and its corresponding Bayesian network for an IM.

The Bayesian network governing updates in the extended IM (or EIM) of [57] allows arcs between automata at time $t$. An automaton with such "same-time" arcs indicates that for its update from $t-1$ to $t$, it can choose an automaton's status at time $t$ to serve as its influencer—in contrast to the IM, where each automaton chooses an automaton's status at time $t-1$ as its influencer.

The same-time influencers in an EIM can be incorporated in an influence network graph by allowing two arc types. A *solid arc* indicates that an automaton can be influenced by the status at time $t-1$ of the automaton from which the arc originates; a *dashed arc* indicates that it can be influenced by the status at time $t$. We require that the same-time subgraph, which comprises just the dashed arcs, be acyclic. The arc weights, as in an IM's influence network graph, indicate the probabilities of choosing as an influencer the respective automata at which the arcs originate, and the arc type (solid or dashed) indicates whether it iterates based on its influencer's current or immediately preceding status. The weights of a node's incoming arcs of all types must sum to 1. An example of an EIM's influence network graph and corresponding EIM-update Bayesian network is illustrated in Fig. 4-5.

The EIM-update Bayesian network is acyclic, because of our restriction that the same-time subgraph be acyclic. Consequently, the automata can be partitioned into subsets that can be considered as updating sequentially, whereby each automaton updates only after all other automata that it could choose as a same-time influencer have updated. In the example of Fig. 4-5, the automata can update in the following order: $\{1\}, \{2, 3\}, \{4\}$.



(a) Influence network graph.

(b) EIM-update Bayesian network for time $t$.

Figure 4-5: Comparing $\Gamma(D')$ and its corresponding Bayesian network in an EIM.

Because same-time influencers are allowed in an EIM, row-stochastic matrices $Z_{jk}$ must be defined that specify the conditional PMF of the current value of automaton $j$, given its choice of automaton $k$ at the *current* time as its influencer. This is analogous to the familiar row-stochastic matrices $A_{jk}$ of the IM, specifying the conditional PMF of the current value of automaton $j$ given its choice of automaton $k$ at the *immediately preceding* time as its influencer.

The extension in [57] of the IM to allow same-time influencers is a specific example of a GIM. This can be demonstrated by construction: a model equivalent to the EIM will be defined in terms of probabilistically chosen influencers and selector matrices, in a manner analogous to how the GIM was defined in Section 4.1. A GIM will subsequently be derived from such a random-update characterization.

Consider an EIM-update Bayesian network, and for each automaton $j$ at time $t$, draw independent random influencers $\bar{\mathrm{i}}_j$ according to the marginal PMFs specified by the weights of the incoming arcs. When the influencer of automaton $j$ at time $t$ is another automaton $k$ at the same time, $\bar{\mathrm{i}}_j = k[t]$ is written. On the other hand, if the influencer of automaton $j$ at time $t$ is automaton $k$ at the previous time $t-1$, $\bar{\mathrm{i}}_j = k[t-1]$ is written. By adopting this notation, previous-time and same-time influencers can be differentiated.

A particular set of realized values for $\bar{\mathrm{i}}_1, \bar{\mathrm{i}}_2, \ldots \bar{\mathrm{i}}_n$ induces a subgraph of the EIM-update Bayesian network, and for each automaton $j$, there exists a unique path, denoted as $\pi_j$, terminating at the automaton $j$'s instance at time $t$ and originating at some automaton's instance at time $t-1$. Such a path is unique, as each automaton at time $t$ has exactly one arc terminating at it in such a realized subgraph. This random path $\pi_j$ can be characterized by the instances of the automata at time $t$ at which its arcs terminate. An example of such a realized subgraph is illustrated in Fig. 4-6. For the realization illustrated in Fig. 4-6, $\pi_4 = \{4, 2, 1\}$ and $\pi_3 = \{3\}$

Similarly, probabilistically chosen selector matrices, $\bar{\mathsf{A}}_j$, can be defined, which are con-

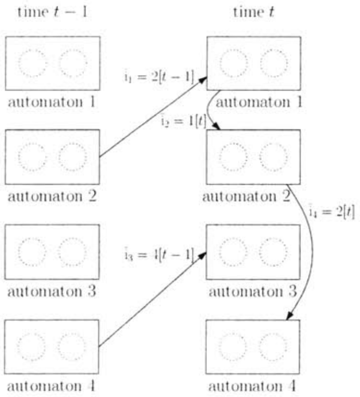

time $t-1$
time $t$
automaton 1
automaton 2
automaton 3
automaton 4
automaton 1
automaton 2
automaton 3
automaton 4
$\bar{\mathsf{i}}_1 = 2[t-1]$
$\bar{\mathsf{i}}_2 = 1[t]$
$\bar{\mathsf{i}}_3 = 4[t-1]$
$\bar{\mathsf{i}}_4 = 2[t]$

Figure 4-6: Realization of influencers in an EIM.

ditionally independent given $\bar{\mathsf{i}}_1, \bar{\mathsf{i}}_2, \ldots \bar{\mathsf{i}}_n$ and satisfy

$$\begin{aligned} \mathbb{E}\left[\,\bar{\mathsf{A}}_j \,\middle|\, \bar{\mathsf{i}}_j = k[t]\,\right] &= Z_{jk} && (4.41)\\ \mathbb{E}\left[\,\bar{\mathsf{A}}_j \,\middle|\, \bar{\mathsf{i}}_j = k[t-1]\,\right] &= A_{jk}\;. && (4.42)\end{aligned}$$

Consider a process defined using this formulation and updated sequentially, much as the EIM from which it is defined, as follows:

$$\begin{aligned} \mathbf{s}_j[t]' = \mathbf{s}_k[t]'\bar{\mathsf{A}}_j &\qquad \text{when } \bar{\mathsf{i}}_j = k[t]\;\;, \\ \mathbf{s}_j[t]' = \mathbf{s}_k[t-1]'\bar{\mathsf{A}}_j &\qquad \text{when } \bar{\mathsf{i}}_j = k[t-1]\;\;. \end{aligned} \tag{4.43}$$

It should be clear by its construction that such a process has the same probabilistic dynamics as the EIM.

Furthermore, a GIM can be defined in terms of this process with its random-update formulation. Probabilistically chosen influencers and selector matrices will be specified for the GIM, and as a consequence of substitution and associativity of matrix multiplication, the GIM will exhibit identical dynamics. Specifically, automaton $j$'s influencer in the GIM, $\mathsf{i}_j$, can be defined as the automaton at which the random path $\pi_j$ originates, and the random selector matrices $\mathsf{A}_j$ of a GIM can be defined as

$$\mathsf{A}_j = \prod_{k\in\pi_j} \bar{\mathsf{A}}_k\;. \tag{4.44}$$

Note that the joint PMF over the influencers $\mathsf{i}_j$ and update matrices $\mathsf{A}_j$, which is the key to defining a GIM, is implicitly defined via the specified joint PMF over the $\bar{\mathsf{i}}_j$ and $\bar{\mathsf{A}}_j$.

## 4.7 Conclusions

We have introduced the GIM as a generalization of the IM that maintains many of its beneficial properties—namely that expected values of many different state vectors can be propagated linearly in time. This enables efficient asymptotic analysis of GIMs via marginal (e.g., pairwise) characterizations of the joint PMF for the network state, with algorithms of complexity only polynomial in the network size. A few examples of GIMs have been illustrated that make use of the more interesting behavior made possible by this generalization. Some of the algebraic and geometric relationships of the joint transition matrices of GIMs have also been described.

Markov chains over permutations were shown to be a special case of the GIM, thereby bridging a connection between the GIM and representation theory for the symmetric group. Furthering the connections of the GIM to group theory could be potentially valuable. It may be possible to leverage some of the power of group theory, e.g., the convergence bounds of [8], to simplify the analysis of GIMs in special cases.

One of the utilities of the IM is that parameter estimation algorithms have been developed, and the success of these algorithms in a variety of problems has been shown [60, 61, 62]. The utility of GIMs would be similarly enhanced by efficient algorithms for estimating unknown parameters from both complete information (a sequence of network states) as well as incomplete information (a non-injective function of a sequence of network states, i.e., a sequence of outputs of a hidden Markov model [63] whose underlying Markov process is a GIM). The next chapter offers a solution to this problem in the general setting of $M_*$-separability.

# Chapter 5

# Parameter Learning under Separability

This chapter focuses on learning the parameters of an $M_*$-separable stochastic network based on an observed realization. Maximum likelihood (ML) estimation of $G$, the underlying Markov chain's transition matrix, is the natural problem to first consider. Although estimating $G$ is intractable when $n$ is large, discussing this problem prepares the reader for the primary focus of this chapter—learning $H_*$, the matrix that propagates the partial information vector $\pi_*$. A proposal to modify the algorithm for the ML estimate of $G$ by replacing Kullback-Leibler (KL) divergences in an objective function with Euclidian ($\ell_2$) distances motivates our algorithm that estimates $H_*$.

At the computational core of this algorithm to estimate $H_*$ is linear least-mean-squares estimation. Because of this, our proposed algorithm offers familiarity and tractability. The algorithm to estimate $H_*$ has computational complexity that is polynomial in the length of the partial information vector $\pi_*$, meaning that it is polynomial in $n$ for our canonical examples of $M^{(r)}$-separability.

Our estimates of $H_*$ exhibit several desirable properties. We show that our estimate of $H_*$ is strongly consistent, i.e., it converges almost surely to its true value, provided that the underlying Markov chain is ergodic. Despite that fact that characterizing the feasibility set for $H_*$ is intractable for many of our interesting examples of $M_*$-separability—necessitating that we ignore feasibility when computing an estimate of $H_*$—iterative estimates of $H_*$ based on observing an ever-increasing trajectory will almost surely become feasible and remain feasible provided that the underlying transition matrix $G$ is strictly positive.

In two computational examples we illustrate the ability to learn $H_*$ for networks whose size forbids traditional analysis. Our experiments demonstrate known results regarding the the rate of mixing of a deck of cards under some standard shuffling techniques, without any reliance on the theory of the symmetric group. In addition, we illustrate our parameter learning algorithm on a more general second example. In each case, our computational experiments require a relatively short trajectory before our estimate is feasible for subsequent observations and is close to the true $H_*$ by a reasonable measure of distance. Lastly, we suggest a criterion that offers a rough measure of how reasonable it may be to assume $M_*$-separability.

## 5.1 Parameter estimation for Markov chains

It is instructive to review parameter learning for time-homogeneous finite-state Markov chains before proceeding to the special case of parameter learning under separability. Let $\mathsf{x}[t]$ denote the scalar state of our Markov chain at time $t$, and assume that its finite alphabet is the positive integers from 1 to $\eta$. Because of the constraints imposed by the Markov property, the statistical model is fully parameterized by the initial probabilities $\sigma_i = \mathbb{P}(\mathsf{x}[0] = i)$ and the transition probabilities $g_{ij}$, the probability that the Markov chain transitions to state $j$ at the next time-step, given that its current state is $i$, i.e., $g_{ij} \triangleq \mathbb{P}(\mathsf{x}[t] = j \,|\, \mathsf{x}[t-1] = i)$. Note that the probability of any event over a finite-time horizon can be expressed as a polynomial expression in terms of the parameters $\{g_{ij}\}$ and $\{\sigma_i\}$. Equivalently, a Markov chain is an algebraic statistical model [54, 55].

We will often represent all parameters by a single vector $\boldsymbol{\theta}$, and denote $\Theta$ as the set of all parameter configurations, i.e., the set of all possible values for $\boldsymbol{\theta}$. The parameter estimation problem for Markov chains reduces to the problem of estimating $g_{ij}$ and $\sigma_i$ for all states $i, j$, based on observing a trajectory (sequence of states) from time 0 to time $t$, i.e., we observe that $\mathsf{x}[0] = x[0]$, $\mathsf{x}[1] = x[1]$, $\ldots$ $\mathsf{x}[t] = x[t]$, which is expressed compactly as $\mathbf{x}_0^t = \boldsymbol{x}_0^t$.

The likelihood of observing the trajectory $\boldsymbol{x}_0^t$ can be expressed in terms of the parameters as:

$$\begin{aligned} \mathbb{P}\left(\mathbf{x}_0^t = \boldsymbol{x}_0^t \,;\, \boldsymbol{\theta}\right) &= \sigma_{x[0]} \prod_{\tau=1}^{t} g_{x[\tau-1]x[\tau]} \\ &= \sigma_{x[0]} \prod_{i,j} g_{ij}^{n_{ij}(\boldsymbol{x}_0^t)} , \end{aligned} \tag{5.1}$$

where $n_{ij}(\boldsymbol{x}_0^t)$ is the count of the number of transitions from $i$ to $j$ over the trajectory $\boldsymbol{x}_0^t$, and we have used the shorthand notation $\boldsymbol{\theta}$ to represent all parameters, i.e.,

$$\boldsymbol{\theta} = [\sigma_1\, \sigma_2 \ldots \sigma_\eta\; g_{11}\; g_{12} \cdots g_{\eta\eta}]' \quad . \tag{5.2}$$

Our parameter vector $\boldsymbol{\theta}$ can be simplified by defining the $\eta \times 1$ probability vector

$$\boldsymbol{\sigma} = [\sigma_1\, \sigma_2 \ldots \sigma_\eta]' \quad , \tag{5.3}$$

and $\eta \times \eta$ row-stochastic matrix

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1\eta} \\ g_{21} & g_{22} & \cdots & g_{2\eta} \\ \vdots & \vdots & \ddots & \vdots \\ g_{\eta 1} & g_{\eta 2} & \cdots & g_{\eta\eta} \end{bmatrix} \quad . \tag{5.4}$$

We express our parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\theta}, G)$.

Typically, the approach in parameter estimation is to find the values of the parameters that maximize the likelihood, the probability of observing the trajectory $\boldsymbol{x}_0^t$ as expressed in (5.1). Such parameter estimates, denoted as $\widehat{\sigma}_i$ and $\widehat{g}_{ij}$, are called maximum likelihood (ML) estimates, and are found as

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta}\; \mathbb{P}\left(\mathbf{x}_0^t = \boldsymbol{x}_0^t \,;\, \boldsymbol{\theta}\right) \quad . \tag{5.5}$$

Up to this point we have not discussed how the parameter space $\Theta$ is specified. In the Markov chain case, defining $\Theta$ is straightforward as our parameters $\sigma_i$ and $g_{ij}$ are probabilities. Thus a possible parameter combination $\boldsymbol{\theta} = (\boldsymbol{\sigma}, G) \in \Theta$, must be nonnegative and satisfy the linear constraints

$$\begin{aligned} \boldsymbol{\sigma}'\mathbb{1} &= \mathbb{1} && (5.6) \\ G\mathbb{1} &= \mathbb{1} \quad , && (5.7) \end{aligned}$$

or equivalently, $\boldsymbol{\sigma}$ must be a probability vector, and $G$ a row-stochastic matrix.

Returning to the computation of ML estimates, by inspection of (5.1), one should set $\widehat{\sigma}_j = 1$ for $j$ such that $x[0] = j$, and $\widehat{\sigma}_i = 0$ for all $i \neq j$ (we use a 'hat' $\widehat{}$ to denote the estimate of a parameter that is computed from an observed trajectory). Because of the factorable form of the likelihood in (5.1) and the constraints on our parameter space $\Theta$, the estimated transition probabilities can be computed separately for each row of $G$. In particular, by denoting

$$\boldsymbol{g}_i = [g_{i1}\ g_{i2}\ \cdots\ g_{i\eta}] \tag{5.8}$$

as the $i$th row of $G$, it follows from the factorable form of (5.1) that

$$\widehat{\boldsymbol{g}}_i = \arg\max_{\boldsymbol{g}_i \in \Delta_\eta} \prod_j g_{ij}^{n_{ij}(\boldsymbol{x}_0^t)} \quad , \tag{5.9}$$

where $\Delta_\eta$ is the set of probability vectors of length $\eta$. We have reduced our ML estimation problem into $\eta$ smaller estimation problems for each state $i$.

It may be intuitively obvious that the ML estimate $\widehat{g}_{ij}(\boldsymbol{x}_0^t)$ can be found via frequency counting:

$$\widehat{g}_{ij}(\boldsymbol{x}_0^t) = \frac{n_{ij}(\boldsymbol{x}_0^t)}{n_{i:}(\boldsymbol{x}_0^t)} \quad , \tag{5.10}$$

where $n_{i:}(\boldsymbol{x}_0^t) \triangleq \sum_j n_{ij}(\boldsymbol{x}_0^t)$. The intuition is that $\widehat{g}_{ij}(\boldsymbol{x}_0^t)$ should be the fraction of times when in state $i$ that the Markov chain immediately transitions to state $j$ over the trajectory $\boldsymbol{x}_0^t$, i.e. the empirical transition probability from $i$ to $j$. To argue this fact rigorously, one can compute the log-likelihood from (5.1) (implicitly ensuring nonnegativity), introduce Lagrange multipliers to enforce the linear constraints on each row of $\widehat{G}$, and take derivatives [64]. However, we feel that this approach is best understood as minimizing a KL divergence (see Lemma 3.1 in [65]). This perspective will be explained in what follows.

The factors of the likelihood given in (5.9) involving transitions out of state $i$ can be interpreted as the likelihood that over $n_{i:}(\boldsymbol{x}_0^t)$ independent and identically distributed samples of a random variable $\mathsf{j}_i$, which denotes the next state of the Markov chain whenever it is in state $i$, $\mathsf{j}_i$ assumes the value $j$ a total of $n_{ij}(\boldsymbol{x}_0^t)$ times. The true distribution of the random variable $\mathsf{j}_i$ would be determined by the probability vector $\boldsymbol{g}_i$. In observing the trajectory $\boldsymbol{x}_0^t$, we obtain an empirical distribution for the random variable $\mathsf{j}_i$, namely

$$\widehat{p}_{\mathsf{j}_i}(j; \boldsymbol{x}_0^t) = \frac{n_{ij}(\boldsymbol{x}_0^t)}{n_{i:}(\boldsymbol{x}_0^t)} \quad , \tag{5.11}$$

that is, the fraction of observed transitions out of $i$ that go immediately to $j$. The notation $\widehat{p}_{\mathsf{x}}(\cdot; \mathbf{y})$ denotes an empirical PMF for $\mathsf{x}$ as obtained from observing $\mathbf{y}$. The empirical distribution expressed as $\widehat{p}_{\mathsf{j}_i}(\cdot; \mathbf{x}_0^t)$ is a random function that depends on the observed trajectory $\mathbf{x}_0^t$. Note that the random variable $\mathsf{j}_i$ has nothing in common with the influencer random variables $\mathrm{i}_j$ of the GIM of Section 4.1.

From Lemma 3.1 of [65], the ML estimate $\widehat{\boldsymbol{g}}_i(\boldsymbol{x}_0^t)$ of (5.9) can be equivalently determined from the empirical distribution $\widehat{p}_{\mathrm{j}_i}(j;\boldsymbol{x}_0^t)$ as follows:

$$\widehat{\boldsymbol{g}}_i(\boldsymbol{x}_0^t) = \arg\min_{q(\cdot)\in\Delta_\eta} D(\,\widehat{p}_{\mathrm{j}_i}(\cdot;\boldsymbol{x}_0^t)\,||\,q(\cdot)\,) \quad , \tag{5.12}$$

where $D(\,p(\cdot)\,||\,q(\cdot)\,)$ is the KL divergence between distributions $p(\cdot)$ and $q(\cdot)$ [49]. We will shortly derive (5.12) in detail.

Minimizing the KL divergence as in (5.12) with respect to its second element is referred to as the *reverse I-projection* of $\widehat{p}_{\mathrm{j}_i}(\cdot;\boldsymbol{x}_0^t)$ onto $\Delta_\eta$ [65]. This offers us a geometric perspective to ML estimation. One begins by computing the empirical distribution from the observed trajectory $\boldsymbol{x}_0^t$ for the next state of the Markov chain whenever it is in state $i$, i.e., $\widehat{p}_{\mathrm{j}_i}(\cdot;\boldsymbol{x}_0^t)$. The ML parameter estimates of the transition probabilities out of state $i$ are given by the reverse I-projection of the empirical distribution onto the probability simplex, $\Delta_\eta$.

At this point, we feel it is important to comment on the two ways of representing a discrete random variable's distribution: a probability vector $\boldsymbol{p}$ of length $\eta$, or a probability mass *function* (PMF) $p(\cdot)$ defined on an alphabet of size $\eta$. We will primarily employ probability vectors, but at times when it is necessary to discuss concepts that are most commonly defined in terms of PMFs, e.g., KL divergence (5.12), we will alternate between notations. Note that a probability vector will be bolded, while a PMF will not be bolded (as it is a scalar-valued function), and for a probability vector $\boldsymbol{p}$, its associated PMF will be represented as $p(\cdot)$ and expressed in terms of its probability vector as $p(j) = p_j$, where $p_j$ is the $j$th entry of $\boldsymbol{p}$.[1] With this warning, hopefully the occasional blurring of the distinctions between probability vectors and PMFs should be straightforward to follow, and will not confuse the reader when our notation requires implicit conversions between PMFs and probability vectors, such as in (5.12), when a PMF on the right hand side is assigned to a probability vector on the left hand side.

The geometric picture of KL divergence and reverse I-projections is made possible by the implicit cost imposed on parameter estimates by an objective function that maximizes the likelihood. This relationship is important to highlight, and by doing so, we will derive the equivalence between reverse I-projections and ML estimates derived from IID observations, Lemma 3.1 of [65].

As $\log(\cdot)$ is a strictly increasing function, the ML estimate for the $i$th row of $G$ must maximize the log of the likelihood (5.9), that is

$$\begin{aligned}\widehat{\boldsymbol{g}}_i &= \arg\max_{\boldsymbol{g}_i\in\Delta_\eta} \sum_j n_{ij}(\boldsymbol{x}_0^t)\log g_{ij} \\ &= \arg\max_{\boldsymbol{g}_i\in\Delta_\eta} \sum_j \widehat{p}_{\mathrm{j}_i}(\,j;\,\boldsymbol{x}_0^t)\log g_{ij} \\ &= \arg\min_{g(\cdot)\in\Delta_\eta} \mathbb{E}_{\widehat{p}_{\mathrm{j}_i}}\left[-\log g(\mathrm{j}_i)\right] \\ &= \arg\min_{g(\cdot)\in\Delta_\eta} \mathbb{E}_{\widehat{p}_{\mathrm{j}_i}}\left[C_{LL}(\mathrm{j}_i, g(\cdot))\right] \quad ,\end{aligned} \tag{5.13}$$

[1] The alphabet of a finite valued random variable that assumes $\eta$ possible values is always assumed to be the integers from 1 to $\eta$. This convention eliminates the complications one would otherwise encounter as the probability vector $\boldsymbol{p}$ would be indexed by a set not equal to the domain on which the associated PMF $p(\cdot)$ is defined.

where in the third equality, we have switched our maximization over probability vectors to a minimization over PMFs, with our dummy variable changing from the probability vector $\boldsymbol{g}_i$ to the PMF $g(\cdot)$. The final equality of (5.13) equates our objective to determining the distribution that minimizes the expected log-loss cost $C_{LL}(\mathsf{j}, q(\cdot))$ defined as

$$C_{LL}(\mathsf{j}, q(\cdot)) \triangleq -\log q(\mathsf{j}) \quad . \tag{5.14}$$

The interested reader is encouraged to consult [66] for a complete explanation and treatment of cost criteria for distributions. The idea of such cost criteria is to specify the penalty for using a distribution $q(\cdot)$ to represent a random variable $\mathsf{j}$. In an information theoretic sense, the expected log-loss cost, $\mathbb{E}\left[C_{LL}(\mathsf{j}, q(\cdot))\right]$, computed using a base 2 logarithm, is the expected number of bits necessary to represent a random variable $\mathsf{j}$, when using the code designed for a random variable with distribution $q(\cdot)$ [47]. Intuitively, it should be evident that such a cost is minimized when using the code designed for the underlying random variable's distribution (this is half of Gibbs' inequality [47], a fundamental result in information theory [47]), and hence we can always consider the nonnegative quantity that is the expected number of *additional* bits necessary to represent the random variable $\mathsf{j}$ when using the code designed for a random variable with distribution $q(\cdot)$, instead of $\mathsf{j}$'s underlying distribution, that is,

$$\begin{aligned}\mathbb{E}_{p_{\mathsf{j}}}\left[C_{LL}(\mathsf{j}, q(\cdot)) - C_{LL}(\mathsf{j}, p_{\mathsf{j}}(\cdot))\right] &= \sum_{j} p_{\mathsf{j}}(j)\Big(\log p_{\mathsf{j}}(j) - \log q(j)\Big) \\ &\triangleq D(\, p_{\mathsf{j}}(\cdot)\,||\,q(\cdot)\,) \quad ,\end{aligned} \tag{5.15}$$

which is the definition of the KL divergence [49]. The KL divergence is obviously nonnegative, and Gibbs' inequality further ensures that $D(\,p(\cdot)\,||\,q(\cdot)\,) = 0$ if and only if $p(\cdot) = q(\cdot)$. However, the KL divergence is not a metric. In general it is not symmetric in its arguments, i.e., $D(\,p(\cdot)\,||\,q(\cdot)\,) \neq D(\,p(\cdot)\,||\,q(\cdot)\,)$, although it is roughly symmetric in its arguments when $p(\cdot)$ and $q(\cdot)$ are very close to one another. In some special circumstances the KL divergence does satisfy a triangular inequality, but in general it does not. It is often referred to as a measure of distance [67], an example of the larger class of $f$-divergences [68, 69]. Minimizing the KL divergence with respect to one of its elements is called a projection, and in our case, a reverse I-projection when the minimization is with respect to the second element.

By combining (5.13) with (5.15), it should be clear that ML estimates can be obtained by a reverse I-projection (5.12). Summarizing, ML estimation imposes a log-loss cost criterion, and consequently, information geometry offers an intuitive perspective into how ML estimates are obtained: as the reverse I-projection of the measured empirical distribution onto probability simplex $\Delta_\eta$. In our case, as our feasibility set $\Delta_\eta$ is otherwise unconstrained, such a reverse I-projection is degenerate: the KL divergence between the measured empirical distribution $\widehat{p}_{\mathsf{j}_i}(\cdot; \boldsymbol{x}_0^t)$ and a distribution in $\Delta_\eta$ can be minimized to 0, yielding the observed empirical distribution as the ML parameter estimates:

$$\begin{aligned}\widehat{g}_{ij}(\boldsymbol{x}_0^t) &= \widehat{p}_{\mathsf{j}_i}(j; \boldsymbol{x}_0^t) \\ &= \frac{n_{ij}(\boldsymbol{x}_0^t)}{n_{i:}(\boldsymbol{x}_0^t)} \quad \text{for all } i,\, j \quad ,\end{aligned} \tag{5.16}$$

where $\widehat{g}_{ij}(\boldsymbol{x}_0^t)$ is the ML estimate for the $j$th element of the probability vector $\boldsymbol{g}_i$ (5.8). Consequently, the intuitive ML estimates suggested in (5.10) are valid.

### 5.1.1 Alternate perspective on ML parameter estimates

The notion of deriving the ML estimates of $G$ via empirical distributions can be extended further to provide an alternate perspective in terms of linear least-mean-squares estimation. This new perspective motivates our suggested procedure for efficient parameter learning under the assumption of separability.

As in the original development of ML parameter estimates for Markov chains, an empirical distribution is defined via the observed trajectory $\boldsymbol{x}_0^t$. For each state $i$, we defined an empirical distribution for $\mathsf{j}_i$, the next state of the Markov chain when its current state is $i$. Each of these empirical distributions can be thought of as a *conditional* empirical distribution, given that the Markov chain is in some state $i$. Thus the subscript on $\mathsf{j}_i$ denotes the conditioning event, and upon dropping the subscript, we can use the trajectory $\boldsymbol{x}_0^t$ to define a joint empirical distribution for random variables $\mathsf{i}$ and $\mathsf{j}$ as follows:

$$\widehat{p}_{\mathsf{i},\mathsf{j}}(i,j;\mathbf{x}_0^t) = \frac{\#\text{ times } \tau : \mathsf{x}[\tau]=i,\ \mathsf{x}[\tau+1]=j}{t} \tag{5.17}$$

$$= \frac{1}{t}\sum_{\tau=0}^{t-1} 1_{\{\mathsf{x}[\tau]=i,\, \mathsf{x}[\tau+1]=j\}} \quad , \tag{5.18}$$

where we have employed standard indicator random variable notation, with $1_A$ being the indicator random variable for the event $A$. The idea is that each transition of the Markov chain provides a realized sample of the pair of random variables $(\mathsf{i},\mathsf{j})$. Note that $\widehat{p}_{\mathsf{j}|\mathsf{i}}(\cdot|\mathsf{i}=i;\boldsymbol{x}_0^t)$ derived from (5.17) is equal to $\widehat{p}_{\mathsf{j}_i}(\cdot;\boldsymbol{x}_0^t)$ of (5.11). We will henceforth represent $\widehat{p}_{\mathsf{j}_i}$ using the clearer conditional probability notation $\widehat{p}_{\mathsf{j}|\mathsf{i}}(\cdot|\mathsf{i}=i;\boldsymbol{x}_0^t)$.

To introduce linear least-mean-squares estimation, we define the random indicator vectors corresponding to $(\mathsf{i},\mathsf{j})$:

$$\mathsf{s}_\mathsf{i} = \boldsymbol{e}_i \quad \text{when } \mathsf{i} = i \tag{5.19}$$

$$\mathsf{s}_\mathsf{j} = \boldsymbol{e}_j \quad \text{when } \mathsf{j} = j \quad . \tag{5.20}$$

In an ergodic Markov chain, the conditional expectation of $\mathsf{s}_\mathsf{j}'$ given $\mathsf{i} = i$ with respect to the true joint distribution of $(\mathsf{i},\mathsf{j})$[2] is the $i$th row of $G$, i.e.,

$$\mathbb{E}\left[\mathsf{s}_\mathsf{j}' \,|\, \mathsf{i} = i\right] = \boldsymbol{e}_i' G \quad , \tag{5.21}$$

or equivalently,

$$\mathbb{E}\left[\mathsf{s}_\mathsf{j}' \,|\, \mathsf{s}_\mathsf{i}\right] = \mathsf{s}_\mathsf{i}' G \quad . \tag{5.22}$$

Following directly from (5.22), it is evident that the Bayesian least-mean-squares (BLS) estimator for $\mathsf{s}_\mathsf{j}$ given $\mathsf{s}_\mathsf{i}$ [48] is

$$\widehat{\mathsf{s}}_{\mathsf{j}\,BLS}'(\mathsf{s}_\mathsf{i}) = \mathsf{s}_\mathsf{i}' G \ , \tag{5.23}$$

and because the BLS estimate is linear, the linear least-mean-squares (LLS) estimator for

[2] The true distribution of $(\mathsf{i},\mathsf{j})$ should be thought of as the distribution of the random variables when the Markov chain is in steady-state, or equivalently, the limiting empirical distribution obtained as $t \to \infty$ (such a limiting distribution exists by renewal theory under the assumption of egodicity [40]). Hence, the true distribution is defined as $p_{\mathsf{i},\mathsf{j}}(i,j) = \overline{\pi}_i g_{ij}$, where $\overline{\pi}_i$ is the steady-state probability of state $i$.

$\mathsf{s_j}$ given $\mathsf{s_i}$ is

$$\widehat{\mathsf{s}}_{\mathsf{j}\,LLS}'(\mathsf{s_i}) \triangleq \mathsf{s_i}' R_{\mathsf{s_i s_i}}^{-1} R_{\mathsf{s_i s_j}} \tag{5.24}$$

$$= \widehat{\mathsf{s}}_{\mathsf{j}\,BLS}'(\mathsf{s_i}) = \mathsf{s_i}' G \quad , \tag{5.25}$$

where $R_{\mathbf{yz}} \triangleq \mathbb{E}\left[\mathbf{y}\mathbf{z}'\right]$ is the cross correlation of $\mathbf{y}$ and $\mathbf{z}$. Note that we mean linear in the strict sense, i.e., not the more familiar *affine* linear least-means-squares estimator of [48]. Hence it should be apparent why we introduced indicator vectors $\mathsf{s_i}$ and $\mathsf{s_j}$—because the transition matrix $G$ is part of the linear least-mean-squares estimate of $\mathsf{s_j}$ given $\mathsf{s_i}$.

Suppose that we substitute the empirical distribution $\widehat{p}_{\mathsf{i,j}}$ in place of the true distribution for our analysis in (5.21)–(5.25). It follows that

$$\widehat{\mathsf{s}}_{\mathsf{j}\,LLS(\widehat{p})}'(\mathsf{s_i}) = \mathsf{s_i}' \widehat{R}_{\mathsf{s_i s_i}}^{-1} \widehat{R}_{\mathsf{s_i s_j}} \tag{5.26}$$

where the expectations in $\widehat{R}_{\mathsf{s_i s_i}}$ and $\widehat{R}_{\mathsf{s_i s_j}}$ are computed with respect to the empirical distribution $\widehat{p}_{\mathsf{i,j}}(\cdot, \cdot; \boldsymbol{x}_0^t)$. It is straightforward to show that

$$\left[\widehat{R}_{\mathsf{s_i s_i}}\right]_{kl} = \begin{cases} \frac{1}{t} n_{k:}(\boldsymbol{x}_0^t) & \text{if } k = l \\ 0 & \text{otherwise} \end{cases}$$

$$\left[\widehat{R}_{\mathsf{s_i s_j}}\right]_{kl} = \frac{1}{t} n_{kl}(\boldsymbol{x}_0^t) \quad , \tag{5.27}$$

and thus by (5.27), we can rewrite our linear least-mean-squares estimate (5.26) as

$$\widehat{\mathsf{s}}_{\mathsf{j}\,LLS(\widehat{p})}'(\mathsf{s_i}) = \mathsf{s_i}' \widehat{G}(\boldsymbol{x}_0^t) \quad , \tag{5.28}$$

where $\widehat{G}(\boldsymbol{x}_0^t)$ is the ML estimate of $G$ based on observing the trajectory $\boldsymbol{x}_0^t$ as given in (5.10).[3]

Hence, we have an alternate method to compute the ML estimate for $G$, albeit by means convoluted in relation to frequency counting: compute the linear least-mean-squares estimate of $\mathsf{s_j}$ given $\mathsf{s_i}$ with respect to the empirical distribution $\widehat{p}_{\mathsf{i,j}}(\cdot, \cdot; \boldsymbol{x}_0^t)$ obtained via the observed trajectory $\boldsymbol{x}_0^t$. In doing so, we compute

$$\widehat{G}(\boldsymbol{x}_0^t) = \widehat{R}_{\mathsf{s_i s_i}}^{-1} \widehat{R}_{\mathsf{s_i s_j}} \quad , \tag{5.29}$$

which holds by (5.28) as $\mathsf{s_i}$ ranges over all of $\mathbb{R}^{\eta}$. We refer to this indirect method of computing the ML estimate of $G$ as the linear least-mean-squares approach.

A cautious reader will note that $\widehat{R}_{\mathsf{s_i s_i}}$ will be invertible if and only if the trajectory $\boldsymbol{x}_0^t$ contains a transition out of every state. However, this limitation is easily overcome by taking a Moore-Penrose generalized inverse [39] in (5.29). In doing so, the transition probability estimates out of the unobserved states are set at 0, and should be interpreted as arbitrary. Similarly, the estimates of the transition probabilities $\widehat{g}_{ij}$ as obtained via frequency counting when there are no observed transitions out of state $i$ are not uniquely defined.

[3] The astute reader will note that such a linear estimate developed from an empirical distribution can be directly computed via linear least squares regression; however, we feel that the random variable perspective via empirical distributions is useful as a comparison to the KL divergence perspective for ML estimates.

### 5.1.2 Resolving the equivalence of the two approaches

It may seem surprising that linear least-mean-squares estimation with respect to empirical distributions provides an alternate means of computing the ML estimate of $G$. Probing why this is so will provide the reader with a valuable perspective on various approaches to estimation in general. The equivalence can best be understood as linear least-mean-squares estimation imposing a quadratic cost criterion. The fact that both the log-loss and quadratic cost criteria are *proper*, a property that we will soon define, ensures that both approaches, reverse I-projection and linear least-mean-squares estimation, yield the ML parameter estimates of $G$.

Recall from (5.13) that the direct approach to ML parameter estimation of Markov chains, one minimizes the expected log-loss cost:

$$\widehat{g}_i(\boldsymbol{x}_0^t) = \arg\min_{q(\cdot)\in\Delta_\eta} \mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[C_{LL}(\mathsf{j}, q(\cdot)) \,|\, \mathsf{i} = i\right]. \tag{5.30}$$

As suggested already, we can envision the linear least-mean-squares approach as minimizing an expected cost with a different cost criterion. In determining the linear least-mean-squares estimator of $\mathsf{s}_\mathsf{j}$ as a function of $\mathsf{s}_\mathsf{i}$ under the empirical distribution $\widehat{p}_{\mathsf{i},\mathsf{j}}(\cdot,\cdot;\boldsymbol{x}_0^t)$, we obtain:

$$\widehat{\mathsf{s}}'_{\mathsf{j}LLS(\widehat{p})}(\mathsf{s}_\mathsf{i}) = \mathsf{s}'_\mathsf{i}\widehat{G}(\boldsymbol{x}_0^t) \quad , \tag{5.31}$$

where $\widehat{G}(\boldsymbol{x}_0^t)$ is the estimate of $G$ via the least-mean-squares approach (5.29). Because $\mathsf{s}_\mathsf{i}$ is an indicator vector, $\widehat{\boldsymbol{g}}_i$, the $i$th row of $\widehat{G}$, is precisely the Bayes least-mean-squares estimate of $\mathsf{s}_\mathsf{j}$ given that $\mathsf{s}_\mathsf{i} = \boldsymbol{e}_i$ under the empirical distribution $\widehat{p}_{\mathsf{i},\mathsf{j}}(\cdot,\cdot;\boldsymbol{x}_0^t)$, that is,

$$\begin{aligned}\widehat{\boldsymbol{g}}_i(\boldsymbol{x}_0^t) &= \widehat{\mathsf{s}}'_{\mathsf{j}LLS(\widehat{p})}(\boldsymbol{e}_i)\\ &= \widehat{\mathsf{s}}'_{\mathsf{j}BLS(\widehat{p})}(\boldsymbol{e}_i)\\ &= \arg\min_{\boldsymbol{g}\in\Delta_\eta} \|\mathsf{s}_\mathsf{j} - \boldsymbol{g} \,|\, \mathsf{i} = i\,\|_2^2\\ &= \arg\min_{\boldsymbol{g}\in\Delta_\eta} \mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[(\mathsf{s}_\mathsf{j} - \boldsymbol{g})'(\mathsf{s}_\mathsf{j} - \boldsymbol{g}) \,|\, \mathsf{i} = i\right]\\ &= \mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[\mathsf{s}_\mathsf{j} \,|\, \mathsf{i} = i\right] \quad .\end{aligned} \tag{5.32}$$

The notation $\|\cdot\|_2$ denotes the $\ell_2$-norm for random vectors, defined as $\|\mathsf{x}\|_2 = \mathbb{E}\left[\mathsf{x}'\mathsf{x}\right]^{1/2}$. We can define a quadratic cost criterion $C_Q(\mathsf{j}, q(\cdot))$ [66] as

$$C_Q(\mathsf{j}, q(\cdot)\,) \triangleq \sum_j (1_{\{\mathsf{j}=j\}} - q(j))^2 \quad . \tag{5.33}$$

Recognizing the equivalence between the quadratic cost in (5.33) and the expression inside the expectation in the fourth equality of (5.32), evidently the estimate of $G$ via the linear least-mean-squares approach is minimizing an expected quadratic cost:

$$\widehat{\boldsymbol{g}}_i(\boldsymbol{x}_0^t) = \arg\min_{g(\cdot)\in\Delta_\eta} \mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[C_Q(\mathsf{j}, g(\cdot)\,)|\mathsf{i} = i\right] \quad . \tag{5.34}$$

We can also illustrate our least-mean-squares estimate of the $i$th row of $G$ without the overhead of empirical distributions, as

$$\widehat{\boldsymbol{g}}_i(\boldsymbol{x}_0^t) = \arg\min_{\boldsymbol{g}\in\Delta_\eta} \sum_{\tau:\, x[\tau-1]=i} \| \boldsymbol{e}_{x[\tau]} - \boldsymbol{g} \|_2^2 \quad . \tag{5.35}$$

We should note here that in all such derivations, there is an implicit dependence on the observed trajectory $\boldsymbol{x}_0^t$ that we have omitted in most cases for notational simplicity.

Notice how (5.30) and (5.34) are the equivalent, except for the particular cost criteria being employed. When the feasibility set is otherwise unconstrained, the distribution that minimizes the expected cost in both cases is the observed empirical conditional distribution of $\mathsf{j}$ given $\mathsf{i} = i$:

$$\begin{aligned}\arg\min_{g(\cdot)\in\Delta_\eta} \mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[C_{LL}(\mathsf{j}, g(\cdot)) \,|\, \mathsf{i} = i\right] &= \arg\min_{g(\cdot)\in\Delta_\eta} \mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[C_{Q}(\mathsf{j}, g(\cdot)) \,|\, \mathsf{i} = i\right] \\ &= \widehat{p}_{\mathsf{j}|\mathsf{i}}(\cdot \,|\, i;\, \boldsymbol{x}_0^t),\end{aligned} \tag{5.36}$$

that is to say, both cost criteria are *proper*. Gibbs' inequality assures that the log-loss cost criteria is proper, and the quadratic cost criteria is proper as it is associated with a metric. This fact explains why the ML estimates are equivalent to the estimates obtained under the least-mean-squares approach, which we will henceforth refer to as LS estimates.

Evidently, we can envision our parameter estimates of the $i$th row of $G$, $\widehat{\boldsymbol{g}}_i$, as obtained by minimizing the expected log-loss cost (5.30) with a reverse I-projection, or by minimizing the expected quadratic cost (5.34) with a Euclidian projection. The ML estimation problem motivates the use of expected log-loss costs; however, when the minimization is performed over all distributions, any proper cost function can be considered, and thus both approaches are consistent. If there are constraints known a priori on the parameter space $\Theta$, only the log-loss criterion would necessarily yield the ML estimate.

As this investigation has shown, ML estimation motivates the use of log-loss costs. Moreover, there is an axiomatic defense of log-loss cost criteria when dealing with nonnegative quantities like probability distributions [67].[4] However, if one were to consider the problem without posing the objective of maximizing the likelihood, one may argue that it may not be immediately obvious that log-loss costs are preferred. Arguably the primary value of ML estimation is that the ML estimate for our ergodic Markov chain will converge to the true $G$ as the duration of the observed trajectory increases, i.e., it is an asymptotically consistent estimate. However, the same holds true for an estimate derived via any proper cost criteria, even in the case when $\Theta$ is restricted. As we will show in the next section, an estimate derived from quadratic cost criterion often may be preferable when it can be solved via a quadratic program or potentially even better, a closed form solution. Because estimates derived from quadratic cost criterion share the same asymptotic properties as the ML estimate, we argue that it can be a reasonable alternative to ML estimation.

## 5.2 Learning $G$ under $M_*$-separability

We now consider the parameter learning problem under $M_*$-separability. It is assumed that an $M_*$-separable network of stochastic automata produces the observed trajectory $\boldsymbol{x}_0^t$ of the network state, with $M_*$ being known. As we have a time-homogeneous Markov

[4] In constrained cases, minimizing an expected quadratic cost criterion may yield solutions with negative entries.

chain, the parameters to be estimated include the joint transition matrix $G$ and the initial probability vector $\boldsymbol{\sigma}$. Much as in the unrestricted case, our focus will be on estimating $G$. As discussed at length in Section 3.4, the transition matrix of a network exhibiting $M_*$-separability can be represented by a number of parameters equal to the dimension of $\mathcal{G}(M_*)$, i.e., the number of free parameters of the linear transformed $G$ (3.3). However, it is seemingly easier to estimate $G$ in our original untransformed space by imposing its simple nonnegativity constraints with the linear constraints of $M_*$-separability.

When $M_*$-separability is assumed, we compute the ML estimate of $G$ as

$$\widehat{G}_{ML} = \underset{G \in \mathcal{G}(M_*)}{\arg\max} \prod_{i,j} g_{ij}^{n_{ij}(\boldsymbol{x}_0^t)} \quad , \tag{5.37}$$

with $\mathcal{G}(M_*)$ being defined by a collection of linear equality and inequality constraints. This computation of the ML estimate of $G$ reduces to solving the following convex program in terms of the log likelihood:

$$\begin{aligned} \max & \sum_{i,j} n_{ij}(\boldsymbol{x}_0^t) \log g_{ij} \\ s.t. \;\; & G\mathbb{1} = \mathbb{1} \\ & \left[M_*^{\perp}\right]' G M_* = 0 \\ & G \geq 0 \quad , \end{aligned} \tag{5.38}$$

where $G \geq 0$ denotes that each entry of $G$ must be nonnegative. Because of the linear constraint $\left[M_*^{\perp}\right]' G M_* = 0$, the estimates of each row of $G$ do not decouple as they do in the otherwise unrestricted Markovian case.

The objective, which is maximizing a likelihood, can be manipulated into an expression that minimizes a convex combination of expected log-loss costs:

$$\begin{aligned} \widehat{G}_{ML}(\boldsymbol{x}_0^t) &= \underset{G \in \mathcal{G}(M_*)}{\arg\max} \sum_{i,j} n_{ij}(\boldsymbol{x}_0^t) \log g_{ij} \\ &= \underset{G \in \mathcal{G}(M_*)}{\arg\max} \sum_{i} \left( \sum_{j} n_{ij}(\boldsymbol{x}_0^t) \log g_{ij} \right) \\ &= \underset{G \in \mathcal{G}(M_*)}{\arg\max} \frac{1}{t} \sum_{i} n_{i:}(\boldsymbol{x}_0^t) \left( \sum_{j} \frac{n_{ij}(\boldsymbol{x}_0^t)}{n_{i:}(\boldsymbol{x}_0^t)} \log g_{ij} \right) \\ &= \underset{G \in \mathcal{G}(M_*)}{\arg\max} \sum_{i} \widehat{p}_{\mathsf{i}}(i; \boldsymbol{x}_0^t) \left( \sum_{j} \widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i; \boldsymbol{x}_0^t) \log g_{ij} \right) \\ &= \underset{G \in \mathcal{G}(M_*)}{\arg\min} \sum_{i} \widehat{p}_{\mathsf{i}}(i; \boldsymbol{x}_0^t) \left( \mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}} \left[ C_{LL}(\mathsf{j}, g_i(\cdot)) | \, \mathsf{i} = i \right] \right) \quad , \end{aligned} \tag{5.39}$$

where in the last line, $g_i(\cdot)$ must be conceived as the PMF associated with the probability vector $\boldsymbol{g}_i$, the $i$th row of $G$.

For those with an intuitive understanding of KL divergences, it may be helpful to manipulate the penultimate equality of (5.39) into a convex sum of KL divergences,

$$\begin{aligned}\widehat{G}_{ML}(\boldsymbol{x}_0^t) &= \operatorname*{arg\,max}_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t)\Bigg( \sum_j \widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i;\boldsymbol{x}_0^t)\log g_{ij} - \widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i;\boldsymbol{x}_0^t)\log \widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i;\boldsymbol{x}_0^t)\Bigg) \\ &= \operatorname*{arg\,min}_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t) D(\,\widehat{p}_{\mathsf{j}|\mathsf{i}}(\,\cdot\,|\,i;\boldsymbol{x}_0^t)\,||\,\boldsymbol{g}_i\,) \qquad (5.40)\end{aligned}$$

where several constant terms in the form of entropies are introduced in the first equality to manipulate the expression to the form of a KL divergence. The probability vector $\boldsymbol{g}_i$ that appears in the final equality should be thought of as a probability mass function.

We can derive an expression for the ML estimate solely in terms of a reverse I-projection, but only after first defining the matrix

$$\widetilde{G} \triangleq \operatorname{diag}\Big(\widehat{p}_{\mathsf{i}}(1;\boldsymbol{x}_0^t), \widehat{p}_{\mathsf{i}}(2;\boldsymbol{x}_0^t), \ldots, \widehat{p}_{\mathsf{i}}(\eta;\boldsymbol{x}_0^t)\Big) G \quad , \qquad (5.41)$$

where diag$(\cdot)$ denotes a diagonal matrix with the given entries on the diagonal. Evidently, $G$ and $\widetilde{G}$ share a linear bijection. Note that $\widetilde{G}$ represents a joint distribution on two variables: all of its entries are nonnegative and sum to 1. The first equality of (5.40) can be manipulated into a reverse I-projection involving $\widetilde{G}$:

$$\begin{aligned}\widehat{G}_{ML}(\boldsymbol{x}_0^t) &= \operatorname*{arg\,min}_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t)\left( \sum_j \widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i;\boldsymbol{x}_0^t)\log\left(\frac{\widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i;\boldsymbol{x}_0^t)\widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t)}{g_{ij}\widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t)}\right)\right) \\ &= \operatorname*{arg\,min}_{G\in\mathcal{G}(M_*)} \sum_i\sum_j \widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t)\widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i;\boldsymbol{x}_0^t)\log\left(\frac{\widehat{p}_{\mathsf{j}|\mathsf{i}}(j|i;\boldsymbol{x}_0^t)\widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t)}{g_{ij}\widehat{p}_{\mathsf{i}}(i;\boldsymbol{x}_0^t)}\right) \\ &= \operatorname*{arg\,min}_{G\in\mathcal{G}(M_*)} D\Big(\,\widehat{p}_{\mathsf{i},\mathsf{j}}(\cdot,\,\cdot;\,\boldsymbol{x}_0^t)\,\Big\|\,\widetilde{G}\,\Big) \quad , \qquad (5.42)\end{aligned}$$

where the matrix $\widetilde{G}$, defined in terms of $G$ in (5.41), must be envisioned as a joint PMF indexed by its rows followed by its columns. Operationally, the form of (5.42) offers the simplest method of deriving the ML estimates. Effectively it is a reverse I-projection with linear constraints. First determine

$$\widetilde{G}_{ML} = \operatorname*{arg\,min}_{\widetilde{G}\in\widetilde{\mathcal{G}}(M_*)} D\Big(\,\widehat{p}_{\mathsf{i},\mathsf{j}}(\cdot,\,\cdot;\,\boldsymbol{x}_0^t)\,\Big\|\,\widetilde{G}\,\Big) \qquad (5.43)$$

where the feasibility set $\widetilde{\mathcal{G}}(M_*)$ can be characterized by the linear constraints obtained by substituting (5.41) into the constraints characterizing $\mathcal{G}(M_*)$ as given in (5.38). Then compute

$$\widehat{G}_{ML}(\boldsymbol{x}_0^t) = \operatorname{diag}\left(\frac{1}{\widehat{p}_{\mathsf{i}}(1;\boldsymbol{x}_0^t)}, \frac{1}{\widehat{p}_{\mathsf{i}}(2;\boldsymbol{x}_0^t)}, \ldots, \frac{1}{\widehat{p}_{\mathsf{i}}(\eta;\boldsymbol{x}_0^t)}\right)\widetilde{G}_{ML} \quad . \qquad (5.44)$$

The form of (5.43), a reverse I-projection over a linear family, is a standard problem, yet this author is not aware of algorithms to solve such a problem efficiently apart from general 'hill-climbing' techniques, e.g., Newton-Ralphson. If the feasibility set was an exponential family, or if the optimization involved an I-projection over a linear family (not a reverse I-projection), an iterative alternating projections algorithm [70, 71, 72] could be employed, of which the efficient and ubiquitous EM algorithm is a special case [73].

### 5.2.1 A quadratic approach

Suppose that one estimates $G$ indirectly via a constrained Bayesian least-mean-squares estimation technique, the constrained counterpart to the approach outlined in Section 5.1.1. In particular, one estimates $G$ by developing the linear least-mean-squares estimate for the next state of the Markov chain given its current state, under the constraints that the operator satisfy the linear constraints of $M_*$-separability (5.38). In the language of quadratic costs, we wish to define our alternative estimate $\widehat{G}_{LS}$, which we refer to as the least squares estimate, as

$$\begin{aligned}\widehat{G}_{LS}(\boldsymbol{x}_0^t) &= \arg\min_{G\in\mathcal{G}(M_*)} \|\mathsf{s}_\mathsf{j}' - \mathsf{s}_\mathsf{i}'G\|_2^2 \\ &= \arg\min_{G\in\mathcal{G}(M_*)} \mathbb{E}_{\widehat{p}_{\mathsf{i},\mathsf{j}}}\left[(\mathsf{s}_\mathsf{j}' - \mathsf{s}_\mathsf{i}'G)(\mathsf{s}_\mathsf{j}' - \mathsf{s}_\mathsf{i}'G)'\right] \\ &= \arg\min_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_\mathsf{i}(i;\boldsymbol{x}_0^t)\mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[(\mathsf{s}_\mathsf{j}' - \mathsf{s}_\mathsf{i}'G)(\mathsf{s}_\mathsf{j}' - \mathsf{s}_\mathsf{i}'G)'|\mathsf{i} = i\right] \\ &= \arg\min_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_\mathsf{i}(i;\boldsymbol{x}_0^t)\|\mathsf{s}_\mathsf{j}' - \boldsymbol{g}_i|\mathsf{i} = i\|_2^2 \\ &= \arg\min_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_\mathsf{i}(i;\boldsymbol{x}_0^t)\mathbb{E}_{\widehat{p}_{\mathsf{j}|\mathsf{i}}}\left[C_Q(\mathsf{j},\boldsymbol{g}_i)|\mathsf{i} = i\right] \quad . \qquad (5.45)\end{aligned}$$

Comparing the last equality of (5.39) to the last equality of (5.45), the objective of minimizing a convex combination of log-loss costs has been transformed into a convex combination of quadratic costs.

As an alternative, one can compare the two approaches—the ML and LS estimates of $G$—on the basis of KL divergences and Euclidian ($\ell_2$) distances. The Cauchy-Schwartz equality for random vectors [50] assures that

$$\|\mathsf{s}_\mathsf{j}' - \widehat{p}_{\mathsf{j}|\mathsf{i}}\,|\mathsf{i} = i\|_2^2 + \|\widehat{p}_{\mathsf{j}|\mathsf{i}} - \boldsymbol{g}_i\,|\mathsf{i} = i\|_2^2 = \|\mathsf{s}_\mathsf{j}' - \boldsymbol{g}_i\,|\mathsf{i} = i\|_2^2 \quad , \qquad (5.46)$$

because $\mathbb{E}\left[\,\mathsf{s}_\mathsf{j}\,|\,\mathsf{i} = i\,\right] = \widehat{p}_{\mathsf{j}|\mathsf{i}}(\cdot|\,i\,)$. Substituting (5.46) into the penultimate equality of (5.45) and simplifying, we obtain

$$\begin{aligned}\widehat{G}_{LS}(\boldsymbol{x}_0^t) &= \arg\min_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_\mathsf{i}(i;\boldsymbol{x}_0^t)\left(\|\mathsf{s}_\mathsf{j}' - \widehat{p}_{\mathsf{j}|\mathsf{i}}\,|\mathsf{i} = i\|_2^2 + \|\widehat{p}_{\mathsf{j}|\mathsf{i}} - \boldsymbol{g}_i\,|\mathsf{i} = i\|_2^2\right) \\ &= \arg\min_{G\in\mathcal{G}(M_*)} \sum_i \widehat{p}_\mathsf{i}(i;\boldsymbol{x}_0^t)\|\widehat{p}_{\mathsf{j}|\mathsf{i}} - \boldsymbol{g}_i\,|\mathsf{i} = i\|_2^2 \quad . \qquad (5.47)\end{aligned}$$

Upon comparing (5.40) to (5.47), it should be evident that by replacing the KL divergences in the objective function for our ML estimate of $G$ (5.40) with $\ell_2$ distances, we obtain the LS estimate of $G$.

There are fundamental reasons to favor log-loss costs and the associated KL divergences as the appropriate measure of distance for nonnegative quantities like distributions; these reasons extend beyond the setting of maximum likelihood estimation [67]. Nevertheless, for computational reasons alone, it is reasonable to consider quadratic costs. Solving (5.45) is a rather tractable convex program with a quadratic objective function. Moreover, as the goal is to estimate the true parameters on the basis that the estimator is *consistent*, meaning that it converges to the true value in some probabilistic sense when the underlying Markov chain is ergodic, either quadratic costs or log-loss costs suffice.

Our estimate of $G$ derived from minimizing a quadratic cost can be made even more tractable. If the nonnegativity constraint is relaxed when computing the LS estimate of $G$, what results is a linear least-mean-squares problem subject to linear equality constraints, which has a closed-form solution [74]. Moreover, an iterative updating form for the estimate when additional observations arrive is straightforward to derive. Provided that the true $G$ is in the interior of the nonnegative orthant, after observing a sufficiently long trajectory, such estimates would eventually satisfy the relaxed nonnegativity constraint.

## 5.3 Learning $H_*$ under $M_*$-separability

Up to this point, we have avoided discussing the complexity of the computations to estimate $G$ under the assumption of $M_*$-separability. Recall that the dimension of the feasibility space $\mathcal{G}(M_*)$ will be a constant fraction of the dimension of the feasibility space in the otherwise unrestricted case: $M_*$-separability enforces a block of 0s in the linear transformation of $G$ illustrated in (3.3), but this block of 0s cannot dominate the matrix, as the *sum* of the rows and columns of such a block of 0s must be $\eta$, the number of rows or columns of $G$. Consequently, the dimension of the feasibility space $\mathcal{G}(M_*)$ will grow exponentially with $n$, regardless of how $M_*$ evolves with $n$. Even regular $\mathcal{M}$-separability faces these issues of the exponential growth of $\mathcal{G}(\mathcal{M})$ as $n$ increases.

In Sections 3.5 and 3.6, the increasingly large relative dimension of $M^{(r)}$-separability and $\mathcal{M}^{(\overline{r})}$-separability as $n$ increases was deemed favorable, as it offers substantial relative generality in our probabilistic model for large $n$. When learning $G$, however, such exponential growth of the parameter space as $n$ increases is a hindrance. When $n$ is large, learning $G$ by either of the suggested approaches—maximum likelihood or least squares—will be intractable. From the perspective of learning $G$, there is no significant computational advantage to assuming $M_*$-separability when $n$ is large, which we remind the reader, is the scenario that motivates separability. This is not to say that learning in general will be intractable, or that the previous discussions of learning $G$ are useless; in fact, $\widehat{G}_{LS}$, the least squares approach to estimating $G$, motivates the subsequent approaches to learning.

We wish to develop a learning algorithm under the assumption of $M_*$-separability that can be tractable for large $n$. At the very least, this requires estimating a set of parameters that is polynomial in $n$. The logical choice is to estimate $H_*$, the matrix that propagates the partial information vector $\boldsymbol{\pi}_*$. Provided that $M_*$ has column rank $O(n^k)$ (which is satisfied by $M^{(r)}$ with $k = r$), then estimating $H_*$ requires estimating $O(n^k)$ parameters, and has the potential of being tractable. By nature of being a much smaller matrix, learning $H_*$ may be more robust than learning $G$. Although $H_*$ does not provide the full information that $G$ provides, it can be used to tractably propagate partial information when $n$ is large (provided $H_*$'s size is polynomial in $n$), which is the value to assuming $M_*$-separability. As discussed in Section 3.1.2, $H_*$ offers substantial insight into $G$'s spectral properties. Hopefully we have sufficiently motivated the idea of estimating $H_*$.

Before developing algorithms for estimating $H_*$, we need to characterize $\mathcal{H}(M_*)$, the set of possible $H_*$ that propagate partial information under $M_*$-separability, much as we did for $\mathcal{G}(M_*)$ in Section 3.2.

### 5.3.1 The feasibility set $\mathcal{H}(M_*)$

If we are to estimate $H_*$, we should understand the feasible set of $H_*$ that one may encounter under $M_*$-separability. In this section, we characterize the set of $H_*$ matrices that solve (3.1) for fixed $M_*$ and any row-stochastic $G$, a set that will be denoted as $\mathcal{H}(M_*)$. From an alternate perspective, $\mathcal{H}(M_*)$ consists of the matrices $H_*$ that propagate partial information when $M_*$-separability is satisfied.

When $M_*$ has full column rank, $\mathcal{H}(M_*)$ can be defined as the image of a function with $\mathcal{G}(M_*)$ as its domain:

$$\mathcal{H}(M_*) = \left\{M_*^{-L} G M_* : G \in \mathcal{G}(M_*)\right\} \quad . \tag{5.48}$$

As the function $M_*^{-L}GM_*$ is linear in $G$, $\mathcal{H}(M_*)$ must be finitely-generated whenever $\mathcal{G}(M_*)$ is finitely-generated. As $\mathcal{G}(M_*)$ is the intersection of a polyhedral cone [29] (the nonnegative orthant) with an affine subspace and in addition is bounded, it must be a finitely-generated convex set, and consequently $\mathcal{H}(M_*)$ must also be a finitely-generated convex set. Note that $I \in \mathcal{H}(M_*)$, just as $I \in \mathcal{G}(M_*)$, for all $M_*$. The linear constraints satisfied by matrices in $\mathcal{G}(M_*)$ can be converted to constraints on $\mathcal{H}(M_*)$ by elimination. Similarly, the cone membership condition (nonnegativity) can be transformed into a cone membership condition in the transformed space. This procedure would have the benefit of describing $\mathcal{H}(M_*)$ in terms of a set of linear equalities and inequalities that must be satisfied, which can then be used to verify membership of a matrix in $\mathcal{H}(M_*)$.

When $M_*$ fails to have full column rank, consisting of $\overline{p}$ columns with rank $p < \overline{p}$, characterizing $\mathcal{H}(M_*)$ in terms of $\mathcal{G}(M_*)$ is a bit more complicated. Because $M_*$ has a null space, for a given $G$, there is an affine subspace of matrices $H_*$ that satisfy (3.1). As has been discussed in Section 3.1.2, for a given $G$ exhibiting $M_*$-separability, one can solve (3.1) for the matrix $H_0$ of minimal Frobenius norm:

$$H_0 = M_*^{\dagger} G M_* \quad . \tag{5.49}$$

The affine subspace of solutions, for a given $M_*$-separable $G$, can then be specified as

$$H_*(Z) = H_0 + N_* Z \quad , \tag{5.50}$$

where $N_*$ is a matrix whose columns serve as a basis for $\mathcal{N}(M_*)$, and $Z \in \mathbb{R}^{(\overline{p}-p)\times\overline{p}}$ parameterizes the affine subspace.

If we define the subspace of $\overline{p} \times \overline{p}$ matrices with columns in $\mathcal{N}(M_*)$ as

$$\overline{\mathcal{N}}(M_*) = \left\{N_* Z : Z \in \mathbb{R}^{(\overline{p}-p)\times\overline{p}}\right\} \quad , \tag{5.51}$$

and define the finitely-generated convex set of possible minimal Frobenius norm matrices for any $G$ in $\mathcal{G}(M_*)$:

$$\mathcal{H}_0(M_*) = \left\{M_*^{\dagger} G M_* : G \in \mathcal{G}(M_*)\right\} \quad , \tag{5.52}$$

we can then express $\mathcal{H}(M_*)$ as

$$\mathcal{H}(M_*) = \mathcal{H}_0(M_*) \oplus \overline{\mathcal{N}}(M_*) \tag{5.53}$$

where $\oplus$ indicates a direct sum.

There also is a rather simple and intuitive way to characterize $\mathcal{H}(M_*)$, without invoking

$\mathcal{G}(M_*)$. Note (3.1), and reverse the typical roles of $G$ and $H_*$. Suppose that $M_*$ and $H_*$ are given. Under what conditions does there exist a row-stochastic matrix $G$ satisfying (3.1)? A row-stochastic matrix $G$ exists if and only if $H_*$ leaves the polytope defined as the convex hull of the rows of $M_*$ left-invariant. Why? Because by (3.1), the $i$th row of $M_*$ times $H_*$ must equal the $i$th row of $G$ multiplied by $M_*$, which is a convex sum of the rows of $M_*$. For $M^{(r)}$, we refer to such a polytope as the *$r$th-order marginal polytope.*

Sets defined in terms of the rows of $M_*$ and the convex hull of its rows were introduced informally in Section 3.3. Recall that $\mathcal{S}$ is the set of indicator vectors for the network (3.71). Let $\mathcal{S}_*$ be the image of $\mathcal{S}$ under the linear transformation given by $M_*$, i.e.,

$$\mathcal{S}_* = \left\{ \boldsymbol{e}_k' M_* : \boldsymbol{e}_k \in \mathcal{S} \right\} \quad . \tag{5.54}$$

In simple terms, $\mathcal{S}_*$ consists of the rows of $M_*$. Let $\Delta_*$ be the convex hull of $\mathcal{S}_*$, that is, the set of potential partial information vectors that may be encountered under $M_*$-separability. Such notation does not indicate the appropriate size of vectors in such sets, as it is assumed that the appropriate sizes can be ascertained from $M_*$.

### 5.3.2 Verifying membership in $\mathcal{H}(M_*)$

Recall that $\mathcal{H}(M_*)$ is the set of matrices that leaves $\Delta_*$ left-invariant. Determine the half space constraints and linear constraints defining $\Delta_*$. Represent these constraints with matrices $A$, $B$, and vectors $\boldsymbol{c}$, $\boldsymbol{d}$ as follows

$$\boldsymbol{m}'A = \boldsymbol{c}' \quad \forall \boldsymbol{m} \in \Delta_* \tag{5.55}$$

$$\boldsymbol{m}'B \geq \boldsymbol{d}' \quad \forall \boldsymbol{m} \in \Delta_* \quad . \tag{5.56}$$

It then follows that $H_* \in \mathcal{H}(M_*)$ if and only if

$$M_* H_* A = \mathbb{1}\boldsymbol{c}' \tag{5.57}$$

$$M_* H_* B \geq \mathbb{1}\boldsymbol{d}' \quad . \tag{5.58}$$

We can make a few important observations. Suppose that $M_*$ has $\overline{p}$ columns, and rank $p \leq \overline{p}$. Naturally, $\Delta_* \subset \mathbb{R}^{\overline{p}}$. There are only two possibilities for the dimension of $\Delta_*$: it must be either $p$ or $p-1$ (depending on whether or not $\mathbb{1} \in \mathcal{R}(M_*)$), meaning that $A$ must have either $\overline{p} - p$ or $\overline{p} - p + 1$ columns, i.e., $O(\overline{p})$ columns. As $\overline{p} - p$ of the constraints on $\Delta_*$ must be strict linear equality constraints stemming from the null space of $M_*$, we can always choose $A$ and $\boldsymbol{c}$ such that $\boldsymbol{c}$ consists of all 0s with the possible exception of a single 1. Only when $\mathbb{1} \in \mathcal{R}(M_*)$ is it possible for $\boldsymbol{c}$ to have a nonzero entry (otherwise we would have a contradiction in (5.57)).

If $M_*$ has full column rank, then there are no equality constraints due to $M_*$ having a null space. Consequently, $A$ will consist of a single column if $\mathbb{1} \in \mathcal{R}(M_*)$, and otherwise there will be no equality constraints characterizing $\mathcal{H}(M_*)$.

These observations have implications on verifying (5.57), which characterizes the affine hull of $\mathcal{H}(M_*)$. As $\Delta_*$ has dimension $p$ or $p-1$, one can choose $p+1$ (or in some cases only $p$) rows of $M_*$ (elements of $\mathcal{S}_*$) whose affine hull contains $\Delta_*$. Let $\widetilde{M_*}$ be the matrix with these $p+1$ rows of $M_*$, and note that verifying (5.57) for a candidate $H_*$ is equivalent to verifying

$$\widetilde{M_*} H_* A = \mathbb{1}\boldsymbol{c}' \quad . \tag{5.59}$$

As all matrices in (5.59) have no more than $\overline{p}$ rows/columns, evaluating such an expression requires computation of $O(\overline{p}^3)$ complexity, which may be tractable provided that $\overline{p}$ is polynomial in $n$ (assured for our canonical cases of $M^{(r)}$-separability). This does assume that $M_*$ is 'preprocessed,' meaning that $p+1$ rows have been identified whose affine hull contains $\Delta_*$. For our canonical examples, such processing is tractable because the rank of $M^{(r)}$ is analytically known in advance. Hence, verifying that a candidate $H_*$ satisfies the linear equality constraints of $\mathcal{H}(M_*)$ may well be tractable.

On the other hand, verifying the linear inequality constraints (5.58) may prove difficult. There are no obvious limits on the number of columns of $B$; furthermore, one cannot reduce $M_*$ to $\widetilde{M}_*$ when inequalities are involved. However, one could lessen the computational demands of checking membership in $\mathcal{H}(M_*)$ by replacing $M_*$ in (5.58) with a matrix consisting of only a subset of the rows of $M_*$, the rows that are the extreme points $\Delta_*$. A potential algorithm to accomplish this is given in [75]. If all $\eta$ rows of $M_*$ are extreme points, then the computational complexity of evaluating (5.58) will be $\Omega(\eta)$, i.e., at least exponential in $n$ (this seemingly weak lower bound considers the potential computational advantages stemming from sparsity). On the other hand, if $B$ has order $O(\overline{p})$ columns and $\Delta_*$ has $\overline{p}$ extreme points, then evaluating (5.58) is of $O(\overline{p}^3)$ complexity. Thus the computational bottleneck in checking membership in $\mathcal{H}(M_*)$ is the linear inequality constraints. At one extreme, the computational demands of checking membership in $\mathcal{H}(M_*)$ will be no worse than polynomial in $\overline{p}$, and at the other extreme, it will be at least exponential in $n$.

Recognizing that separability is defined with respect to a subspace (Corollary 3), one may think to choose $M_*$ judiciously to minimize the number of extreme points in the convex hull of its rows, thereby reducing the potential computation of checking membership in $\mathcal{H}(M_*)$. This is futile, as evident by linear algebra. For example, suppose that the $j$th row of $M_*$ can be expressed as a convex combination of the other rows of $M_*$, meaning that the $j$th row of $M_*$ is not an extreme point of $\Delta_*$. Define the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{j-1}, -1, \lambda_{j+1}, \ldots, \lambda_\eta)$, where for $i \neq j$, $\lambda_i \geq 0$ are the weights in the convex combination associated with the $j$th row of $M_*$. Evidently, $\boldsymbol{\lambda}' M_* = 0$, meaning that $\boldsymbol{\lambda}$ is in the left null space of $M_*$. Any other matrix $M_\#$ such that $\mathcal{R}(M_*) = \mathcal{R}(M_\#)$ must have $\boldsymbol{\lambda}$ in its left null space, and consequently, its $j$th row will not be an extreme point of the convex hull of its rows.

We note that for $M^{(r)}$-separability, each row of $M^{(r)}$ is an extreme point of $\Delta^{(r)}$. Consequently, checking the linear inequality constraints of $\mathcal{H}(M^{(r)})$ via (5.58) requires computations that are at least exponential in $n$. Even though the matrices in $\mathcal{H}(M^{(r)})$ are of size $O(n^r)$, they are the set of matrices that leave a convex polytope invariant that is generated by $\Theta(c^n)$ extreme points, which is makes checking membership in $\mathcal{H}(M_*)$ intractable for large $n$.

Note that the complexity of checking membership in $\mathcal{H}(M_*)$ may be very different from the complexity of checking membership in $\Delta_*$. For $r = 1$, membership of $\boldsymbol{\pi}^{(1)} \in \Delta^{(1)}$ is straightforward to check: each partition of $\boldsymbol{\pi}^{(1)}$ that corresponds to a probability vector must sum to 1 and be nonnegative, i.e., $O(n)$ constraints. When $r = 2$, the bivariate marginal probability vectors composing the vectors in $\Delta^{(2)}$ must be nonnegative and sum to 1, but in addition they must be *consistent*, i.e., the bivariate marginal for $(\mathsf{x}_1, \mathsf{x}_2)$ and the bivariate marginal for $(\mathsf{x}_2, \mathsf{x}_3)$ must both specify the same univariate marginal for $\mathsf{x}_2$. Each univariate marginal is implicitly defined in $O(n)$ bivariate marginals; consequently, verifying consistency is of $O(n^2)$ complexity when $r = 2$. For general $r$, we can argue that the computational complexity of checking consistency is $O(n^r)$, which includes checking consistency of the $(r-1)$-variate marginals, followed by the $(r-2)$-variate marginals, and so forth. What we have failed to show, however, is whether or not nonnegativity and consistency of marginals is sufficient for characterizing the $r$th-order marginal polytope $\Delta^{(r)}$.

For $r = 1$, the answer is yes, but for $r > 1$, the answer is no. When $r > 1$, we refer to the polytope of nonnegative vectors composed of a sequence of consistent marginal probability vectors as the *relaxed rth-order marginal polytope*, which is a strict superset of $\Delta^{(r)}$. When $r > 1$, there exist *pseudo* $r$th-order marginal vectors exhibiting consistency, but for which there does not exist a joint distribution that produces such consistent marginals.

The explanation as to why nonnegativity and consistency are not sufficient to characterize the $r$th-order marginal polytope for $r > 1$ is explained in Vorob'ev's and Kellerer's work on extension of consistent probability measures [42, 43, 44]. There is a graphical approach to check when consistency and nonnegativity are sufficient to characterize a marginal polytope, which can be explained as follows: associate a node which each marginal distribution given. Next, draw an edge connecting each pair of nodes that define a common lower-order marginal (and hence would need to be checked for consistency). If the resulting graph is a tree, i.e., no cycles, then nonnegativity and consistency are sufficient to characterize the marginal polytope. Obviously this explains why we have sufficiency for $r = 1$ (no edges would be drawn) but not for $r > 1$ (for any pair of connected nodes, there exists a third node connected to each that completes a 3-cycle). We note that this characterization of a *junction tree* is equivalent what we refer to as the *common orthonormal basis assumption*, which is discussed in Appendix A.2, and offers a linear algebraic perspective of such ideas.

Readers familiar with graphical models may note the similarities of our graphical construction to junction trees for Markov random fields [41]; in fact these ideas are equivalent. As shown in [76], when one has a junction tree of marginal information, a joint distribution producing such marginals can be specified as the product of the marginals associated with each node, divided by the common lower-order marginals associated with each edge. In the language of graphical models, these would be the clique marginals and separator marginals, respectively. The joint distribution defined in terms of these clique and separator marginals would the the joint distribution of maximum entropy that satisfies such marginals [70], and equivalently, the unique joint distribution that satisfies the given marginals and exhibits the conditional independence relationships encoded by the Markov random field that can be 'backed out' from the junction tree (each of the Markov random field's cliques, maximal fully-connected subsets of nodes, would correspond to a node in the junction tree) [76].

Summarizing our observations, it is evident that checking membership in $\mathcal{H}(M_*)$ will be of computational complexity that is polynomial in the number of extreme points of $\Delta_*$, meaning that for the canonical examples of separability that we often consider, $M^{(r)}$ or $\mathcal{M}^{(\overline{r})}$-separability, this complexity is exponential in $n$. On the other hand, checking membership in the the relaxed $r$th-order marginal polytope is only polynomial in $n$. This realization motivates the suggested approach of ignoring the inequality constraints of the feasibility set from which we estimate $H_*$, but then concerning ourselves with maintaining membership in $\Delta_*$ (or its relaxation) when propagating partial information vectors via a potentially infeasible estimate of $H_*$.

### 5.3.3 Approaches to estimating $H_*$

By estimating $H_*$, one is estimating linear functions of the transition probabilities of an $M_*$-separable Markov chain. The mapping from $G$ to $H_*$ is not injective, meaning that in general, $H_*$ cannot characterize $G$. In general, to associate a unique $G \in \mathcal{G}(M_*)$ with a given $H_* \in \mathcal{H}(M_*)$, one must know the values of the two additional matrices $H_{(12)}$ and $H_{(22)}$ represented in (3.3). In most instances, for a particular $H_*$ there is a set of transition matrices $G \in \mathcal{G}(M_*)$ all associated with the same $H_*$. This may create a problem for maximum likelihood estimation. One wishes to find the parameters that maximize a likelihood that is not fully defined in terms of the parameters one is considering. If we use (3.3) to express the likelihood (5.1) in terms of $H_*$, $H_{(12)}$, and $H_{(22)}$, we do not obtain a likelihood that factors into two terms, one involving $H_*$ and the other involving $H_{(12)}$ and $H_{(22)}$ (nor do the constraints decouple). This fact presents a challenge as to how to proceed with estimating $H_*$.

One could find the $H_*$, $H_{(12)}$, and $H_{(22)}$ that maximize the probability of the observed trajectory. Equivalently, one can compute $\widehat{G}_{ML}$ and then report

$$\widehat{H}_{ML} = M_*^{\dagger} \widehat{G}_{ML} M_* \quad . \tag{5.60}$$

As suggested by our notation, we refer to $\widehat{H}_{ML}$ in (5.60) as the ML estimate of $H$. However, such a computation would be undesirable, on account of having first to determine $\widehat{G}_{ML}$.

Alternatively, one may consider a Bayesian formulation, since a convex set of row-stochastic matrices $G$ may produce the same $H_*$. One could define a measure on $\mathcal{G}(M_*)$, and thereby implicity a measure on $\mathcal{H}(M_*)$.[5] One could compute the maximum a posteriori (MAP) estimate from such a formulation. Again, this would seemingly not be tractable for large $n$, as it would again require computations involving $G$.

These two approaches to estimating $H_*$ will not be computationally feasible when $n$, the number of automata, is large. Hence, we must develop an alternative approach to estimating $H_*$. Our motivation lies in the observation that under $M_*$-separability, $H_*$ is the matrix that propagates the expected value of the linear transformation of the state indicator vector, recall (2.32), i.e.,

$$\mathbb{E}\left[\, \mathsf{s}_*[t]' \,|\, \mathsf{s}_*[t-1] \,\right] = \mathsf{s}_*[t-1]' H_* \quad . \tag{5.61}$$

Similar to the approach in Section 5.1.1, define random variables $\mathsf{s}_{\mathsf{i}*}$ and $\mathsf{s}_{\mathsf{j}*}$ jointly distributed according to our time-homogeneous stochastic process as follows: .

$$p_{\mathsf{s}_{\mathsf{i}*},\mathsf{s}_{\mathsf{j}*}}(s_{*(1)},\ s_{*(2)}) \triangleq \mathbb{P}\left(\, \mathsf{s}_*[t-1] = s_{*(1)},\ \mathsf{s}_*[t] = s_{*(2)} \,\right) \tag{5.62}$$

for $s_{*(1)}$, $s_{*(1)} \in \mathcal{S}_*$. Such a joint probability distribution is well-defined whenever the the underlying Markov chain is ergodic and initialized in steady-state.

Evidently from (5.61),

$$\begin{aligned} \arg\min_{f(\cdot)\in\mathcal{B}} \|\mathsf{s}_{\mathsf{j}*}' - f(\mathsf{s}_{\mathsf{i}*}) \,|\, \mathsf{s}_{\mathsf{i}*} \|_2 &= \mathbb{E}\left[\, \mathsf{s}_{\mathsf{j}*}' \,|\, \mathsf{s}_{\mathsf{i}*} \,\right] \\ &= \mathsf{s}_{\mathsf{i}*}' H_* \end{aligned} \tag{5.63}$$

where in the first equality, $\mathcal{B}$ is the set of measurable functions. By the second equality in

[5] Recall that if $M_*$ fails to have full column rank, there is a subspace of matrices $H_*$ associated with a particular $G$ (3.24). To simplify matters, we only consider the $H_*$ matrices of minimal Frobenius norm (3.20), so as to always associate a unique $H_*$ with a given $G \in \mathcal{G}(M_*)$.

(5.63), it follows that if one were to restrict $\mathcal{B}$ to only linear functions, we would have

$$\begin{aligned} H_* &= \arg\min_{H \in \mathbb{R}^{p \times p}} \|\mathsf{s}_{\mathsf{j}*}' - \mathsf{s}_{\mathsf{i}*}' H\|_2 \\ &= R_{\mathsf{s}_{\mathsf{i}*}\mathsf{s}_{\mathsf{i}*}}^{-1} R_{\mathsf{s}_{\mathsf{i}*}\mathsf{s}_{\mathsf{j}*}} \quad , \end{aligned} \tag{5.64}$$

as long as $\mathcal{S}_*$ does not lie in a strict subspace of $\mathbb{R}^p$ (this assured when $M_*$ has full column rank). Our (5.64) states that $H_*$ can be obtained via a linear least-mean-squares minimization between $\mathsf{s}_{\mathsf{j}*}$ and linear functions of $\mathsf{s}_{\mathsf{i}*}$.

### Least squares estimator $\widehat{H}_{LS}$

At this moment, our approach to estimating $H_*$ should be clear. By observing a trajectory $\mathsf{x}_0^t$, we receive a sequence of samples of $\mathsf{s}_{\mathsf{j}*}$ and $\mathsf{s}_{\mathsf{i}*}$. Although the samples are not independent, as long as we have an ergodic chain, by renewal theory [40], the empirical distribution

$$\widehat{p}_{\mathsf{s}_{\mathsf{i}*},\mathsf{s}_{\mathsf{j}*}}(s_{*(1)}, \ s_{*(2)}; \ \mathsf{x}_0^t) \triangleq \frac{1}{t}\sum_{\tau=1}^{t} 1_{\{\mathsf{s}_*[\tau-1]=s_{*(1)}, \mathsf{s}_*[\tau]=s_{*(2)}\}} \tag{5.65}$$

will converge almost surely to the true distribution given in (5.62). It follows that the empirical cross correlation matrices computed from the empirical distribution will converge almost surely to the true cross correlation matrices. In doing so, we can compute an estimate of $H_*$ that will be strongly consistent, that is, will converge almost surely to the true $H_*$ as $t$, the length of the trajectory, increases.[6] This is summarized in the following theorem.

**Theorem 11.** *When $M_*$ has full column rank, define*

$$\begin{aligned} \widehat{H}_{LS}(\mathsf{x}_0^t) &\triangleq \arg\min_{H \in \mathbb{R}^{p \times p}} \sum_{\tau=1}^{t} \|\mathsf{s}_*[\tau]' - \mathsf{s}_*[\tau-1]' H\|_2^2 \\ &= \arg\min_{H \in \mathbb{R}^{p \times p}} \mathbb{E}_{\widehat{p}}\left[ (\mathsf{s}_{\mathsf{j}*}' - \mathsf{s}_{\mathsf{i}*}' H)(\mathsf{s}_{\mathsf{j}*}' - \mathsf{s}_{\mathsf{i}*}' H)' \right] \\ &= \widehat{R}_{\mathsf{s}_{\mathsf{i}*}\mathsf{s}_{\mathsf{i}*}}^{-1} \widehat{R}_{\mathsf{s}_{\mathsf{i}*}\mathsf{s}_{\mathsf{j}*}} \quad , \end{aligned} \tag{5.66}$$

*where the expectations in second and third equalities are with respect to the empirical distribution on $\mathsf{s}_{\mathsf{i}*}$ and $\mathsf{s}_{\mathsf{j}*}$ defined as in (5.65).*

*Provided that the underlying Markov chain's transition matrix $G$ is ergodic, by renewal theory, it follows that*

$$\widehat{H}_{LS}(\mathsf{x}_0^t) \to H_* \quad \textit{almost surely as } t \to \infty \tag{5.67}$$

*meaning that $\widehat{H}_{LS}(\mathsf{x}_0^t)$ is a strongly consistent estimator.*

We refer to $\widehat{H}_{LS}(\mathsf{x}_0^t)$ as the least squares estimator of $H_*$. Note that determining $\widehat{H}_{LS}(\mathsf{x}_0^t)$ requires computation of $O(p^3)$ complexity.

[6] In order for linear least mean squares estimates to be well conditioned, which is necessary as matrix inverses must be computed, the values of $\mathsf{s}_{\mathsf{i}*}$ should span $\mathbb{R}^p$. This can be assured with high probability for a sufficiently large number of samples when $M_*$ has full column rank $p$.

### Constrained least squares estimator $\widehat{H}_{\underline{LS}}$

One may think to improve our estimator by taking into consideration the linear equality constraints characterizing $\mathcal{H}(M_*)$. Recall that when $M_*$ has full column rank, there exist such linear equality constraints if and only if $\mathbb{1} \in \mathcal{R}(M_*)$. For this special case, define the set of matrices satisfying such linear equality constraints as

$$\mathcal{H}_{=}(M_*) \triangleq \{H : M_* H \boldsymbol{a} = \mathbb{1} c\} \quad , \tag{5.68}$$

with $\boldsymbol{a} = A$ and $c = \boldsymbol{c}$ from (5.55) (because $M_*$ has full column rank, $A$ has at most one column, which we represent as the vector $\boldsymbol{a}$ and $\boldsymbol{c}$ is a scalar, which we represent by $c$). Note that $\mathcal{H}_{=}(M_*)$ is the affine hull of $\mathcal{H}(M_*)$. We can then define an alternate estimator as

$$\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t) \triangleq \underset{H \in \mathcal{H}_{=}(M_*)}{\arg\min} \sum_{\tau=1}^{t} \left\| \mathsf{s}_*[\tau]' - \mathsf{s}_*[\tau - 1]' H \right\|_2^2 \quad . \tag{5.69}$$

As mentioned in the context of estimating $G$, linearly constrained linear least-mean-squares problems have closed-form solutions [74]. Because such linear equality constraints can be expressed in terms of matrices with $O(p)$ rows/columns, the computational complexity of computing constrained solutions is $O(p^3)$. By the same renewal theory arguments, the constrained estimator $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$ is also asymptotically consistent.

The reader may wonder: why a quadratic cost and not a log-loss cost? As explained in Section 5.2.1, when nonnegative quantities are concerned, log-loss costs and KL divergences are the only cost functions that satisfy some rather reasonable properties [67]. One may note that under $M^{(r)}$-separability, all vectors being propagated are indeed nonnegative (they are probability distributions). Apart from the complication that $M^{(r)}$ does not have full column rank, it would seem that KL divergences would be the preferred measure of distance to minimize when estimating $H^{(r)}$. However, should one attempt to pose the problem in terms of minimizing KL divergences, one will see that it is not possible to pose the problem as such, at least not involving a form analogous to (5.66). Should one condition on a particular value of $\mathsf{i}$, the cost function of (5.66) does not involve a single row of the $H$. This property was essential for expressing the ML estimate of $G$ as a minimization problem involving KL divergences, c.f. (5.39), (5.40). There are additional reasons for our departure from log-loss costs. Our definition of separability does not require that $\Delta_*$ be in the positive orthant. When $\Delta_* \not\subset \mathbb{R}_+^p$, log-loss costs will be undefined for nonpositive values.

The choice of quadratic costs is partially based on tractability: the estimate of $H_*$ in (5.66) is straightforward to compute. But moreover, should one not be restricted to considering nonnegative quantities, quadratic costs themselves are the only cost functions that satisfy some reasonable properties [67], the most important of which is being proper, which assures that our estimator of $H_*$ is strongly consistent.

## 5.4 Additional considerations for estimators

At this point, there are several lingering questions regarding $\widehat{H}_{LS}$ and $\widehat{H}_{\underline{LS}}$. For example, how does one estimate $H_*$ when $M_*$ does not have full column rank? Secondly, are there any guarantees that these estimators will be feasible for finite $t$? Thirdly, if one is uncertain as to whether or not the underlying Markov chain is $M_*$-separable, is there an efficient means of determining whether or not such an assumption is reasonable? These questions will be considered over the following sections.

### 5.4.1 Estimating $H_*$ when $M_*$ does not have full column rank

As was suggested at the end of Section 3.1.2, there is no disadvantage to tracking the dynamics of the partial information expressed in a space of minimal dimension. Thus, one should always consider $M_*$ with full column rank. Nevertheless, one may feel compelled to solve the linear least-mean-squares problem in the more natural space, e.g. in the space propagating univariate marginals under $M^{(1)}$-separability. In general, this can be accomplished in a lower dimensional space defined by a matrix $\widetilde{M}_*$ with full column rank $p$ such that $\mathcal{R}(\widetilde{M}_*) = \mathcal{R}(M_*)$. The idea is as follows: when an $\eta \times \overline{p}$ matrix $M_*$ fails to have full column rank, there is redundancy in the representation of the partial information vectors in the form of linear constraints that are satisfied. One wishes to work in $p$-dimensional space, compute a linear least-mean-squares estimate $\widehat{\widehat{H}}_{LS}$ in $\mathbb{R}^p$, that when transformed into a matrix propagating partial information in $\mathbb{R}^{\overline{p}}$, is a linear least-mean-squares estimate $\widehat{H}_{LS}$ in $\mathbb{R}^{\overline{p}}$. The key is for $\ell_2$ distances to be invariant under the transformation from $\mathbb{R}^{\overline{p}}$ to $\mathbb{R}^p$ and vice-versa. This is easily accomplished. For the $\eta \times \overline{p}$ matrix $M_*$ with rank $p$, represent an orthonormal basis for the row space of $M_*$ as the columns of a $\overline{p} \times p$ matrix $V$. Define the $\eta \times p$ matrix

$$\widetilde{M}_* \triangleq M_* V \quad , \tag{5.70}$$

and note that the column spaces of $M_*$ and $\widetilde{M}_*$ are equal. As $V$ is orthonormal, one can easily verify that 2-norms and $\ell_2$-distances are invariant under the transformation by $V$.

Naturally, $\widehat{H}_{LS}$, the estimate for $H_*$ with respect to $M_*$, can be derived from $\widehat{\widehat{H}}_{LS}$, the estimate of the matrix propagating partial information with respect to $\widetilde{M}_*$ calculated via (5.66), because

$$\widehat{H}_{LS} = V \widehat{\widehat{H}}_{LS} V' \quad . \tag{5.71}$$

The same process can be mirrored in the linearly constrained case for the estimator $\widehat{\widehat{H}}_{\underline{LS}}$. It is straightforward to show that by respecting the linear equality constraints characterizing $\mathcal{H}_=(\widetilde{M}_*)$, we will automatically respect all of the linear equality constraints characterizing $\mathcal{H}_=(M_*)$ when $\widehat{H}_{\underline{LS}}$ is derived from $\widehat{\widehat{H}}_{\underline{LS}}$ via (5.71).

### 5.4.2 Feasibility of $\widehat{H}_{LS}(\mathsf{x}_0^t)$

Theorem 11 argues that a sequence of estimators $\widehat{H}_{LS}(\mathsf{x}_0^t)$, $\widehat{H}_{LS}(\mathsf{x}_0^{t+1})$, $\widehat{H}_{LS}(\mathsf{x}_0^{t+2})$, ... will converge almost surely to the true $H_*$ under fairly general conditions. Obviously $H_* \in \mathcal{H}(M_*)$. But is it possible that all of our estimates $\widehat{H}_{LS}(\mathsf{x}_0^t)$ for finite times $t$ will be infeasible?

Although the answer may be unclear in the case of $\widehat{H}_{LS}(\mathsf{x}_0^t)$, it should be clear for $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$, as summarized by the following theorem.

**Theorem 12.** *When $H_*$ is in the relative interior of $\mathcal{H}(M_*)$, almost surely, our constrained estimator $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$ will eventually become feasible and remain feasible, i.e., there exists a time $\overline{t}$ such that $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t) \in \mathcal{H}(M_*)$ for all $t \geq \overline{t}$.*

Theorem 12 follows based on an analysis argument that considers the following: $\mathcal{H}(M_*)$ is a finitely-generated convex set, $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$ is always in the affine hull of $\mathcal{H}(M_*)$, and $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$

converges almost surely to $H_*$, a point in the relative interior of $\mathcal{H}(M_*)$. Can the same be said for $\widehat{H}_{LS}(\mathsf{x}_0^t)$, when it cannot seemingly be assured that $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$ will converge to $H_*$ in the affine hull of $\mathcal{H}(M_*)$?

Indeed, we will show that any least-mean-squares estimator $\widehat{H}_{LS}(\mathsf{x}_0^t)$ will almost surely satisfy the linear equality constraints characterizing $\mathcal{H}(M_*)$ for sufficiently large $t$, even though such an estimator is computed without regard to such linear equality constraints (note that when $\mathbb{1} \notin \mathcal{R}(M_*)$, there are no linear equality constraints characterizing $\mathcal{H}(M_*)$ as $M_*$ is assumed to have full column rank). This ensures that the sequence of estimators $\widehat{H}_{LS}(\mathsf{x}_0^t)$, $\widehat{H}_{LS}(\mathsf{x}_0^{t+1})$, $\widehat{H}_{LS}(\mathsf{x}_0^{t+2})$ also approaches $H_*$ in the affine hull of $\mathcal{H}(M_*)$. If one knew the distance of $H_*$ from the relative boundary of $\mathcal{H}(M_*)$ in the affine subspace and the speed of convergence of $\widehat{H}_{LS}(\mathsf{x}_0^t)$ to $H_*$, one could develop probabilistic guarantees for the feasibility of $\widehat{H}_{LS}(\mathsf{x}_0^t)$.

Our argument relies on a key fact: affine linear least-mean-squares estimators must satisfy all linear equalities satisfied by the random variables that they estimate. This is summarized in the following theorem.

**Theorem 13.** *Suppose that $\alpha_1 \mathsf{w}_1 + \alpha_2 \mathsf{w}_2 = b$ for random variables $\mathsf{w}_1, \mathsf{w}_2$ and scalars $\alpha_1$, $\alpha_2$, $b$. It follows that $\alpha_1 \widehat{\mathsf{w}_1}_{ALLS}(\mathsf{y}) + \alpha_2 \widehat{\mathsf{w}_2}_{ALLS}(\mathsf{y}) = b$, where $\widehat{\mathsf{w}}_{i\,ALLS}(\mathsf{y})$ is the affine linear least-mean-squares estimator of $\mathsf{w}_i$ given $\mathsf{y}$.*

*Proof.* Let $\mathsf{z} \triangleq \alpha_1 \mathsf{w}_1 + \alpha_2 \mathsf{w}_2$. Consider the form of the affine linear least-mean-squares estimator of $\mathsf{z}$ given $\mathsf{y}$ [48]:

$$\widehat{\mathsf{z}}_{LLS}(\mathsf{y}) = \mathbb{E}\left[\,\mathsf{z}\,\right] + \frac{\mathrm{cov}(\mathsf{z},\mathsf{y})}{\mathrm{cov}(\mathsf{y},\mathsf{y})}\left(\mathsf{y} - \mathbb{E}\left[\,y\,\right]\right) \quad . \tag{5.72}$$

By linearity of expectation,

$$\widehat{\mathsf{z}}_{LLS}(\mathsf{y}) = \alpha_1 \widehat{\mathsf{w}_1}_{LLS}(\mathsf{y}) + \alpha_2 \widehat{\mathsf{w}_2}_{LLS}(\mathsf{y}) \quad . \tag{5.73}$$

As $\mathsf{z} = \boldsymbol{b}$, evidently $\widehat{\mathsf{z}}_{LLS}(\mathsf{y}) = \boldsymbol{b}$. □

We can use this theorem to show that almost surely, $\widehat{H}_{LS}(\mathsf{x}_0^t)$ will eventually satisfy all of the linear equalities of $\mathcal{H}(M_*)$ (which only exist when $\mathbb{1} \in \mathcal{R}(M_*)$). First, define $\mathcal{S}_*^t(\mathsf{x}_0^t)$ as the subset of $\mathcal{S}_*$ (recall (5.54)) that is observed as samples of $\mathsf{s}_{\mathsf{i}*}$ over the trajectory $\mathsf{x}_0^t$, i.e.,

$$\mathcal{S}_*^t(\mathsf{x}_0^t) = \left\{\mathsf{s}_{\mathsf{x}}[\tau]' M_* : \tau \in [0, t-1]\right\} \quad . \tag{5.74}$$

We wish to invoke Theorem 13 for the linear least-mean-squares estimator of $\mathsf{s}_{\mathsf{j}*}$ given $\mathsf{s}_{\mathsf{i}*}$, as computed from the empirical distribution over i, j that is specified by $\mathsf{x}_0^t$. This estimator can be expressed in terms of $\widehat{H}_{LS}(\mathsf{x}_0^t)$ as

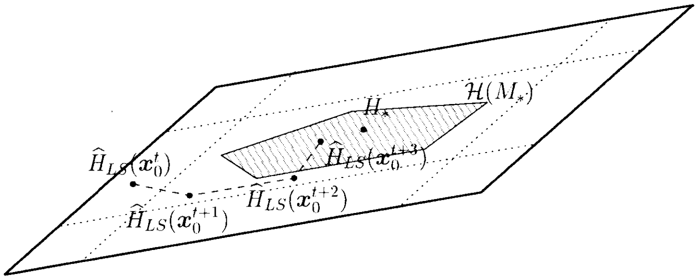
$$\widehat{\mathsf{s}_{\mathsf{j}*}}_{LLS}(\mathsf{s}_{\mathsf{i}*}\,;\,\mathsf{x}_0^t)' = \mathsf{s}_{\mathsf{i}*}' \widehat{H}_{LS}(\mathsf{x}_0^t) \quad . \tag{5.75}$$

Such an estimator must satisfy

$$\begin{aligned} \widehat{\mathsf{s}_{\mathsf{j}*}}_{LLS}(\mathsf{s}_{\mathsf{i}*}\,;\,\mathsf{x}_0^t)' \boldsymbol{a} &= c \\ \boldsymbol{m}' \widehat{H}_{LS}(\mathsf{x}_0^t) \boldsymbol{a} &= c \quad \forall\, \boldsymbol{m} \in \mathcal{S}_0^t(\mathsf{x}_0^t), \end{aligned} \tag{5.76}$$

where $\boldsymbol{a} = A$ and $c = \boldsymbol{c}$ are from (5.55), and characterize the linear equality constraints of $\mathcal{H}(M_*)$ (5.57).

Figure 5-1: Sequence of estimators must converge to $H_*$ in the affine hull of $\mathcal{H}(M_*)$ (the pentagon represents $\mathcal{H}(M_*)$)


$\mathcal{H}(M_*)$
$H_*$
$\widehat{H}_{LS}(\boldsymbol{x}_0^t)$
$\widehat{H}_{LS}(\boldsymbol{x}_0^{t+3})$
$\widehat{H}_{LS}(\boldsymbol{x}_0^{t+2})$
$\widehat{H}_{LS}(\boldsymbol{x}_0^{t+1})$

We argue that (5.76) must be satisfied as follows: when $\mathbb{1} \in \mathcal{R}(M_*)$, the linear least-mean-squares estimate of $\mathsf{s}_{j*}$ given $\mathsf{s}_{i*}$ is equivalent to the affine linear least-mean-squares estimate. This fact follows from Corollary 6, thereby allowing us to invoke Corollary 13 to ensure (5.76).

Examining (5.57), (5.59), and (5.76), $\widehat{H}_{LS}(\mathsf{x}_0^t)$ satisfies all linear equalities characterizing $\mathcal{H}(M_*)$ if the affine hull of $\mathcal{S}_*^t(\mathsf{x}_0^t)$ equals the affine hull of $\mathcal{S}_*$. This will occur almost surely when $G$ is ergodic, for an observed trajectory $\mathsf{x}_0^t$ with as short of a duration as $t = p$. For this particular time $\bar{\mathsf{t}}$ (existing almost surely) whereby the affine hull of $\mathcal{S}_*^{\bar{\mathsf{t}}}(\mathsf{x}_0^{\bar{\mathsf{t}}})$ equals the affine hull of $\mathcal{S}_*$, $\widehat{H}_{LS}(\mathsf{x}_0^t)$ for all $t \geq \bar{\mathsf{t}}$ will satisfy the linear equalities characterizing $\mathcal{H}(M_*)$. This allows us to claim the following corollary that relies on Theorem 13.

**Corollary 11.** *When $\mathbb{1} \notin \mathcal{R}(M_*)$, $\widehat{H}_{LS}(\mathsf{x}_0^t) = \widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$ for all times $t$.*

*When $\mathbb{1} \in \mathcal{R}(M_*)$, there exist linear equality constraints characterizing $\mathcal{H}(M_*)$, which are only considered when computing $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$. Nevertheless, provided that the underlying transition matrix $G$ is ergodic, almost surely there exists a stopping time*[7] *$\bar{\mathsf{t}}$ determined by $\mathsf{x}_0^\infty$ such that $\widehat{H}_{LS}(\mathsf{x}_0^t) = \widehat{H}_{\underline{LS}}(\mathsf{x}_0^t)$ for all $t \geq \bar{\mathsf{t}}$. At this time $\bar{\mathsf{t}}$, the affine hull of $\mathcal{S}_*^t(\mathsf{x}_0^t)$ equals the affine hull of $\mathcal{S}_*$, or equivalently, $R_{\mathsf{s}_{i*},\mathsf{s}_{i*}}$ of* (5.66) *is invertible.*

When $R_{\mathsf{s}_{i*},\mathsf{s}_{i*}}$ lacks full rank, we can compute LS estimates of $H_*$ by evaluating (5.66) using a Moore-Penrose pseudoinverse. However, such estimates of $H_*$ are poor, in the sense that they map partial information vectors in unobserved subspaces to 0 (it is known that partial information vectors span all of $\mathbb{R}^p$ whenever $M_*$ has full column rank). Hence, estimates of $H_*$ are only deemed reasonable when $R_{\mathsf{s}_{i*},\mathsf{s}_{i*}}$ has full rank, assuring by Corollary 11 that $\widehat{H}_{\underline{LS}}(\mathsf{x}_0^t) = \widehat{H}_{LS}(\mathsf{x}_0^t)$. Thus, we argue that it is unnecessary to consider the linear equality constraints of $\mathcal{H}(M_*)$ when computing an LS estimate, as such equality constraints will automatically be satisfied for reasonable estimates. Lastly, we have the following corollary following from Theorem 12 and Corollary 11.

**Corollary 12.** *When $H_*$ is in the relative interior of $\mathcal{H}(M_*)$, almost surely there exists a time $\bar{\mathsf{t}}$ such that for all $t \geq \bar{\mathsf{t}}$, $\widehat{H}_{LS}(\mathsf{x}_0^t)$ will be feasible.*

Thus, for either estimator, we are assured the set-up schematically illustrated in Fig. 5-1, where the sequence of estimators converges to $H_*$ in the affine hull of $\mathcal{H}(M_*)$.

[7] As defined in [37], a random variable $\bar{\mathsf{t}}$ is a stopping time with respect to the stochastic process $\mathsf{x}_0^\infty$ if the event $\{\bar{\mathsf{t}} = \tau\}$ can be determined by $\mathsf{x}_0^\tau$, i.e., it is in the $\sigma$-field generated by $\mathsf{x}_0^\tau$.

### 5.4.3 Validating $M^{(r)}$-separability

Suppose that we estimate $H_*$ via (5.66) for a stochastic network that is not necessarily $M_*$-separable. In computing the estimate of $H_*$, is it possible to affirm or invalidate the assumption of $M_*$-separability? We first note two observations limiting our assertions.

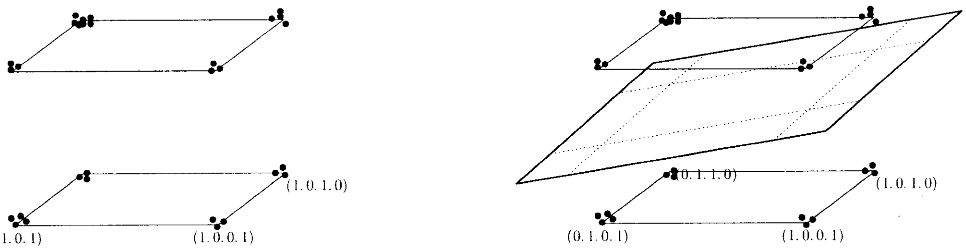
When there exist strictly positive transition matrices in $\mathcal{G}(M_*)$, it is possible to observe any sequence of states. Therefore, in these cases, which include $M^{(r)}$-separability, one cannot invalidate the assumption of $M_*$-separability based on an observed trajectory. On the other hand, one could characterize typical trajectories that occur with high probability, and with high confidence rule out $M_*$-separability based on the observed trajectory. However, because the number of states is exponential in $n$ and $M_*$-separability has the same order of parameters as the unconstrained case, maintaining a catalogue of typical trajectories or properties characterizing typical trajectories from $M_*$-separable networks is seemingly intractable.

As an alternative approach, when the estimator $\widehat{H}_{LS}(\mathbf{x}_0^t)$ has seemingly converged, which will be the case even when $M_*$-separability does not hold, one verifies whether or not it is feasible or close to feasible. If the estimate is far from feasible, this would suggest that $M_*$-separability fails to hold. Checking the linear equality constraints characterizing $\mathcal{H}(M_*)$ is tractable, and by the same arguments given in Section 5.4.2, such linear equality constraints characterizing $\mathcal{H}(M_*)$ will eventually be satisfied by the estimator $\widehat{H}_{LS}(\mathbf{x}_0^t)$ even when $M_*$-separability fails to hold (our arguments never relied on an assumption of $M_*$-separability. But as verifying the linear inequality constraints of $\mathcal{H}(M_*)$ is intractable for large $n$, this approach to evaluating the appropriateness of an assumption of $M_*$-separability, in general, will be intractable.

To consider what is tractable, let's examine the process of computing the estimate $\widehat{H}_{LS}$. Recall from Section 3.3 the visual perspective on $M_*$-separability. The LS estimate of $H_*$, independently for each coordinate function (an entry of the partial information vector), fits, in a least squares sense, a linear function to the observed data. In the case of $M^{(r)}$-separability, all of the observed data is in the form of 0s or 1s. Schematically, we have a picture as illustrated in Fig. 5-2 for each coordinate function. If $M_*$-separability holds, then the true mean for the binary observations at each element of the domain (a state of the network) passes through a plane. However, in order for the LS estimate of $H_*$ to converge to the true $H_*$, it is not necessary that the empirical mean for each of these binary random variables, one associated with each state, to have already converged. Because the deviations from the true mean for each observation are independent (when considered on the basis of a single coordinate function), when $p \ll \eta$, the estimate (the least squares linear fit) may well converge without observing a transition out of every state.

The pictures of Fig. 5-2 suggest attempting to track the reduction in the mean square error offered by the linear least-mean-squares estimate of $\mathsf{s}_{j*}$ given $\mathsf{s}_{i*}$ relative to the uninformed estimate $\mathbb{E}\left[\mathsf{s}_{j*}\right]$, which can be conceived as a level plane. We would like to compute such errors with respect to the true distributions of $\mathsf{s}_{i*}$ and $\mathsf{s}_{j*}$, as it is the true mean square errors that are informative. In practice, however, one computes mean square error estimates based on the empirical distribution for $\mathsf{s}_{i*}$ and $\mathsf{s}_{j*}$ derived from the observed trajectory $\mathbf{x}_0^t$, which will almost surely converge to the true mean square errors by renewal theory for an underlying Markov chain that is ergodic.

For simplicity, we assume that $\mathbb{1} \in \mathcal{R}(M_*)$, thereby ensuring that a 'level' plane can obtained as the linear estimator of $\mathsf{s}_{j*}$ given $\mathsf{s}_{i*}$ (as only when $\mathbb{1} \in \mathcal{R}(M_*)$ are the linear and affine linear estimators equivalent). The reduction in mean-square error gauges the degree



(a) All observed samples (dots) are binary, either 0 (lower plane) or 1 (upper plane)

(b) Independently for each coordinate function, a best-fit plane in the least-squares sense is fit to the observed data

Figure 5-2: Schematic illustration of the linear fitting by a coordinate function of the LS estimate of $H_*$ (consult Section 3.3 for a explanation of such visualizations)

of 'tilt' of the planes for each coordinate function. The intuition is that when there is no reduction in mean-square error (all planes are level), then assuming $M_*$-separability and computing an estimate of $H_*$ is uninformative.

Of course, it is possible for $M_*$-separability not to hold, but there be a reduction in mean-square error. On the flip side, there will be no reduction in mean-square error under $M_*$-separability, if the updated partial information at the next time step from all states is unchanged. In such an instance, however, estimating $H_*$ is of little use (a rank 1 matrix with identical columns). By considering the reduction in the mean square error, we do not have an absolute certificate for $M_*$-separability, but one that is useful, nonetheless.

This measure will be tractable provided that estimating $H_*$ is tractable; it can be estimated from the familiar expression for the error covariance of the affine linear least-mean-squares estimator,

$$\text{mean square error reduction estimate} = \text{trace}\left(\widehat{\Lambda}_{\mathsf{s}_{j*}\mathsf{s}_{i*}}\widehat{\Lambda}^{-1}_{\mathsf{s}_{i*}\mathsf{s}_{i*}}\widehat{\Lambda}_{\mathsf{s}_{i*}\mathsf{s}_{j*}}\right) \quad , \tag{5.77}$$

where $\widehat{\Lambda}_{\mathsf{s}_{i*}\mathsf{s}_{j*}} \triangleq \mathbb{E}_{\widehat{p}}\left[\mathsf{s}_{i*}\mathsf{s}_{j*}'\right] - \mathbb{E}_{\widehat{p}}\left[\mathsf{s}_{i*}\right]\mathbb{E}_{\widehat{p}}\left[\mathsf{s}_{j*}'\right]$ is the empirical cross-covariance of $\mathsf{s}_{i*}$ and $\mathsf{s}_{j*}$ obtained from the observed trajectory $\mathbf{x}_0^t$.

Reduction in mean-square error is only informative when there is no reduction. It does not invalidate the possibility of $M_*$-separability being satisfied, but rather pronounces it irrelevant. What would be more desirable is a measure that could actually invalidate $M_*$-separability. The following theorem motivates one possible approach.

**Theorem 14.** *Assume a network is $M_*$-separable. Then for any $M_\#$ such that $\mathcal{R}(M_*) \subset \mathcal{R}(M_*)$, the reduction in mean-square error of $\mathsf{s}_{j*}$ given $\mathsf{s}_{i*}$ must equal the reduction in mean-square error of $\mathsf{s}_{j*}$ given $\mathsf{s}_{i\#}$,*[8] *where such reductions in expected square-error are calculated*

[8] The true joint distribution of $\mathsf{s}_{i\#}$ and $\mathsf{s}_{j*}$ would be given by the joint distribution of $\mathsf{s}_\#[\tau]' \triangleq \mathsf{s}_\mathsf{x}[\tau]'M_\#$ and $\mathsf{s}_*[\tau+1]' \triangleq \mathsf{s}_\mathsf{x}[\tau+1]'M_*$ for a well-mixed, time-homogeneous, ergodic Markov chain.

*from the true distributions. Equivalently,*

$$trace\left(\Lambda_{\mathsf{s}_{j*}\mathsf{s}_{i*}}\Lambda^{-1}_{\mathsf{s}_{i*}\mathsf{s}_{i*}}\Lambda_{\mathsf{s}_{i*}\mathsf{s}_{j*}}\right) = trace\left(\Lambda_{\mathsf{s}_{j*}\mathsf{s}_{i\#}}\Lambda^{-1}_{\mathsf{s}_{i\#}\mathsf{s}_{i\#}}\Lambda_{\mathsf{s}_{i\#}\mathsf{s}_{j*}}\right) \quad . \tag{5.78}$$

This theorem follows from the definition of $M_*$-sufficiency (Definition 2), which by Theorem 1 is equivalent to $M_*$-separability.

In practice one can use Theorem 14 to invalidate $M_*$-separability in a way that may often be tractable. For example, suppose one wishes to see if an assumption of $M^{(r)}$-separability is invalid. After computing the estimate in the reduction of the mean square error for the linear least-mean-squares estimate of $\mathsf{s}^{(r)}[\tau+1]$ given $\mathsf{s}^{(r)}[\tau]$, as calculated from the observed trajectory $\mathsf{x}_0^t$, one then estimates the reduction of the mean square error for the linear least-mean-squares estimate of $\mathsf{s}^{(r)}[\tau+1]$ given $\mathsf{s}^{(r+1)}[\tau]$ (note that $\mathcal{R}(M^{(r)}) \subset \mathcal{R}(M^{(r+1)})$). If the estimated reduction is greater in the latter case, and the trajectory $\mathsf{x}_0^t$ is sufficiently long such that the estimates computed from the empirical distribution are fairly accurate, then Theorem 14 suggests that $M^{(r)}$-separability is not satisfied. Note that such a test offers sufficient conditions for $M^{(r)}$-separability to be invalid, but not necessary conditions.

## 5.5 Computational examples

We will consider two examples, the first being card shuffling. This example permits us to draw comparisons to results that can be derived from symmetric group theory. A second more general example involves a generalized influence model. Both examples illustrate the tractability of $M^{(1)}$-separability when traditional analysis of the underlying transition matrix $G$ is intractable.

### 5.5.1 Shuffling a deck of cards

We consider a shuffling technique referred to as *random transpositions*:

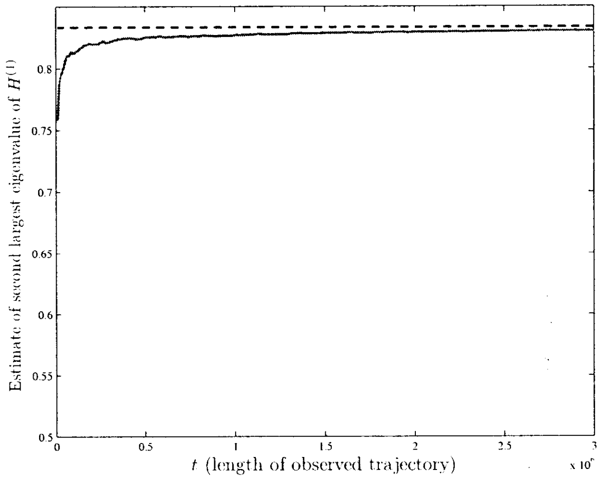
- choose two cards uniformly and independently at random, and
- swap the positions of the two cards.

Note that with probability $\frac{1}{n}$, where $n$ is the number of cards in the deck, a realized shuffle will not rearrange any of the cards.

As is often the case when shuffling a deck of cards, we are interested in the number of shuffles necessary until the deck is well shuffled. It is easy to argue that the underlying transition matrix must be symmetric, and consequently, the steady-state probability vector is uniform. We consider a deck as being well-shuffled after a sequence of random transpositions when the probability that the deck is in any of its $n!$ arrangements is roughly $\frac{1}{n!}$, regardless of the initial arrangement of the cards. Although there are several approaches to analyzing the convergence of Markov chains to steady-state [40], we will limit our focus to the second largest eigenvalue of the underlying transition matrix $G$.

We will consider the representation of the deck of cards as $n$ automata, each corresponding to different position in the deck, i.e., the positions representation where each automaton indicates the particular card in the given position. As explained in Section 4.5, shuffling a deck of cards is a GIM, and hence exhibits $M^{(r)}$-separability for all $1 \leq r < n$. Alternatively, this can be argued on the basis of symmetric group theory [11, 9]. Our initial goal will be to estimate the transition matrix $H^{(1)}$ that propagates the univariate marginal PMFs.

Figure 5-3: Convergence of the second largest eigenvalue of $\widehat{H}^{(1)}$ for random transpositions shuffling ($n = 12$).



For our specific example, we consider a deck of $n = 12$ cards. With permutations being maintained, there are $12! \approx 4.79 \times 10^8$ possible states. The matrix $H^{(1)}$ is $144 \times 144$, and hence tractable. We randomly generate a trajectory $\mathsf{x}_0^t$ with $t = 3 \times 10^6$—more than two orders of magnitude less than the size of the state space. We then iteratively compute estimates of $H^{(1)}$, based on $\mathsf{x}_0^\tau$, where $\tau$ varies up to $t = 3$ million. For the card shuffling example, we plot the second largest relevant eigenvalue[9] of our estimate $\widehat{H}^{(1)}$, and note its apparent convergence (this is computed by defining a matrix with the same range space as $M^{(1)}$ with full column rank). Because eigenvalues are continuous functions, we know that the 2nd largest relevant eigenvalue of $\widehat{H}^{(1)}$ will converge almost surely to the second largest relevant eigenvalue of $H^{(1)}$, by Theorem 11 (using the full rank matrix $\widetilde{M}^{(1)}$).

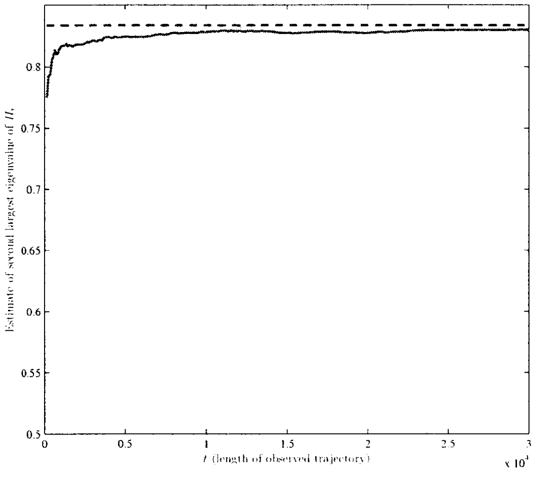
By our arguments in Section 3.1.2, we know that the second largest relevant eigenvalue of $\widehat{H}^{(1)}$ will converge to an eigenvalue of $G$ that will be less than or equal to its second largest eigenvalue. By symmetric group theory [9], one can show that the second largest eigenvalue of $G$ for random transpositions shuffling is $\frac{n-2}{n}$. As our simulation results illustrate in Fig. 5-3, the second largest eigenvalue of our estimates $\widehat{H}^{(1)}$ appears to converge to $\frac{n-2}{n} = \frac{10}{12} \approx 0.83$, which is illustrated as the horizontal line. Naturally this approach can be applied to any shuffling technique, however the approximate lower bound on the second largest eigenvalue that we compute will not necessarily be tight, as appears to be the case for random transpositions shuffling.

By taking a deeper look at the representations of shuffling as a GIM as discussed in Section 4.5, it should be evident that estimates of the second largest eigenvalue of $H^{(1)}$ can be obtained even more easily. From (4.37), recall that $\overline{H}^{(1)}$ of the cards representation, can be represented as

$$\overline{H}^{(1)} = I \otimes \mathbb{E}[\mathsf{A}] \quad , \tag{5.79}$$

[9] Relevant eigenvalues are eigenvalues that can be excited by partial information vectors $\boldsymbol{\pi}^{(1)}$, which lie in a particular subspace as $M^{(1)}$ fails to have full column rank.

Figure 5-4: Convergence of the second largest eigenvalue of $\widehat{H}_i$ for random transpositions shuffling ($n = 12$).



which is similar to our $H^{(1)}$ of the positions representation.

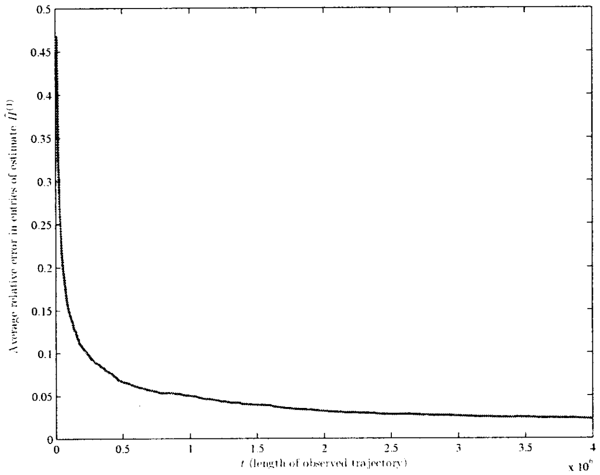
As argued in Section 4.5, $H^{(1)}$'s second largest relevant eigenvalue by modulus is the second largest eigenvalue by modulus of $\mathbb{E}[\mathsf{A}]$. The $n \times n$ matrix $\mathbb{E}[\mathsf{A}]$ is the matrix that propagates $\boldsymbol{\pi}_i$ under the cards representation, for any automaton $i$—recall that under the cards representation, the card shuffling GIM can be thought of as a GIM of coupled Markov chains that is $M_i$-separable for all $i$, with $H_i = \mathbb{E}[\mathsf{A}]$ being the matrix that propagates the partial information $\boldsymbol{\pi}_i$.

Rather than estimating the $n^2 \times n^2$ matrix $H^{(1)}$, we can estimate the $n \times n$ matrix $H_i$ and obtain an approximate lower bound for the second largest eigenvalue by modulus of $G$. Moreover, as each $H_i = \mathbb{E}[\mathsf{A}]$, we can use the updates of the status vectors for all automata to construct an empirical distribution for $\mathbf{s}_i[\tau]$ and $\mathbf{s}_i[\tau + 1]$ that will be used to estimate $H_i$. In Fig. 5-4 we do just that, plotting $\widehat{H}_i$'s second largest eigenvalue by modulus as derived from an observed trajectory $\mathbf{x}_0^t$ with $t$ varying up to $3 \times 10^4$—a trajectory with two orders of magnitude fewer samples than illustrated in Fig. 5-3. The apparent convergence to the second largest eigenvalue of $G$ is improved in Fig. 5-3, when the estimated bound is computed from $\widehat{H}_i$ of the cards representation rather than $\widehat{H}^{(1)}$ of the positions representation. Although our results only assure that we are computing an approximate lower bound, symmetric group theory assures us that our lower bound is tight [8]. Naturally, it would be desirable to understand analytically when such an approximate lower bound will be tight.

### 5.5.2 General GIM example

We consider a second simulated example of estimating the $H^{(1)}$ matrix of a GIM. The GIM is generated as a random convex combination of three IMs, thereby offering it a compact representation even when $n$ is large. The size of the model considered involved $n = 20$ automata, each with $m = 2$ statuses. Note that such a stochastic network has over 1

Figure 5-5: Convergence of $\widehat{H}^{(1)}$ for general GIM example.



million states, and thus is intractable by traditional analysis.

We computed progressive estimates of $H^{(1)}$, as obtained from a trajectory $\mathbf{x}_0^t$ with $t$ varying up to 4 million—the same order of magnitude as the number of states in the Markov chain. We ran several simulations, and all such runs exhibited the same general behavior. Fig. 5-5 illustrates the convergence of the estimate of $H^{(1)}$ for one such simulation run. The duration of the observed trajectory from which estimates of $H^{(1)}$ are computed varies along the $t$-axis. The dependent axis provides the average relative error between our estimate $\widehat{H}^{(1)}(\mathbf{x}_0^t)$ and the actual $H^{(1)}$, calculated as follows:

$$\text{average relative error between } H^{(1)} \text{ and } \widehat{H}^{(1)}(\mathbf{x}_0^t) \triangleq \frac{1}{m^2n^2}\sum_{i,j}\frac{\left[\left|H^{(1)} - \widehat{H}^{(1)}(\mathbf{x}_0^t)\right|\right]_{ij}}{\left[\left|H^{(1)}\right|\right]_{ij}} \quad . \tag{5.80}$$

For the examples that we consider ($n$ moderately large), the majority of the entries of $H^{(1)}$ are rather small, and hence failing to normalize the error can be misleading; consequently, we feel relative error is a fair measure of error. For completeness, we should add that we restricted ourselves to $H^{(1)}$ matrices of minimal Frobenius norm, so as to avoid the issues of $H^{(1)}$ not being unique in the case of $M^{(1)}$-separability (as $M^{(1)}$ fails to have full column rank). As evident from Fig. 5-5, the convergence of our estimate $\widehat{H}^{(1)}$ to $H^{(1)}$ is pretty good, even after having only observed a number of transitions roughly equal to the size of the state space.

## 5.6 Conclusions

This chapter offers many insights into parameter learning under the assumptions of $M_*$-separability. Initially, this chapter focuses on the problem of estimating $G$ under assumptions of $M_*$-separability. Although estimating $G$ is intractable for the interesting cases of $M_*$-separability, the discussion is valuable, as it motivates our proposed algorithm for estimating $H_*$ under assumptions of $M_*$-separability. We argue that this proposed algorithm is tractable: it has computational complexity that is polynomial in $n$ whenever the size of $H_*$ is polynomial in $n$. Moreover its computational routine, linear least-means-squares estimation, is familiar and well-understood. We prove several interesting properties regarding our tractable estimator under some fairly general conditions, namely that it will eventually be feasible, and that it will converge to the true $H_*$, i.e., it is asymptotically consistent. In addition, we offer a general, approach for verifying the appropriateness of assumptions of $M_*$-separability. We closed the chapter by illustrating the performance of our parameter learning algorithm on two examples.

Arguably, Markov chains have achieved their greatest successes as HMMs, made possible by the powerful algorithms to solve the standard HMM estimation problems [77, 78, 79, 80, 81, 82, 83]. The usefulness of $M_*$-separability could be similarly bolstered by tractable algorithms to solve these same problems, in regimes when the standard algorithms are intractable. Proposed in [14] is a tractable approach to determine the state trajectory of an MLSS (in many cases equivalent to separability) based on observed outputs. In this chapter, a tractable parameter learning algorithm is proposed. Ideally, there would be a way to combine these methods and solve both estimation algorithms simultaneously, in a way analogous to what is possible in HMMs, but tractable for large values of $n$. However, as only a subset of the model parameters are estimated in the parameter learning algorithm herein proposed, and as the algorithm to estimate the state in [14] is not an ML estimation algorithm, there are additional complications for the special case of $M_*$-separability. Possibly these complications could be resolved in a more restricted problem setting.

# Chapter 6

# Spatial perspectives on separability

Thus far, we have focused on stochastic networks evolving as time-homogeneous Markov chains. As has been shown, separability offers substantial computational advantages by permitting the propagation of partial information. However, it is possible to consider more general Markov models than Markov chains. The focus of this chapter is to extend the idea of separability to general Bayesian networks. Separability in the case of time-homogenous Markov chains was aimed at propagating partial information forward in time. For general Bayesian networks, however, we will be interested in the reduction in complexity of the parameterizations of conditional PMFs induced by algebraic forms similar to those encountered under separability. The conditions that emerge in our development below are very similar to what we obtained in the context of $M_*$-separability, and we therefore refer to the associated property by the label "spatial separability." We shall point out the distinctions from our previous use of the term "separability" at the appropriate points in the chapter.[1]

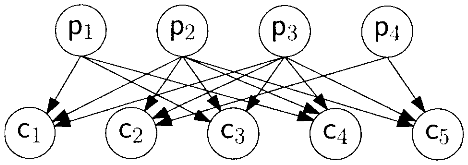
## 6.1 Model description

As before, we have a collection of random variables. There need not be any notion of time; hence we refer to these general models as *spatial.* The joint PMF of $r$ finite-valued random variables is specified, in general, by a number of parameters (joint probabilities) that is exponential in $r$. Consequently, large models are intractable without assuming additional structure. A common approach to simplify the representation of such models is to assume conditional independence relationships among the random variables. Obviously, time-homogeneous Markov chains are an example of a specific form of conditional independence relationships that permit a joint distribution of random variables over any interval of time to be parameterized by the joint transition matrix (representing the conditional PMF for the updated values of the random variables given their values in the immediate past) and an initial distribution of such random variables.

Assumptions of more general forms of conditional independence are seemingly appropriate in a wide-variety of problems, as evident by the widespread application of graphical models, primarily Markov random fields and Bayesian networks. For the models considered in this chapter, it will be assumed that there exists an underlying Bayesian network that encodes the conditional independence relationships of the random variables. We will not provide a tutorial or background on Bayesian networks; we encourage the reader to consult

[1] Our notion of separability should also be distinguished from various notions of graph separation associated with the literature on Bayesian networks and graphical models.

Figure 6-1: Basic building block of directed acyclic graphs: parents with children.



[41] or [84] for excellent introductory treatments of the material. In brief, Bayesian networks use a directed acyclic graph to encode conditional independence relations among a set of random variables. Each node of the graph corresponds to a unique random variable, and the pattern of directed arcs encodes the conditional independence relationships. The connection between conditional independence and graphs exists because of the ability for a graph to represent a consistent set of conditional independence relations among a set of random variables. This ability is a consequence of the parallels between the axioms of conditional independence relations and separation among sets of nodes in a graph, as detailed in [85] and discussed in [41].

Fig. 6-1 illustrates a basic example of a directed acyclic graph encoding conditional independence relations. In this example, each node corresponds to a random variable $\mathsf{p}_i$ (a parent) or $\mathsf{c}_i$ (a child). Provided that all of the parents illustrated in Fig. 6-1 $d$-separate the children, the children $\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3, \mathsf{c}_4, \mathsf{c}_5$ are conditionally independent given the parents $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3, \mathsf{p}_4$. Such an assumption of conditional independence imposes a factorable structure on the local conditional PMF. In our example,

$$\mathbb{P}\left( c_1, c_2, c_3, c_4, c_5 \mid p_1, p_2, p_3, p_4 \right) = \prod_{i=1}^{5} \mathbb{P}\left( c_i \mid p_1, p_2, p_3, p_4 \right) \quad , \tag{6.1}$$

where the lowercase arguments of the conditional PMFs clarify the identity of each conditional PMF (we will employ the same shorthand notation in subsequent conditional PMFs). After assuming conditional independence relations, only the conditional PMF of each child given its parents must be defined. If we assume that the alphabet size for each of the random variables is $m$, the representation of the conditional PMF of the children given the parents has been reduced by conditional independence from an $m^4 \times m^5$ table to five $m \times m^3$ tables (since each child has three parents). Moreover, because of the factorable form of the joint PMF for the parents and children, distributed algorithms can be used to perform inference computations. This factorable form is the 'structure' that conditional independence imposes on the conditional PMF of several children given their parents.

Evidently, such assumptions of conditional independence permit the joint distribution on the random variables to be expressed as a product of conditional distributions, with each conditional distribution corresponding to a different subgraph consisting of a node and its parents. As clear in our example, the complexity of the representation of a Bayesian network is exponential in the size of the largest subgraph corresponding to a conditional PMF for a child given its parents. In addition, the computational complexity of performing inference upon such a Bayesian network is at least exponential in the size of the largest such subgraph.[2] There may be instances where the number of parents of a given node is so

[2] The computation complexity for inference will be exponential in the size of the largest clique of the

large that such a Bayesian network is intractable. Imposing additional structure on these conditional PMFs of a child given its parents can bring tractability to such probabilistic models. If assuming additional conditional independence relations is inappropriate, one must consider an alternate methodology to lend such conditional distributions tractability. There are many approaches one may follow; we consider an approach based on the notion of separability, which offers a general methodology. Conditional independence relations can be thought of as imposing macro-structure on the probabilistic model, while separability can be thought of as imposing micro-structure. We motivate the idea of separability by first discussing Fisher sufficiency (which shares much in common beyond its name with the notion of sufficiency/separability defined for time-homogeneous Markov chains in Section 2.2.2; these connections will be made clear in Section 6.2).

### 6.1.1 Fisher sufficiency

Fisher sufficiency, also known as statistical sufficiency, is a property defined in [27] for classical statistics. Fisher sufficiency imposes additional structure on the conditional PMF of a child given its parents. Its natural analog in Bayesian statistics is discussed in [86] as an example of conditional independence. Although the typical inference set-up where Fisher sufficiency is invoked involves estimating the parents based on observations of the children (inferring upwards in our graphical model), for simplicity we'll illustrate Fisher sufficiency in the context of inferring the children given the parents.

Fisher sufficiency in the Bayesian case can be explained by considering the following: the conditional PMF of $\mathsf{c}_1$ given its parents $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ can be represented as a function with arguments $c_1, p_1, p_2, p_3$, representing the possible values for $\mathsf{c}_1, \mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$, respectively. Suppose that this function $\mathbb{P}(\, c_1 \,|\, p_1, p_2, p_3 \,)$ can be expressed as a function of only two arguments: $c_1$ and a *function* of $p_1, p_2, p_3$:

$$\mathbb{P}(\, c_1 \,|\, p_1, p_2, p_3 \,) = \tilde{g}(\, c_1, f(p_1, p_2, p_3)) \quad .^3 \tag{6.2}$$

In such cases, we can define a new random variable $\mathsf{q}_1 = f(\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3)$ and introduce it into our graphical model as shown in Fig. 6-2, where we have used multi-line arrows to denote that the random variable $\mathsf{q}_1$ is a *function* of its parents and thus uniquely determined by its parents (in contrast to $\mathsf{c}_1$, which probabilistically depends on $\mathsf{q}_1$). The random variable $\mathsf{q}_1$ is said to be *sufficient* (in the sense of [27]) for inferring $\mathsf{c}_1$, meaning that $\mathsf{q}_1$ contains all the information necessary for inferring $\mathsf{c}_1$ from the parents $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$. As evident from the graphical model of Fig. 6-2(b), $\mathsf{q}_1$ being statistically sufficient is equivalent to $\mathsf{c}_1$ being conditionally independent of its parents $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ when $\mathsf{q}_1$ is given. Thus Fisher sufficiency equates to the existence of the conditional independence relationships encoded in Fig. 6-2(b), whereby $\mathsf{q}_1$ is a function of its parents. Fisher sufficiency can be interpreted as assuming additional forms of conditional independence with the introduction of a new random variable $\mathsf{q}_1$.

It may seem that introducing $\mathsf{q}_1$ into the graphical model only introduces additional complexity—an additional random variable. However, assume that the cardinality of $\mathsf{q}_1$ (denoted by $m_\mathsf{q}$) is much less than $m$. By expanding the graphical model to consider

moralized (edges made undirected, parents married) and triangulated graph [76]. The size of the largest clique will be at least as large as the greatest number of parents of a given node.

[3] Naturally this is always the case when the function $f$ is invertible; the interesting cases (when Fisher sufficiency actually imposes additional structure) are for functions $f$ that are not invertible.

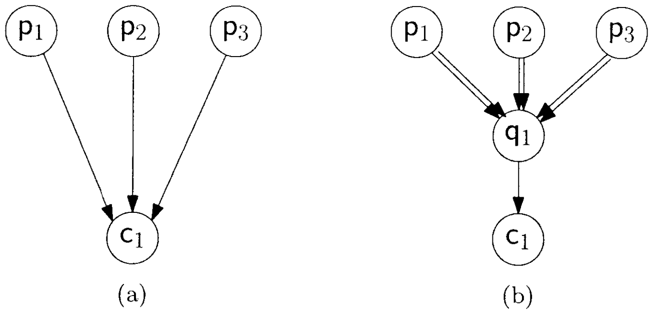
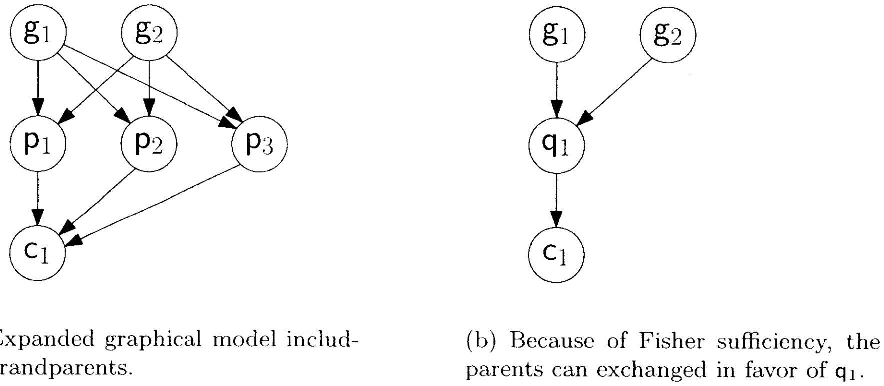


Figure 6-2: When $q_1$ is sufficient for $c_1$ given its parents, the subgraph in (a) can be modified to include $q_1$, as illustrated in (b).

grandparents (assumed to be of cardinality $m$), the advantages offered by Fisher sufficiency become evident. Consider Fig. 6-3(a), which includes the grandparents. Because of Fisher sufficiency, the parents $p_1, p_2, p_3$ can be exchanged for $q_1$, without changing the probabilistic relationship between the child $c_1$ and its grandparents $g_1$, $g_2$. What results from such an exchange, as illustrated in Fig. 6-3(b), is a graphical model with a potentially simpler representation. Without Fisher sufficiency, the graphical model of Fig. 6-3(a) would require tables of sizes $m^3 \times m^2$ and $m \times m^3$ to represent the conditional PMF of the child given the grandparents. On the other hand, the graphical model in Fig. 6-3(b) can be represented by two tables of respective sizes $m_q \times m^2$ and $m \times m_q$, which is a reduction in complexity provided that $m_q$ is much less than $m^3$.



(a) Expanded graphical model including grandparents.

(b) Because of Fisher sufficiency, the parents can exchanged in favor of $q_1$.

Figure 6-3: Simplifying a graphical model when a sufficient random variable is identified.

The reduction in complexity of the representation of the conditional PMF of the child given its parents can be analyzed with linear algebra. The random variable $q_1$, which is sufficient for performing inference on the children, is defined as some function of the parents $p_1, p_2, p_3$. Consequently, its PMF is a linear transformation of the joint PMF of the parents $p_1, p_2, p_3$. This holds regardless of the form of the function $f$ defining $q_1$. To see this, consider the mapping of an indicator vector[4] for $p_1, p_2, p_3$ to an indicator vector for $q_1$; this mapping is a representation of the function $f$. The mapping between indicator vectors

[4] An indicator vector is a vector consisting of 0s and a single 1 whose position indicates the value of $p_1, p_2, p_3$; the length of the vector is the cardinality of the set of possible values for $p_1, p_2, p_3$, so $m^3$ in our example.

necessarily must have a representation as a matrix, denoted as $M_{\mathsf{q}_1}$. This matrix $M_{\mathsf{q}_1}$ is a selector matrix, a row-stochastic matrix consisting of only 0s and 1s. This selector matrix $M_{\mathsf{q}_1}$ is the matrix—the linear transformation—that maps a probability vector for $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ to a probability vector for $\mathsf{q}_1$, that is,

$$\boldsymbol{\pi}'_{\mathsf{q}_1} = \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} M_{\mathsf{q}_1} \quad . \tag{6.3}$$

This linear transformation $M_{\mathsf{q}_1}$ is a representation of the conditional PMF for $\mathsf{q}_1$ given $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ that corresponds to the subgraph in the upper section of Fig. 6-2(b).

Returning to the conditional PMF for $\mathsf{c}_1$ given $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ [Fig. 6-2(a)], this conditional PMF is the relation that produces the PMF of $\mathsf{c}_1$ when the joint PMF of $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ is known. When these PMFs are represented as probability vectors $\boldsymbol{\pi}_{\mathsf{c}_1}$ and $\boldsymbol{\pi}_{\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3}$, respectively, the conditional PMF can be represented by a row stochastic matrix denoted as $B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3}$ and updating the PMF for $\mathsf{c}_1$ given a PMF for its parents reduces to matrix multiplication:

$$\boldsymbol{\pi}'_{\mathsf{c}_1} = \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} \quad . \tag{6.4}$$

That $\mathsf{q}_1$ is sufficient for $\mathsf{c}_1$ given $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ is equivalent to the fact that when using the joint PMF of $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ to update a PMF for $\mathsf{c}_1$, as illustrated in (6.4), such an update is unchanged for any joint PMF $\boldsymbol{\pi}_{\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3}$ that produces the same PMF $\boldsymbol{\pi}_{\mathsf{q}_1}$ for $\mathsf{q}_1$. This necessitates that the left null space of $M_{\mathsf{q}_1}$ is contained within the left null space of $B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3}$, or equivalently,

$$\left[M_{\mathsf{q}_1}^{\perp}\right]' B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} = 0 \quad . \tag{6.5}$$

By (6.5), Fisher sufficiency imposes a linear constraint on $B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3}$, the conditional PMF of the child $\mathsf{c}_1$ given its parents. Our generalized notion of separability for spatial models is when, as in this example, the conditional PMF of a child given its parents satisfies a linear constraint. Note that this characterization resembles $M_*$-separability for time-homogeneous Markov chains, with the exception that separability in the spatial case is defined with respect to a *single* random variable, and not the collection of random variables defining the network state. Fisher sufficiency offers one form of a linear constraint, namely (6.5), with the restriction that $M_{\mathsf{q}_1}$ be a selector matrix. The notion of spatial separability that we will introduce for Bayesian networks (though "sufficiency" would have been a viable alternative, and perhaps more appropriate in the current setting) will admit linear constraints like (6.5) with the relaxation that $M_{\mathsf{q}_1}$ need not be a selector matrix.

### 6.1.2 General example of spatial separability

Suppose that the parents and children of Fig. 6-1 represent a pride of lions. Two of the parent nodes ($\mathsf{p}_2$, $\mathsf{p}_3$) are associated with male adults and two with female adults ($\mathsf{p}_1$, $\mathsf{p}_4$). The child nodes ($\mathsf{c}_1$, $\mathsf{c}_2$, $\mathsf{c}_3$, $\mathsf{c}_4$, $\mathsf{c}_5$) correspond to the offspring of the four parents over a period of several years. Suppose that each random variable represents the associated lion's genotype for a particular trait (a more realistic graphical model would have additional offspring nodes corresponding to observed random variables, the phenotypes). Considering the structure of the graphical model as illustrated in Fig. 6-1, each cub's mother is known, but each cub's father is undetermined. If there were to exist a model of the unconditional likelihood for each male adult being each cub's father (say, based on observations of sexual contact), then as suggested by nature, it is reasonable to assume the conditional independence relations encoded by the graphical model of Fig. 6-1, namely, that the cubs' genotypes are conditionally independent given the genotypes of the adult lions.

In order to define the Bayesian network of Fig. 6-1, one must determine the conditional PMFs for each cub's genotype given the genotypes of its mother and the two adult males in the pride. In order to be fully defined, such a conditional PMF needs to be specified for every combination of possible genotypes for the three relevant parents (its mother lion and both adult males). However, for this example, it should be evident that in updating a cub's genotype based on some joint PMF over the genotypes of the three relevant adults, it should only depend on the two bivariate marginal PMFs over the genotypes of the mother and the potential father. The full joint PMF for the genotypes of the three relevant adults is more information than what is necessary; a linear transformation of the full joint PMF in the form of two pairwise joint PMFs is sufficient. Notice that in this example we have something rather different from Fisher sufficiency; it is not a random variable $\mathsf{q}_1$ that is sufficient, but rather some partial information in the form of bivariate marginals PMFs. Under Fisher sufficiency, such partial information assumes the form of a univariate marginal PMF for $\mathsf{q}_1$.

Focusing on the conditional PMF corresponding to the subgraph of Fig. 6-2(a), then for any given bivariate marginal PMFs for the genotypes of each possible parent pair, $\boldsymbol{\pi}_{\mathsf{p}_1,\mathsf{p}_2}$ and $\boldsymbol{\pi}_{\mathsf{p}_1,\mathsf{p}_3}$, the child's genotype,

$$\boldsymbol{\pi}'_{\mathsf{c}_1} = \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} \tag{6.6}$$

must be unchanged for any joint PMF of $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ producing such bivariate marginal PMFs. All such joint PMFs over $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ producing such bivariate marginal PMFs must satisfy the linear equation

$$\left[\, \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_2} \;\; \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_3} \right] = \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} M_* \quad , \tag{6.7}$$

where $M_*$ maps the joint PMF of the genotypes of the three relevant adults to the partial information in the form of the given bivariate marginal PMFs $\boldsymbol{\pi}_{\mathsf{p}_1,\mathsf{p}_2}$ and $\boldsymbol{\pi}_{\mathsf{p}_1,\mathsf{p}_3}$. For consistency with our notation for time-homogeneous Markov chains, let

$$\boldsymbol{\pi}'_* \triangleq \left[\, \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_2} \;\; \boldsymbol{\pi}'_{\mathsf{p}_1,\mathsf{p}_3} \right] \quad . \tag{6.8}$$

denote the partial information vector.

The invariance in the update of the child given only partial information about the full joint PMF of its parents imposes a linear constraint on the conditional PMF of the child given its parents, namely that

$$\mathcal{R}(B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3}) \subset \mathcal{R}(M_*) \quad , \tag{6.9}$$

or equivalently,

$$\left[M_*^{\perp}\right]' B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} = 0 \quad . \tag{6.10}$$

Note that (6.10) generalizes the Fisher sufficiency condition (6.5), in that $M_*$ is not a selector matrix.

When the linear transformation $M_*$ maps the full joint PMF for the parents to partial information of the form of a sequence of marginal PMFs, as is the case in the lion example, the conditional PMF takes a special additive form. Concretely, expressing the conditional PMF of the child given its three parents as

$$\mathbb{P}\left(c_1 \,|\, p_1, p_2, p_3\right) \quad ,$$

the linear constraint imposed by this example of separability (6.10) entails that there exist functions $g_1$ and $g_2$ such that

$$\mathbb{P}\left(c_1 \,|\, p_1, p_2, p_3\right) = g_1\left(c_1, p_1, p_2\right) + g_2\left(c_1, p_1, p_3\right) \quad , \tag{6.11}$$

where each function $g_1$ and $g_2$ has as its arguments a strict subset of the arguments of the conditional PMF of the child given its parents. This additive decomposition[5] is a consequence of the form of $M_*$. To see this, note that by (6.9), there must exist a matrix $H_*$ such that

$$\begin{aligned} B_{\mathsf{c}_1|\mathsf{p}_1,\mathsf{p}_2,\mathsf{p}_3} &= M_* H_* \\ &= \begin{bmatrix} M_{\mathsf{p}_1,\mathsf{p}_2} & M_{\mathsf{p}_1,\mathsf{p}_3} \end{bmatrix} \begin{bmatrix} H_{*(1)} \\ H_{*(2)} \end{bmatrix} \\ &= M_{\mathsf{p}_1,\mathsf{p}_2} H_{*(1)} + M_{\mathsf{p}_1,\mathsf{p}_3} H_{*(2)} \end{aligned} \tag{6.12}$$

from which the claim of such an additive form (6.11) follows.

### 6.1.3 Defining separability for spatial networks

When separability was defined for time-homogeneous Markov chains, it was defined with respect to a *network* of random variables characterizing the state of the Markov chain. In the spatial setting, separability will be defined with respect to a *random variable.*

**Definition 4.** *A random variable* $\mathsf{c}$ *in a Bayesian network is spatial* $M_*$*-separable with respect to its parents* $\mathbf{p} = (\mathsf{p}_1, \ldots \mathsf{p}_n)$ *provided that its updated distribution, as calculated from any valid joint distribution on its parents,*

$$\boldsymbol{\pi}'_{\mathsf{c}} = \boldsymbol{\pi}'_{\mathbf{p}} B_{\mathsf{c}|\mathbf{p}} \tag{6.13}$$

*can be expressed as a linear function of only the partial information* $\boldsymbol{\pi}'_* \triangleq \boldsymbol{\pi}'_{\mathbf{p}} M_*$.

Note that by Definition 4, a random variable $\mathsf{c}$ is spatial $M_i$-separable with respect to its parents $\mathbf{p}$ if and only if $\mathsf{p}_i$ is Fisher sufficient for inferring $\mathsf{c}$. Within this relatively new setting of Bayesian networks, $M_i$ is defined analogously to how it was defined for networks of stochastic automata, with the exception that the set of parents $\mathbf{p}$ (usually implicitly specified) now serves the role of the network state $\mathbf{x}$, i.e., $\boldsymbol{\pi}'_{\mathsf{p}_i} = \boldsymbol{\pi}'_{\mathbf{p}} M_i$. We will similarly use other shorthand notation for $M_*$ matrices originally developed in Sections 2.1.5 and 2.1.8 with respect to some parents $\mathbf{p}$.

In the case of spatial separability, note that $M_*$-separability implies $M_\#$-separability when $\mathcal{R}(M_*) \subset \mathcal{R}(M_\#)$. This was *not* the case with network separability.

There are many equivalent ways to characterize when $\mathsf{c}$ is spatial $M_*$-separable. Several of these characterizations are given in the following corollaries, which mirror the original development of equivalences to $M_*$-separability for networks of automata (Section 2.2).

**Corollary 13.** *A random variable* $\mathsf{c}$ *in a Bayesian network is spatial* $M_*$*-separable if and only if the matrix representing the conditional distribution of* $\mathsf{c}$ *given its parents* $\mathbf{p}$, $B_{\mathsf{c}|\mathbf{p}}$ *can*

[5] A variation on this additive decomposition for a conditional PMF is referred to in [30] as separability. The linear algebra connection to the notions in [30] was demonstrated in [31], but without a direct discussion of the linear constraints imposed by separability.

*be expressed as*

$$B_{\mathsf{c}|\mathbf{p}} = M_* H_* \tag{6.14}$$

*for some matrix $H_*$.*

From Corollary 13, it follows that some specific instances of spatial $M_*$-separability induce an additive form for the conditional distribution of $\mathsf{c}$ given $\mathbf{p}$. This is summarized in the following corollary.

**Corollary 14.** *Suppose that $M_*$ is a matrix that maps the joint PMF over the parents, $\boldsymbol{\pi}_{\mathbf{p}}$, to a partial information vector $\boldsymbol{\pi}_*$ that is a sequence of marginal distributions. Equivalently, assume that $M_*$ is a horizontal concatenation of selector matrices. Under the assumption of spatial $M_*$-separability, the conditional distribution of $\mathsf{c}$ given its parents $\mathbf{p}$ can be expressed as a summation of functions, one associated with each marginal distribution $\boldsymbol{\pi}_{\mathsf{p}_{i_1},\mathsf{p}_{i_2},\ldots,\mathsf{p}_{i_r}}$ in $\boldsymbol{\pi}_*$, whose only arguments are $c, p_{i_1}, p_{i_2}, \ldots, p_{i_r}$, which correspond to possible values of the random variables $\mathsf{c}, \mathsf{p}_{i_1}, \mathsf{p}_{i_2}, \ldots, \mathsf{p}_{i_r}$, respectively.*

To illustrate an example of the spatial $M_*$-separability described in Corollary 14, suppose that

$$\begin{aligned} \boldsymbol{\pi}'_* &= \boldsymbol{\pi}'_{\mathbf{p}} M_* \\ &= \begin{bmatrix} \boldsymbol{\pi}'_{\mathsf{p}_1} & \boldsymbol{\pi}'_{\mathsf{p}_2,\mathsf{p}_3} & \boldsymbol{\pi}'_{\mathsf{p}_3,\mathsf{p}_4} \end{bmatrix} \quad , \end{aligned} \tag{6.15}$$

where $\mathbf{p}$ are the parents of some random variable $\mathsf{c}$. Then by Corollary 14, spatial $M_*$-separability for $\mathsf{c}$, with $M_*$ characterized as in (6.15), ensures that the conditional distribution of $\mathsf{c}$ given $\mathbf{p}$ can be expressed as

$$\mathbb{P}(c \mid \boldsymbol{p}) = f_1(c, p_1) + f_2(c, p_2, p_3) + f_3(c, p_3, p_4) \tag{6.16}$$

for some functions $f_1(\cdot), f_2(\cdot), f_3(\cdot)$.

Next we define spatial $M_*$-sufficiency, a notion seemingly more general than spatial $M_*$-separability. Spatial $M_*$-sufficiency is characterized by a family of constraints that must be satisfied. As will be claimed in Theorem 15, a random variable $\mathsf{c}$ being spatial $M_*$-sufficient is actually equivalent to it being spatial $M_*$-separable.

**Definition 5.** *In a Bayesian network, a random variable $\mathsf{c}$ is spatial $M_*$-*sufficient *if the distribution for $\mathsf{c}$, as calculated by total probability from the conditional distribution for $\mathsf{c}$ given its parents $\mathbf{p}$ and some joint distribution over the parents $\boldsymbol{\pi}_{\mathbf{p}}$, will be unchanged for all joint distributions $\boldsymbol{\pi}_{\mathbf{p}}$ that produce the same partial information $\boldsymbol{\pi}'_* = \boldsymbol{\pi}'_{\mathbf{p}} M_*$.*

The next theorem claims that if the conditional distribution for $\mathsf{c}$ given $\mathbf{p}$ can be expressed in terms of $\boldsymbol{\pi}_*$, then it can be done so as a linear function of $\boldsymbol{\pi}_*$. The proof of this follows the same arguments equating $M_*$-separability and $M_*$-sufficiency for networks of automata evolving as time-homogeneous Markov chains (Theorem 1).

**Theorem 15.** *In a Bayesian network, a random variable $\mathsf{c}$ is spatial $M_*$-sufficient if and only if it is spatial $M_*$-separable.*

### 6.1.4 Relating spatial $M_*$-separability to past research

We have already discussed how spatial $M_*$-separability for Bayesian networks is intimately connected to Fisher sufficiency. Our notion of spatial separability is also related to Pfeffer's notion of separability [30], as well as the influence model [12, 13]. Let us begin with the former.

Pfeffer's idea of separability is equivalent to our spatial notion of $M^{(1)}$-separability. Note that we previously discussed Pfeffer's separability in the context of $M^{(1)}$-separability for networks evolving as time-homogeneous Markov chains (Section 2.2.1), as well as with respect to the specific case of a GIM (Section 4.1.1). However, it is in the current setting of Bayesian networks that Pfeffer originally defined his notion of separability (which was defined with respect to a particular random variable).

Compare the definition of $M^{(1)}$-separability for networks evolving as time-homogeneous Markov chains (Definition 1) to the definition of spatial $M^{(1)}$-separability for a particular random variable $\mathsf{c}$ in a Bayesian network (Definition 4). In making such a comparison, one may be misled to believe that network $M^{(1)}$-separability is equivalent to each random variable being spatial $M^{(1)}$-separable (in the Bayesian network sense). However, for a set of random variables evolving over discrete time (a stochastic network), if we assume that the instance of each random variable at any given time is spatial $M^{(1)}$-separable, we are inherently assuming a particular structure for the Bayesian network encoding the conditional independence relations among the instances of the random variables at different times. In particular, we would be assuming that all the random variables at a given time are conditionally independent, given their values in the immediate past.

Such conditional independence is not assumed in general for our Markovian stochastic networks. Assuming spatial $M^{(1)}$-separability for each instance of a random variable amongst a set that is collectively evolving as a Markov chain is more restrictive than the network exhibiting $M^{(1)}$-separability. However, the reader may recognize that the additional assumptions of conditional independence transform an $M^{(1)}$-separable network into an influence model. This fact is argued in [31].

## 6.2 Interpretations of spatial $M^{(1)}$-separability

As clear from Corollary 14, spatial $M^{(1)}$-separability with respect to a random variable $\mathsf{c}$ induces a special additive parametrization for the conditional distribution of $\mathsf{c}$ given its parents $\mathbf{p}$. This particular parametrization, spatial $M^{(1)}$-separability, can be related to two familiar notions: Fisher sufficiency and conditional independence. These connections will be developed in what follows.

Recall our discussion in Section 6.1.1 and the particular form of the conditional density of $\mathsf{c}$ given $\mathbf{p}$ that is induced under Fisher sufficiency. Suppose that $\mathsf{p}_1$ is sufficient for inferring $\mathsf{c}$ given its parents, or equivalently, $\mathsf{c}$ and all parents except $\mathsf{p}_1$ are conditionally independent, given $\mathsf{p}_1$. If this holds, it follows that

$$\mathbb{P}(\,c\,|\,\boldsymbol{p}\,) = \mathbb{P}(\,c\,|\,p_1, p_2, \ldots, p_n\,) = \mathbb{P}(\,c\,|\,p_1\,) \quad , \tag{6.17}$$

i.e., the conditional distribution of $\mathbf{c}$ given $\mathbf{p}$ is simply the conditional distribution of $\mathbf{c}$ given $\mathbf{p}_1$, a function of only two variables.

Suppose one considers a mixture model, i.e., a convex combination of probability models, consisting of Fisher sufficient models such as (6.17). The resulting conditional distribution of $\mathsf{c}$ given $\mathbf{p}$ would have the following form:

$$\mathbb{P}(\,c\,|\,\boldsymbol{p}\,) = \sum_{i=1}^{n} \alpha_i f_i(c, p_i) \quad , \tag{6.18}$$

which should be recognized as an instance of spatial $M^{(1)}$-separability. Conversely, it can be shown that any instance of spatial $M^{(1)}$-separability can be expressed in the form of the right hand side of (6.18), i.e., as a convex combination of nonnegative functions $f_i$ that produce a PMF for $\mathsf{c}$ when $p_i$ is fixed.[6] A proof of this fact is given in [31].

Can such a claim be extended? Is $\mathsf{c}$ being spatial $M^{(2)}$-separable also equivalent to a mixture model of Fisher sufficient models? Consider a mixture model of instances where two of $\mathsf{c}$'s parents are together Fisher sufficient:

$$\mathbb{P}(c \,|\, \boldsymbol{p}) = \sum_{i,j} \alpha_{ij} f_{ij}(c, p_i, p_j) \quad . \tag{6.20}$$

By Corollary 14, the mixture of Fisher sufficient models as given in (6.20) is spatial $M^{(2)}$-separable. However, it is unclear whether or not every spatial $M^{(2)}$-separable model (or every $M^{(r)}$-separable model, $r > 1$) can be expressed as a mixture of Fisher sufficient models. The apparent difficulty is assuring the existence of a nonnegative mixture when spatial $M^{(2)}$-separability is assumed. Evidently, spatial $M^{(r)}$-separability is closely associated with mixtures of Fisher sufficient models.

The connection that we wish to draw between spatial $M^{(1)}$-separability and conditional independence is more abstract. Consider a peculiar instance of conditional independence: that all of the parents $\mathsf{p}_1, \ldots \mathsf{p}_n$ are conditionally independent given the child $\mathsf{c}$. Such an instance of conditional independence is equivalent to the conditional distribution of the parents $\mathbf{p}$ given the child $\mathsf{c}$ having the following factorable form

$$\mathbb{P}(p_1, \ldots, p_n \,|\, c) = \prod_{i=1}^{n} g_i(p_i, c) \quad . \tag{6.21}$$

By comparing the abstract forms of (6.18) and (6.21), we see that spatial $M^{(1)}$-separability is the additive analogue to the product forms of conditional independence. Many additive parameterizations of conditional densities correspond to instances of spatial $M^{(r)}$-separability and its irregular generalizations, e.g., (6.16).

## 6.3 Advantages for inference

In many respects, the primary advantage of Bayesian networks and Markov random fields is not their potentially tractable representations via product forms, but rather the tractable inference that their product forms admit. We would briefly like to discuss how incorporating spatial $M_*$-separability into a Bayesian network may offer substantial computational benefits for inference, beyond the advantages of the structure of the Bayesian network

[6] In creating a mixture of Fisher sufficient models, each of the form of (6.17), one may be tempted to write

$$\mathbb{P}(c \,|\, \boldsymbol{p}) = \sum_{i=1}^{n} \alpha_i \mathbb{P}(c \,|\, p_i) \quad , \tag{6.19}$$

which would be misleading, as the probability law on the left hand side would not be the same as the probability laws governing the conditional PMFs on the right hand side. In particular, on the left hand side, we have the conditional PMF for the child *after* forming the mixture; each of the terms on the right hand side is a conditional PMF for the child given a single parent, *prior to* forming the mixture. To be expressed using the same probability law, each conditional PMF on the right hand side needs to be further conditioned on the event that the $i$th mixture is chosen.

alone. However, in other instances, additional assumptions of spatial $M_*$-separability for the micro-structure of the Bayesian network seemingly fail to offer any computational gains.

In a sense, there are two kinds of inference problems in Bayesian networks: the simple inference problems, where some evidence at the nodes of the oldest ancestors is propagated through the network using the conditional PMFs for each child given its parents, and those inference problems that are not so simple, e.g., some evidence regarding the children is observed, which one wants to use to update the probabilities on the ancestors. Typically only the difficult problems are referred to as inference. The simple inference problem for Bayesian networks is a generalization of information propagation that has been discussed extensively in the setting of stochastic networks evolving as a Markov chain. Hence, we can consider in the general setting of Bayesian networks many of the same examples of $M_*$-separability that were introduced for networks, including the propagation of moments, cross-correlations, etc. What we have not discussed thus far are the difficult inference problems.

Inference problems involve computing conditional probabilities; any conditional probability that one would like to compute can be expressed as a ratio of two probabilities obtained by summing specific arguments of the joint PMF over all of their possible values. In essence, difficult inference computations involve summing the joint PMF over its arguments. Belief propagation or the sum-product algorithm [87], [76], [32], the common names for a specific inference algorithm on graphical models, is essentially an efficient distributed method to compute sums over arguments of the joint PMF, that simultaneously computes probabilities conditioned on some observed evidence via the passing of messages (or 'beliefs') between neighboring nodes.

The computational complexity of belief propagation is equivalent to the computational complexity of summing over all arguments of the joint PMF. For an arbitrary joint PMF on $r$ random variables $\mathsf{x}_1, \ldots, \mathsf{x}_r$, each with an alphabet of size $m$, this computational complexity is $\Theta(m^r)$, as there are $m^r$ possible terms to sum.

If, however, $\mathsf{x}_1, \ldots, \mathsf{x}_r$ constitute a Markov chain, the task of summing over the arguments of the joint PMF is simplified:

$$\sum_{x_1,\ldots x_r} \mathbb{P}(\, x_1, \ldots x_r \,) = \sum_{x_r, x_{r-1}} \mathbb{P}(\, x_r \,|\, x_{r-1} \,) \sum_{x_{r-2}} \mathbb{P}(\, x_{r-1} \,|\, x_{r-2} \,) \ldots \sum_{x_1} \mathbb{P}(\, x_2 \,|\, x_1 \,) \, \mathbb{P}(\, x_1 \,) \quad . \tag{6.22}$$

Because of the factorable form of the Markov chain, evaluating the right hand side of (6.22) requires computation of complexity only $\Theta(rm^2)$, i.e., only exponential in the maximum number of arguments in the factors of the joint PMF in product form.

This fact—that the computational complexity of belief propagation is exponential in the maximum number of arguments for any of the factors of the joint PMF—holds true for general Bayesian networks, provided that the summations over the arguments of the joint PMF can be done in a chain-like, or more generally, tree-like manner (the running-intersection property of junction trees [41]). Otherwise, the network must be triangulated [41], and the computational complexity of belief propagation is exponential in the largest clique of the triangulated graph.

Suppose that we have a Bayesian network on random variables $\mathsf{z}$, where there is one particular large factor in the joint PMF: the conditional PMF for $\mathsf{c}$ given its $r$ parents $\boldsymbol{p} = (\mathsf{p}_1, \ldots, \mathsf{p}_r)$ (a function with $r+1$ arguments), and all other factors involve $\leq \bar{r}$ arguments, with $r \gg \bar{r}$. The computational complexity for inference on such a Bayesian network is exponential in $r$, as the maximum number of arguments in the factors is $r+1$.

If the random variable $\mathsf{c}$ is spatial $M^{(2)}$-separable, meaning that $\mathbb{P}(\, c \,|\, \boldsymbol{p} \,)$ can be expressed as

$$\mathbb{P}(\, c \,|\, \boldsymbol{p} \,) = \sum_{i,j} f_{ij}(c, p_i, p_j) \quad , \tag{6.23}$$

then the joint PMF over all random variables $\mathbf{z}$ (inclusive of both $\mathsf{c}$ and $\mathbf{p}$) can be expressed as

$$\begin{aligned} \mathbb{P}(\, \boldsymbol{z} \,) &= \mathbb{P}(\, c \,|\, \boldsymbol{p} \,) g(\boldsymbol{z}) \\ &= \Big(\sum_{i,j} f_{ij}(c, p_i, p_j)\Big) g(\boldsymbol{z}) \\ &= \sum_{i,j} f_{ij}(c, p_i, p_j) g(\boldsymbol{z}) \quad , \end{aligned} \tag{6.24}$$

where $g(\boldsymbol{z})$ includes all the factors remaining in $\mathbb{P}(\, \boldsymbol{z} \,)$ after $\mathbb{P}(\, c \,|\, \boldsymbol{p} \,)$ is factored out.

From (6.24), summing over all possible values of $\mathbf{z}$ in the joint PMF can be accomplished by individually summing each of the $\binom{r}{2}$ terms of the right hand side of (6.24). By assuming that $\mathsf{c}$ is spatial $M^{(2)}$-separable, and provided that $\overline{r} \ll r$, the computational complexity for inference has been reduced from $\Theta(m^r)$ to $O(r^2 m^{\overline{r}})$.

### 6.3.1 Limitations to the benefits of spatial separability

As illustrated in the last example, when a node with a large number of parents exhibits spatial $M^{(2)}$-separability, the computational gains for inference can be substantial. However, in the next example, assuming spatial $M^{(1)}$-separability for all nodes with a large number of parents offers no additional tractability.

We consider a Bayesian network with nodes organized into different generations. Let $d$ and $p$ be parameters that specify the size and connectivity of the Bayesian network. Within generation $i$, there are $d^i$ nodes. Each node in the $i$th generation has $p^{i-1}$ parents in the $(i-1)$th generation, with $p < d$. We do not consider finer descriptions of the connectivity of the Bayesian network graph.

If we consider $g$ generations, the computational complexity for inference on such a graph will be $\Omega(d^g m^{p^{g-1}})$, as there are $d^g$ factors of the joint PMF with $p^{g-1}$ arguments (corresponding to the random variables in the $g$th generation), before any necessary triangulation. Suppose that spatial $M^{(1)}$-separability is assumed for each of the $d^g$ random variables in the $g$th generation. Proceeding as we did in the previous example and expanding the additive forms induced by spatial $M^{(1)}$-separability for each of the random variables in the $g$th generation, what results will be an expanded expression for the joint PMF over all random variables that is a sum of $(p^{g-1})^{d^g}$ terms. As we assume that $d > p$ for our model to be reasonable, the number of terms after assuming spatial $M^{(1)}$-separability and expanding exceeds the original complexity. This failure in obtaining tractability by assuming spatial $M^{(1)}$-separability stems from the fact that the number of terms in the expansion will be exponential in the number of random variables for which spatial $M^{(1)}$-separability is assumed. Thus, assuming spatial $M^{(1)}$-separability, or spatial $M^{(r)}$-separability for any value of $r$, only offers computational benefits for inference if the number of random variables for which spatial $M^{(1)}$-separability is assumed is minimal.

## 6.4 Conclusions

This chapter offers a brief introduction to spatial separability, which extends our notion of separability to the context of general Bayesian networks. The Bayesian network specifies the macro-structure of a set of random variables, while spatial separability imposes micro-level structure, assuming special parameterized forms for the conditional PMFs that serve as the factors of the joint PMF. The connections to Fisher sufficiency, Pfeffer's separability, and the influence model were highlighted. Furthermore, we illustrated connections with mixtures of Fisher sufficient models, as well as conditional independence. An unsolved problem is the potential equivalence of spatial $M^{(r)}$-separability with mixtures of Fisher sufficient models of varying degrees.

We closed by discussing the problem of inference in the context of spatial separability. There are prototypical examples of graphical models assuming a particular parameterized form for the conditional PMFs—Boltzmann machines [88], Ising models [89], and the noisy-OR model [90], to name a few. However, general approaches for parameterizations apart from finer factorable forms (graphical models that are more sparse)—a common approach in variational methods [91]—are seemingly lacking. In a sense the parameterizations discussed in [92] do extend beyond finer factorable forms, but only as mixtures of fully-factorable forms (note that spatial $M^{(r)}$-separability can be conceived of as an extension of such mixture models). Spatial separability offers one alternative approach with substantial generality.

The are many interesting future directions for extending spatial separability. Exploring the connections between mean field theory [93] and spatial separability, particularly the instances involving the propagating of means, could potentially be beneficial. Integrating with spatial separability the approximate and optimization-based approaches to inference, e.g., variational methods, may be of value.

# Chapter 7

# Conclusion

## 7.1 Summary

This thesis introduces separability, a general modeling approach for large finite-state probabilistic models described in terms of a set of evolving random variables. Separability encompasses any instance where partial probabilistic information about the state of a Markov chain can be propagated, e.g., moments or marginal characterizations of the distribution of the state. Separability is intended for situations where the probabilistic description of the state of the Markov chain is intractable due to the exponential explosion in the size of its state space. As many examples of meaningful partial information are only polynomial in $n$ (the number of random variables characterizing the state), such partial information can be tractable even when $n$ is large. Separability offers a potentially tractable modeling paradigm in the large-$n$ regime; we illustrated examples of separability that permit efficient representation, analysis, simulation, and inference. We connected separability with several models present in the literature, including Pfeffer's separability [30], the influence model [12, 13], moment-linear stochastic systems [14, 15], Fisher sufficiency [27], probabilistic models on the symmetric group [9], and mass action kinetics [38].

After formally defining separability, we derived several important equivalences regarding separability, including the fact that whenever partial information can be propagated exactly, it can be done so linearly. Without loss of generality, we could focus on the linear propagation of partial information and use the tools of linear algebra to characterize separability. We followed this path in Chapter 3, deriving many interesting insights into separability based on linear algebra. It was demonstrated that the spectral characteristics of the matrices propagating partial information are inherited from the transition matrices of the underlying Markov chain. Separability was shown to be an instance of subspace invariance for linear systems, but relating to invariance in a nonstandard way. In Section 3.3, we offered a valuable visualization of the constraints of separability. The remainder of Chapter 3 was focused on determining the dimension of canonical instances of separability, that is, the number of free parameters necessary to specify a particular separable model. In doing so, we developed several interesting techniques for analyzing matrices composed of blocks of Kronecker products, as well as offered a linear algebraic perspective on the constraints of the running-intersection property of graphical models. It was shown that in many cases, separability is a relatively general modeling formulism for large $n$. We illustrated that as a consequence of this, in many cases, a high percentage of imposed constraints can be matched with separable models when $n$ is large.

Chapter 4 introduced the generalized influence model (GIM), a parametric model that includes as special cases both the symmetric group and the influence model. Being a parametric model, the GIM offers general, tractable representations for a specific instance of separability. Proofs of many of the properties of the symmetric group and the influence model were simplified in the general setting of the GIM. Many interesting geometric relationships of the IM, the GIM, and separability were derived and illustrated. We offered many examples of GIMs beyond influence models and the symmetric group, which take advantage of the additional modeling generality offered by the GIM.

The problem of tractable learning for separable models was considered in Chapter 5. We began by developing approaches for estimating the transition matrix of the underlying Markov chain. Even though such estimation algorithms are intractable for interesting instances of separability, the approaches for solving the full parameter learning problem were valuable, as they motivated the approach for estimating the matrices that propagate the partial information. Our proposed learning algorithm, of computational complexity polynomial in $n$, was motivated by replacing Kullback-Leibler divergences with 2-norms, ensuring that the core computational routine of the learning algorithm would be both tractable and familiar—linear least-mean-squares estimation. We established that our learning algorithm would be asymptotically consistent, that iterative estimates would almost surely eventually become and stay feasible, and we offered several ways to evaluate the appropriateness of assumptions of separability, computed as part of our learning algorithm.

The final chapter of significance extended the notions of separability to general Bayesian networks. After discussing how spatial separability is related to mixture models of Fisher sufficiency, and also serves as an additive analog to the product forms of conditional independence, we briefly discussed the potential benefits for inference when separability is assumed.

## 7.2 Future research

- **Identify additional models where assumptions of separability are appropriate** For many years, Markov had only two examples of Markov chains. We hope that our examples of separability can be similarly augmented over time, offering new opportunities and problems for applying the rich theory of separability.

- **Existence of nonnegative $H^{(r)}$ / Potential equivalence of $M^{(r)}$-separability in Bayesian networks with mixtures of Fisher sufficient models, for $r > 1$** Although there was little discussion of this fact, one can guarantee the existence of a nonnegative $H^{(1)}$ under the assumption of $M^{(1)}$-separability (a proof of this is given in [31]). Guarantees for the existence of nonnegative $H^{(r)}$ for $r > 1$, under the assumption of $M^{(r)}$-separability, is an open question. Similarly, the existence of nonnegative additive forms under $M^{(r)}$-separability in Bayesian networks, for $r > 1$, is also an open question. These problems are closely connected to the existence of nonnegative realizations in linear systems theory [94], realizations of hidden Markov models (HMMs) [95, 96], and the existence of nonnegative matrices that leave cones invariant [29, 97].

- **Relationship between the GIM and $\mathcal{M}^{(n)}$-separability** Empirically, the dimension of the set of transition matrices that are GIMs and those that exhibit $\mathcal{M}^{(n)}$-separability are equal for all cases that have been evaluated. The sets themselves are not equal, however; the set of matrices exhibiting $\mathcal{M}^{(n)}$-separability has additional extreme points. Is there a way for such additional extreme points to be incorporated into a simple parametrization that generalizes the GIM?

- **Separability's connections to the symmetric group** The first connections between the GIM and the symmetric group were demonstrated in Section 4.5. We believe such connections can be developed further. There are many powerful results from group theory; furthering or extending such results in potentially more general settings, such as general GIMs or general instances of $M_*$-separability, could be rather beneficial. On the other hand, the linear algebraic perspectives offered by $M_*$-separability offer many new interpretations of standard results from the symmetric group. Our discussions on parameter learning may yield much insight into parameter learning for the symmetric group, which is briefly mentioned in [10]. Furthermore, the structure of $M_*$-separability may also offer new insights into the Fourier based approximations pursued in [10]. At the very least, $M_*$-separability offers the natural generalization to the propagation of such partial information in cases beyond the symmetric group.

- **Parameter learning of $M_*$-separable models under more general settings** In Chapter 5 we proposed a solution to the parameter learning problem under the assumption of $M_*$-separability, when given full observations of the state. Extending this learning algorithm to situations with only partial state information would be valuable. Theoretically, this could be achieved by imposing additional restrictions on algorithms for learning HMMs, i.e., Baum-Welch [56]; however, practically, it is essential that such learning algorithms be tractable for the interesting instances of $M_*$-separability that are intractable under traditional methods of analysis.

- **Connecting instances of separability involving the propagation of expectations to mean field theory** As evident in the specific examples of mass action kinetics, and the more general mean field theory, laws of large numbers permit the close approximation of propagating expectations in a variety of models. Specific instances of $M_*$-separability precisely characterize when such expectations can be propagated. Evidently, these models become roughly separable upon reaching a sufficiently large size. It would be interesting to detail the evolution of the algebraic structure of such models as they increase in size and become roughly separable. For a general class of models, one may be able to show how laws of large numbers push such models towards the approximate propagation of moments.

- **Developing variational methods for separability** We have briefly discussed instances where separability is approximately satisfied. In addition to developing approximation techniques for models that are roughly separable, it would be valuable to use variational methods to leverage the structure of separability and develop techniques for bounding probabilities. This approach would be analogous to how finer factorable forms for the joint PMF are used in [91, 90] to obtain bounds on probabilities via substantially less complex inference computations. In our case, we would consider the additive parameterizations of separability rather than the product parameterizations associated with assuming additional conditional independence relations.

# Appendix A

# Kronecker Products

We begin by defining the Kronecker (scalar) product and mentioning some of its key properties. We proceed to develop some of the key theorems for Kronecker products. Much of the content here can be found in [28], however, our approach to the material focuses only on those concepts that are pertinent to our discussions.

**Definition 6.** *The Kronecker product of $A$ (a $p \times q$ matrix) and $B$ (an $r \times s$ matrix), denoted as $A \otimes B$, is defined as the following $pq \times qs$ matrix represented in block form:*

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1q}B \\ a_{21}B & a_{22}B & \dots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2} & \dots & a_{pq}B \end{bmatrix} \quad , \tag{A.1}$$

*where $a_{ij}$ is the $(i,j)$ element of $A$.*

The definition above is what is also used to define Kronecker products involving vectors and scalars (simply degenerate matrices). Occasionally, $A \otimes B$ is referred to as $A$ Kroneckered with $B$.

**Remark 1.** *$A \otimes B$ contains every distinct product of an element in $A$ with an element in $B$.*

**Remark 2.** *Each column of $A \otimes B$ consists of a Kronecker product of a column of $A$ with a column of $B$. All such Kronecker products involving a column of $A$ Kroneckered with a column of $B$ are represented as columns of $A \otimes B$.*

**Remark 3.** *Each row of $A \otimes B$ consists of a Kronecker product of a row of $A$ with a row of $B$. All such Kronecker products involving a row of $A$ Kroneckered with a row of $B$ are represented as rows of $A \otimes B$.*

We now state as fact some basic properties of Kronecker products.

## A.1 Properties of Kronecker products

1. **No Commutative Property**: In general $A \otimes B \neq B \otimes A$. However, by Remark 1, each entry in $A \otimes B$ appears as an entry in $B \otimes A$.

2. **Associative Property**:

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C$$

3. **Distributive Property**:

$$\begin{aligned} (A + B) \otimes C &= (A \otimes B) + (B \otimes C) \\ A \otimes (B + C) &= (A \otimes B) + (A \otimes C) \\ A \otimes (B + C) \otimes D &= A \otimes B \otimes D + A \otimes C \otimes D \end{aligned} \tag{A.2}$$

The distributive property of Kronecker products over addition follows directly from the distributive property of multiplication over addition.

4. **Mixed-product Property**:

$$(A_1 \otimes B_1)(A_2 \otimes B_2) = A_1 A_2 \otimes B_1 B_2 \tag{A.3}$$

for appropriately-sized matrices $A_1$, $A_2$, $B_1$, $B_2$.

5. **Identity**: $A \otimes 1 = 1 \otimes A = A$ (note that we have a scalar 1, not a vector of all 1s ($\mathbb{1}$)

6. **Eigenvalues and Eigenvectors**: Suppose $A$ and $B$ are square matrices. Let $\lambda$ and $\boldsymbol{x}$ be an eigenvalue and associated eigenvector of $A$. Let $\sigma$ and $\boldsymbol{v}$ be an eigenvalue and associated eigenvector of $B$. Then $\lambda\sigma$ and $\boldsymbol{x} \otimes \boldsymbol{v}$ is an eigenvalue and associated eigenvector of $A \otimes B$. In fact, all eigenvalues and associated eigenvectors of $A \otimes B$ can be composed as such from the eigenvalues and eigenvectors of $A$ and $B$.

7. **Transpose**:

$$(A \otimes B)' = A' \otimes B' \tag{A.4}$$

8. **Vectorization**:

$$\text{vec}(ABC) = (C' \otimes A)\text{vec}(B) \tag{A.5}$$

where vec$(\cdot)$ is the linear operation that converts a $p \times q$ matrix into a $pq \times 1$ vector, with the first column of the matrix serving as the first $p$ entries of the vector, the second column as the next $p$ entries of the vector, and so forth.

9. **Row Vectorization**:

$$\text{rvec}(ABC) = \text{rvec}(B)(A' \otimes C) \tag{A.6}$$

where rvec$(\cdot)$ is the linear operation that converts a $p \times q$ matrix into a $1 \times pq$ row vector, with the first row of the matrix serving as the first $q$ entries of the vector, the second row as the next $q$ entries of the vector, and so forth. Note that $\text{vec}(A)' = \text{rvec}(A')$.

**Theorem 16.** *If $\mathcal{R}(A) = \mathcal{R}(\widehat{A})$ and $\mathcal{R}(B) = \mathcal{R}(\widehat{B})$ then $\mathcal{R}(A \otimes B) = \mathcal{R}(\widehat{A} \otimes \widehat{B})$.*

*Proof.* We will only prove that $\mathcal{R}(A \otimes B) = \mathcal{R}(\widehat{A} \otimes B)$ when $\mathcal{R}(A) = \mathcal{R}(\widehat{A})$; the full statement follows by similarly proving that $\mathcal{R}(A \otimes B) = \mathcal{R}(A \otimes \widehat{B})$ when $\mathcal{R}(B) = \mathcal{R}(\widehat{B})$. It will be sufficient to show that any column in $\widehat{A} \otimes B$, say $\widehat{\boldsymbol{a}}_j \otimes \boldsymbol{b}_k$, can be expressed as a linear combination of the columns of $A \otimes B$.

As $\mathcal{R}(A) = \mathcal{R}(\widehat{A})$, it follows that $\widehat{\boldsymbol{a}}_j$ can be expressed as a linear combination of the columns of $A$:

$$\widehat{\boldsymbol{a}}_j = \sum_i \alpha_i \boldsymbol{a}_i \quad , \tag{A.7}$$

for some coefficients $\alpha_i$. By the distributive property (A.2) and the definition of the Kronecker product (Definition 6), it follows that

$$\begin{aligned}\widehat{\boldsymbol{a}}_j \otimes \boldsymbol{b}_k &= \left(\sum_i \alpha_i \boldsymbol{a}_i\right) \otimes \boldsymbol{b}_k \\ &= \sum_i \alpha_i \left(\boldsymbol{a}_i \otimes \boldsymbol{b}_k\right) \quad , \end{aligned} \tag{A.8}$$

which proves that any column in $\widehat{A} \otimes B$ can be expressed as a linear combination of the columns of $A \otimes B$. □

**Corollary 15.** *If $\mathcal{R}(A) \subset \mathcal{R}(\widehat{A})$ and $\mathcal{R}(B) \subset \mathcal{R}(\widehat{B})$ then $\mathcal{R}(A \otimes B) \subset \mathcal{R}(\widehat{A} \otimes \widehat{B})$.*

*Proof.* By our assumptions, we can always construct matrices $\widetilde{A}$ and $\widetilde{B}$ such that

$$\widetilde{A} \triangleq [A \ \ *] \quad , \tag{A.9}$$

$$\widetilde{B} \triangleq [B \ \ *] \quad , \tag{A.10}$$

where $\mathcal{R}(\widetilde{A}) = \mathcal{R}(\widehat{A})$ and $\mathcal{R}(\widetilde{B}) = \mathcal{R}(\widehat{B})$, and $*$ denotes additional columns (unspecified for our purposes). By Theorem 16, it follows that

$$\mathcal{R}(\widetilde{A} \otimes \widetilde{B}) = \mathcal{R}(\widehat{A} \otimes \widehat{B}) \quad . \tag{A.11}$$

Moreover, by Remark 2, all columns in $A \otimes B$ appear as columns of $\widetilde{A} \otimes \widetilde{B}$. The claim follows. □

**Theorem 17.** *Let $A$ and $B$ each have individually orthogonal columns, i.e. for columns $\boldsymbol{a}_i$, $\boldsymbol{a}_j$ of $A$,*

$$\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle = \boldsymbol{a}_i' \boldsymbol{a}_j = \begin{cases} \delta_i^{(A)} \neq 0 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad ,$$

*and similarly for the columns of $B$. It follows that $A \otimes B$ has orthogonal columns.*

*Proof.* Consider any two columns of $A \otimes B$: $\boldsymbol{a}_i \otimes \boldsymbol{b}_k$ and $\boldsymbol{a}_j \otimes \boldsymbol{b}_l$. By the mixed-product property of Kronecker products (A.3), it follows that

$$\begin{aligned}\langle \boldsymbol{a}_i \otimes \boldsymbol{b}_k, \boldsymbol{a}_j \otimes \boldsymbol{b}_l \rangle &= \boldsymbol{a}_i' \boldsymbol{a}_j \otimes \boldsymbol{b}_k' \boldsymbol{b}_l \\ &= \begin{cases} \delta_i^{(A)} \delta_k^{(B)} \neq 0 & \text{if } i = j \text{ and } k = l \\ 0 & \text{otherwise} \end{cases} \quad . \end{aligned} \tag{A.12}$$

□

By following the proof above but specifying that $\delta_i^{(A)} = \delta_k^{(B)} = 1$ for all $i, k$, we can make the following more specific claim.

**Corollary 16.** *If $A$ and $B$ each have individually orthonormal columns, it follows that $A \otimes B$ has orthonormal columns.*

The next two corollaries follow by combining Theorem 16 with Theorem 17.

**Corollary 17.** *Let $\mathcal{V}_A = \{\boldsymbol{v}_1^{(a)}, \boldsymbol{v}_2^{(a)}, \ldots\}$ be a set of basis vectors for $\mathcal{R}(A)$, and let $\mathcal{V}_B = \{\boldsymbol{v}_1^{(b)}, \boldsymbol{v}_2^{(b)}, \ldots\}$ be a set of basis vectors for $\mathcal{R}(B)$. Then*

$$\mathcal{V}_{A \otimes B} \triangleq \{\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \ : \ \forall \ \boldsymbol{v}_1 \in \mathcal{V}_A, \ \boldsymbol{v}_2 \in \mathcal{V}_B\} \tag{A.13}$$

*is a set of basis vectors for $\mathcal{R}(A \otimes B)$.*

**Corollary 18.** *If $A$ has rank $\eta_A$ and $B$ has rank $\eta_B$, then $A \otimes B$ has rank $\eta_A \eta_B$.*

Our theorems and corollaries, up to this point, have involved the Kronecker product of a pair of matrices. Many examples will force us to consider Kronecker products of a sequence of more than two matrices. Induction and the associative property for Kronecker products can be used to generalize these theorems and corollaries for sequences with an arbitrary number of matrices. For completeness, we list each of these generalizations.

**Theorem 16b.** *If $\mathcal{R}(A) = \mathcal{R}(\widehat{A})$, $\mathcal{R}(B) = \mathcal{R}(\widehat{B})$, $\ldots$ $\mathcal{R}(Z) = \mathcal{R}(\widehat{Z})$, then*

$$\mathcal{R}(A \otimes B \otimes \ldots \otimes Z) = \mathcal{R}(\widehat{A} \otimes \widehat{B} \otimes \ldots \otimes \widehat{Z}) \quad . \tag{A.14}$$

**Corollary 15b.** *If $\mathcal{R}(A) \subset \mathcal{R}(\widehat{A})$, $\mathcal{R}(B) \subset \mathcal{R}(\widehat{B})$, $\ldots$ $\mathcal{R}(Z) \subset \mathcal{R}(\widehat{Z})$, then*

$$\mathcal{R}(A \otimes B \otimes \ldots \otimes Z) \subset \mathcal{R}(\widehat{A} \otimes \widehat{B} \otimes \ldots \otimes \widehat{Z}) \quad . \tag{A.15}$$

**Theorem 17b.** *If $A$, $B$, $\ldots$, $Z$ each have individually orthogonal columns, then*

$$A \otimes B \otimes \ldots \otimes Z \quad , \tag{A.16}$$

*will have orthogonal columns.*

**Corollary 16b.** *If $A$, $B$, $\ldots$, $Z$ each have individually orthonormal columns, then*

$$A \otimes B \otimes \ldots \otimes Z \quad , \tag{A.17}$$

*will have orthonormal columns.*

**Corollary 17b.** *Let $\mathcal{V}_A$ be a set of basis vectors for $\mathcal{R}(A)$, let $\mathcal{V}_B$ be a set of basis vectors for $\mathcal{R}(B)$, and so forth, with $\mathcal{V}_Z$ be a set of basis vectors for $\mathcal{R}(Z)$. Then*

$$\mathcal{V}_{A \otimes \ldots \otimes Z} \triangleq \{v_1 \otimes v_2 \ : \ \forall \ \boldsymbol{v}_a \in \mathcal{V}_A, \ \boldsymbol{v}_b \in \mathcal{V}_B, \ \ldots, \ \boldsymbol{v}_z \in \mathcal{V}_Z, \} \tag{A.18}$$

*is a set of basis vectors for $\mathcal{R}(A \otimes B \otimes \ldots \otimes Z)$.*

**Corollary 18b.** *If $A$ has rank $\eta_A$, $B$ has rank $\eta_B$, $\ldots$, $Z$ has rank $\eta_z$, then $A \otimes B \otimes \ldots \otimes Z$ has rank $\eta_A \eta_B \ldots \eta_Z$.*

## A.2 Computing the rank of block Kronecker matrices

We have shown in Corollary 18b that for a matrix specified in terms of Kronecker products of smaller matrices, one can determine its rank in terms of the constituent matrices' ranks. Many of our problems require determining the rank of block matrices, in which each block is expressed as a Kronecker product of matrices; such matrices are referred to as *block Kronecker matrices.* This particular problem arises when trying to determine the rank of $M^{(r)}$ (Sections 3.4.1 and 3.4.2), as well as when attempting to calculate the number of free parameters in specifying the transition matrix of an $M_*$-separable model (Section 3.2.1), or an $\mathcal{M}$-separable model (Section 3.2.2). We describe here an analytical approach to derive, under specific conditions, a mathematical expression for the rank of a matrix in block form, where each block is expressed as a sequence of Kronecker products, in terms of the constituent matrices' ranks.

### A.2.1 General approach

Consider the $p \times q$ matrix

$$\begin{bmatrix} A_1 & A_2 & & A_r \\ \otimes & \otimes & & \otimes \\ B_1 & B_2 & & B_r \\ \otimes & \otimes & \cdots & \otimes \\ \vdots & \vdots & & \vdots \\ \otimes & \otimes & & \otimes \\ Z_1 & Z_2 & & Z_r \end{bmatrix}, \tag{A.19}$$

where each constituent matrix $A_i$ has $\eta_A$ rows, each matrix $B_i$ has $\eta_B$ rows, and so on, meaning that $p = \eta_A \eta_B \cdots \eta_Z$ (we express chains of Kronecker products vertically, in part to conserve space, but also to emphasize the perspective that each column of the resulting matrix is a Kronecker product of columns of the constituent matrices). It is assumed that there is a constant upper bound on the number of rows and columns of each constituent matrix, i.e., there are $O(\log p)$ matrices in each Kronecker product, and a total of $r = O(\log q)$ blocks.

*Common Orthonormal Bases Assumption*: The set of matrices $\{A_i\}_{i=1}^{r}$ each with $\eta_A$ rows satisfies the common orthonormal basis assumption provided that there exists an orthonormal basis $\mathcal{V}_A$ for $\mathbb{R}^{\eta_A}$ such that for every $i$, $\mathcal{R}(A_i)$ can be expressed as the span of a subset of the basis vectors in $\mathcal{V}_A$ .

If the orthonormal basis assumption individually holds for the set of matrices $\{A_i\}_{i=1}^{r}$, for the set of matrices $\{A_i\}_{i=1}^{r}$, ..., and for the set of matrices $\{Z_i\}_{i=1}^{r}$, then there is always an $O(q(\log q)(\log p))$ algorithm to determine the rank of the matrix in (A.19). For many of the examples that we will consider when such an assumption holds, one can also derive a simple analytical expression for such a matrix's rank in terms of the ranks of the constituent matrices. The approach is as follows:

1. Replace each $A_i$ with a matrix $\widehat{A}_i$, whose columns consist of the subset of basis vectors from $\mathcal{V}_A$ that ensures $\mathcal{R}(A_i) = \mathcal{R}(\widehat{A}_i)$. Do the same for each $B_i$, and so forth, up to and including $Z_i$. In this modified matrix, each constituent matrix, e.g., $\widehat{A}_i$ or $\widehat{Z}_j$, will have individually orthonormal columns. By Theorem 16b, this new matrix has the

same range (and rank) as the original matrix. This step will require $O((\log q)(\log p))$ operations.

2. By composing the matrix as described, the columns in each block will be orthonormal (see Theorem 16b). Moreover, as each column can be expressed as

$$\boldsymbol{v}_A \otimes \boldsymbol{v}_B \otimes \ldots \boldsymbol{v}_Z \quad , \tag{A.20}$$

with $\boldsymbol{v}_A \in \mathcal{V}_A$, $\boldsymbol{v}_B \in \mathcal{V}_B$, ..., $\boldsymbol{v}_Z \in \mathcal{V}_Z$, the columns in different blocks will either be either equal or orthonormal. Therefore, the rank of the matrix in (A.19) will equal the number of unique columns in the modified matrix. To count the number of unique columns in the modified matrix, build a list of the columns in the new matrix, identifying each column by its constituent basis vectors, i.e., by a length $r$ vector that indicates the particular basis vector in $\mathcal{V}_A$, the particular basis vector in $\mathcal{V}_B$, and so on, which generate the column as represented in the form of (A.20). Building this list requires $O(q \log p)$ operations (and also storage of size $O(q \log p)$).

3. Sort the list and remove duplicates, thereby determining the number of unique columns in the modified matrix and consequently the rank of the matrix of (A.19). This would require $O(q(\log q)(\log p))$ operations.

**Remark 4.** *In order to ensure that the aforementioned common orthonormal bases assumptions is satisfied, a sufficient condition is for there to exist a nesting of the range spaces for the $A_i$, a nesting of the range spaces of the $B_i$, and so forth. Specifically, there should exist a permutation $\pi_A(\cdot)$ on the integers 1 through $n$, such that*

$$\mathcal{R}(A_{\pi_A(1)}) \subset \mathcal{R}(A_{\pi_A(2)}) \subset \ldots \subset \mathcal{R}(A_{\pi_A(n)}) \quad . \tag{A.21}$$

*This ensures a common orthonormal basis in $\mathbb{R}^{\eta_A}$ for each of the $A_i$ that can be constructed as follows: first find an orthonormal basis for $\mathcal{R}(A_{\pi_A(1)})$, then augment this basis to form a (potentially) larger orthonormal basis for $\mathcal{R}(A_{\pi_A(2)})$, and continue repeating this process while iterating through the nested range spaces. A common orthonormal basis is possible because of the nesting of the range spaces.*

*Similarly, there should exist a permutation $\pi_B(\cdot)$ on the integers 1 through $n$ such that*

$$\mathcal{R}(B_{\pi_B(1)}) \subset \mathcal{R}(B_{\pi_B(2)}) \subset \ldots \subset \mathcal{R}(B_{\pi_B(n)}) \quad , \tag{A.22}$$

*which would permit one to construct an orthonormal basis that can be used to represent $\mathcal{R}(B_i)$ for each $i$. Similarly, there should exist permutations from which one can define a nesting of the range spaces for all constituent matrices of a common level in different blocks, up to and including the $Z_i$ matrices.*

**Remark 5.** *The common orthonormal basis assumption is closely related to extensions of consistent probability measures, which themselves share an equivalence to Markov random fields. From the perspective of linear algebra, the common orthonormal basis assumption characterizes the running intersection property of Markov random fields [41], as well as the sufficient conditions for the extendability of measures [42, 43, 44]. A Markov random field can be transformed into a junction tree, whose nodes correspond to the cliques in the original Markov random field, and any two nodes in the junction tree are connected if they share the same random variables (as single nodes in the original Markov random field could belong* *to several cliques). We identify each node in the junction tree with a clique in the Markov random field. Let $\mathcal{C}$ denote the set of cliques in the Markov random field, and equivalently, the set of nodes in the junction tree. Each node in the junction tree $C \in \mathcal{C}$ has an associated clique marginal $\phi_C$, which for the case of finite alphabet random variables, can be obtained by a marginalizing operation from the joint probability vector via a matrix $M_C$. We have a junction tree (i.e., no cycles) if and only if the set of matrices $M_C$ for $C \in \mathcal{C}$ satisfies the common orthonormal basis assumption.*

*From the perspective of extending measures, suppose for a set of random variables $\mathrm{x}_1, \ldots, \mathrm{x}_n$, one is given a set of* consistent *marginal distributions, $\phi_i$ for $i \in \mathcal{I}$, over specified random variables. For example, $\phi_1$ may be a marginal distribution for $\mathrm{x}_1$ and $\mathrm{x}_2$, while $\phi_2$ may be a marginal distribution for $\mathrm{x}_2$, $\mathrm{x}_3$, and $\mathrm{x}_5$, and so forth. In order to be* consistent, *marginal distributions over common random variables must specify the same marginals. For example, under the proposed forms for $\phi_1$ and $\phi_2$ just given, both marginal distributions must specify the same marginal distribution for $\mathrm{x}_2$ in order to be consistent. Both Vorob'ev and Kellerer sought to identify necessary and sufficient conditions such that* for any *consistent marginals of the form $\phi_i$, $i \in \mathcal{I}$, there exists a joint distribution with such specified marginals. We say that the marginal information $\phi_i$, $i \in \mathcal{I}$ can be* extended. *What they discovered is that such marginals $\phi_i$ for $i \in \mathcal{I}$ must correspond to clique marginals in a junction tree. Again, from the perspective of linear algebra, if $M_i$ is the marginalizing matrix that provides the marginal distribution $\phi_i$, then the set of marginals $\phi_i$ for $i \in \mathcal{I}$ can always be extended if and only if $M_i$ for $i \in \mathcal{I}$ satisfy the common orthonormal basis assumption. Markov random fields offer a direct means to represent one such joint distribution that satisfies such marginals (the one of maximum entropy): the product of the clique marginals, divided by the product of the separator marginals [76].*

### A.2.2 Analytical expressions for special cases

There a few specific theorems to introduce regarding the rank of block Kronecker matrices that in essence follow the outlined approach. These theorems will be helpful when trying to calculate the dimension of the set of stochastic transition matrices exhibiting $M_*$-separability or $\mathcal{M}$-separability.

**Theorem 18.** *For any matrices $A_1$, $A_2$, $B_1$, $B_2$, each with $\eta$ rows, such that $\mathcal{R}(A_1) \subset \mathcal{R}(A_2)$, the matrix*

$$\left[\; A_1 \otimes B_1 \quad A_2 \otimes B_2 \;\right] \tag{A.23}$$

*will have rank $\rho_{A_1}(\rho_{B_1} + \rho_{B_2} - \widetilde{\rho}_{B_1,B_2}) + (\rho_{A_2} - \rho_{A_1})\rho_{B_2}$, where $\rho_X = \operatorname{rank}(X)$ and $\widetilde{\rho}_{X,Y} = \widetilde{\rho}_{Y,X}$ is the dimension of $\mathcal{R}(X) \cap \mathcal{R}(Y)$ (the intersection of the column spaces of $X$ and $Y$). Equivalently, its left null space will be of dimension $\eta^2 - (\rho_{A_1}(\rho_{B_1} + \rho_{B_2} - \widetilde{\rho}_{B_1,B_2}) + (\rho_{A_2} - \rho_{A_1})\rho_{B_2})$.*

*Proof.* By considering Theorem 16, we can assume that $A_1$ and $B_2$ each individually have orthonormal columns, and also because of Theorem 16, in the place of (A.23), we can equivalently compute the rank of

$$\left[\; A_1 \otimes B_1 \quad [A_1 \;\; A_{2\perp 1}] \otimes B_2 \;\right] \quad , \tag{A.24}$$

where

$$\mathcal{R}\Big( [A_1 \;\; A_{2\perp 1}] \Big) = \mathcal{R}(A_2) \quad . \tag{A.25}$$

Because of the assumed nesting $\mathcal{R}(A_1) \subset \mathcal{R}(A_2)$, there exists such a matrix $A_{2\perp 1}$ satisfying (A.25), and moreover, we can define it to be a matrix with orthonormal columns such that

$$\mathcal{R}(A_{2\perp 1}) = \mathcal{R}(A_2) \cap \mathcal{R}(A_1^{\perp}) \quad . \tag{A.26}$$

We can rearrange (A.24) and again invoke Theorem 16 to construct the following matrix with the same range space as (A.23):

$$\left[ \begin{array}{cc} A_1 \otimes B_{1,2} & A_{2\perp 1} \otimes B_2 \end{array} \right] \quad , \tag{A.27}$$

where $B_{1,2}$ is a matrix with orthonormal columns, satisfying $\mathcal{R}([B_1 \;\; B_2]) = \mathcal{R}(B_{1,2})$. As $A_1$, $B_{1,2}$, $A_{2\perp 1}$, and $B_2$ each have individually orthonormal columns, by Theorem 16, each block of (A.27) has individually orthonormal columns. Moreover, as each column in $A_1$ is orthogonal to each column in $A_{2\perp 1}$ by design (A.26), it follows that all columns in the matrix of (A.27) are orthonormal. Hence, its rank, and consequently the rank of (A.23) must be its number of columns, which can be counted as:

$$\rho_{A_1}(\rho_{B_1} + \rho_{B_2} - \widetilde{\rho}_{B_1,B_2}) + (\rho_{A_2} - \rho_{A_1})\rho_{B_2} \quad . \tag{A.28}$$

□

A similar claim (with a slightly different resulting expression for the rank) can be developed for when $\mathcal{R}(B_1) \subset \mathcal{R}(B_2)$.

**Corollary 19.** *For any matrices $A_1$, $A_2$, $B_1$, $B_2$, each with $\eta$ rows, such that $\mathcal{R}(B_1) \subset \mathcal{R}(B_2)$, the matrix*

$$\left[ \begin{array}{cc} A_1 \otimes B_1 & A_2 \otimes B_2 \end{array} \right] \tag{A.29}$$

*will have rank $\rho_{B_1}(\rho_{A_1} + \rho_{A_2} - \widetilde{\rho}_{A_1,A_2}) + (\rho_{B_2} - \rho_{B_1})\rho_{A_2}$. Equivalently, its left null space will be of dimension $\eta^2 - (\rho_{B_1}(\rho_{A_1} + \rho_{A_2} - \widetilde{\rho}_{A_1,A_2}) + (\rho_{B_2} - \rho_{B_1})\rho_{A_2})$.*

As the proof is so similar to that of Theorem 18, it is omitted.

Theorem 18 and Corollary 19 require a nesting of the range spaces for the matrices at one level, i.e., either $\mathcal{R}(B_1) \subset \mathcal{R}(B_2)$ or $\mathcal{R}(A_1) \subset \mathcal{R}(A_2)$, but not both. Recall that the general approach outlined for determining the rank of block Kronecker matrices depends on the common orthonormal bases assumption, which was discussed in Section A.2.1. This common orthonormal bases assumption required a nesting of the range spaces at all levels. However, as is evident in Theorem 18 and Corollary 19, such nesting is technically required at all levels but one, provided that the dimension of the intersections of the range spaces at the final level are known.

We have a last theorem to share for a final special case, which will be useful when trying to determine the dimension of the set of transition matrices exhibiting $\mathcal{M}$-separability.

**Theorem 19.** *If $\mathcal{R}(A_1) \subset \mathcal{R}(A_2) \subset \ldots \subset \mathcal{R}(A_r) \subset \mathbb{R}^{\eta}$, and $\mathcal{R}(B_r) \subset \mathcal{R}(B_{r-1}) \subset \ldots \subset \mathcal{R}(B_1) \subset \mathbb{R}^{\eta}$, then*

$$\left[ \begin{array}{cccc} A_1 & A_2 & & A_r \\ \otimes & \otimes & \ldots & \otimes \\ B_1 & B_2 & & B_r \end{array} \right] \tag{A.30}$$

*will have rank $\rho_{A_1}\rho_{B_1} + \sum_{l=2}^{r}(\rho_{A_l} - \rho_{A_{l-1}})\rho_{B_l}$. Equivalently, its left null space will be of dimension $\eta^2 - (\rho_{A_1}\rho_{B_1} + \sum_{l=2}^{r}(\rho_{A_l} - \rho_{A_{l-1}})\rho_{B_l})$.*

*Proof.* We follow an approach similar to Theorem 18, where, motivated by Theorem 16, we will define a new matrix with the same range space and orthonormal columns. First, define $\widehat{A}_1$ as a matrix with orthonormal columns such that

$$\mathcal{R}(\widehat{A}_1) = \mathcal{R}(A_1) \quad . \tag{A.31}$$

Next, for each $i \geq 2$, define $\widehat{A}_i = \begin{bmatrix}\widehat{A}_{i-1} & A_{i\perp i-1}\end{bmatrix}$ as a matrix with orthonormal columns such that

$$\mathcal{R}(\widehat{A}_i) = \mathcal{R}(A_i) \quad . \tag{A.32}$$

This is possible because of the nesting, i.e., $\mathcal{R}(A_{i-1}) \subset \mathcal{R}(A_i)$. Note that the columns of each matrix $\widehat{A}_i$ are a subset of the orthonormal columns of $\widehat{A}_r$.

Similarly, define $\widehat{B}_r$ as a matrix with orthonormal columns such that

$$\mathcal{R}(\widehat{B}_r) = \mathcal{R}(B_r) \quad . \tag{A.33}$$

For each $i < r$, define $\widehat{B}_i = \begin{bmatrix}\widehat{B}_{i+1} & B_{i\perp i+1}\end{bmatrix}$ as a matrix with orthonormal columns such that

$$\mathcal{R}(\widehat{B}_i) = \mathcal{R}(B_i) \quad . \tag{A.34}$$

Note that the columns of each matrix $\widehat{B}_i$ are a subset of the orthonormal columns of $\widehat{B}_1$.

By Theorem 16, the matrix

$$\begin{bmatrix} \widehat{A}_1 & \begin{bmatrix}\widehat{A}_1 & \widehat{A}_{2\perp 1}\end{bmatrix} & \begin{bmatrix}\widehat{A}_2 & \widehat{A}_{3\perp 2}\end{bmatrix} & & \begin{bmatrix}\widehat{A}_{r-1} & \widehat{A}_{r\perp r-1}\end{bmatrix} \\ \otimes & \otimes & \otimes & \cdots & \otimes \\ \widehat{B}_1 & \widehat{B}_2 & \widehat{B}_3 & & \widehat{B}_r \end{bmatrix} \tag{A.35}$$

has the same range space as (A.30). As $\widehat{A}_i \otimes \widehat{B}_i = \widehat{A}_i \otimes \begin{bmatrix}\widehat{B}_{i\perp i+1}\widehat{B}_{i+1}\end{bmatrix}$ contains as its columns all of the columns of $\widehat{A}_i \otimes \widehat{B}_{i+1}$, such duplicate columns in (A.35) can be removed for each $i$, to define a new matrix with the same range:

$$\begin{bmatrix} \widehat{A}_1 & \widehat{A}_{2\perp 1} & \widehat{A}_{3\perp 2} & & \widehat{A}_{r\perp r-1} \\ \otimes & \otimes & \otimes & \cdots & \otimes \\ \widehat{B}_1 & \widehat{B}_2 & \widehat{B}_3 & & \widehat{B}_r \end{bmatrix} \quad . \tag{A.36}$$

The matrix of (A.36) has orthonormal columns, which is easily verified. Therefore, the rank of the matrix illustrated in (A.36), which equals the rank of the matrix of (A.30), is simply its number of columns:

$$\rho_{A_1}\rho_{B_1} + \sum_{l=2}^{r}(\rho_{A_l} - \rho_{A_{l-1}})\rho_{B_l} \quad . \tag{A.37}$$

□

# Appendix B

# Asymptotic Equalities

## B.1 Asymptotic notation definitions

There are several asymptotic equivalences invoked in Section 3.5 that we be carefully proved here. We begin with the definition of big $O$ notation, big $\Theta$ notation, and asymptotic equality ($\sim$):[98]

**Definition 7.** *A function $g(n) = O(f(n))$ if there exist constants $N$ and $M$ such that*

$$|g(n)| \leq M|f(n)| \quad \forall\, n \geq N \quad . \tag{B.1}$$

**Definition 8.** *A function $g(n) = \Theta(f(n))$ if there exist constants $N$, $M_0 > 0$, $M_1$ such that*

$$M_0|f(n)| \leq |g(n)| \leq M_1|f(n)| \quad \forall\, n \geq N \quad . \tag{B.2}$$

**Definition 9.** *For functions $f(n)$ and $g(n)$, $f(n) \sim g(n)$ if*

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} \to 1 \quad . \tag{B.3}$$

## B.2 Bounds on binomial coefficients and binomial probabilities

Our first focus will be on binomial coefficients, $\binom{n}{r}$, for fixed $r$. An upper bound on $\binom{n}{r}$ is straightforward,

$$\begin{aligned} \binom{n}{r} &= \frac{n(n-1)\dots(n-r+1)}{r!} \\ &\leq \frac{n^r}{r!} \\ &= \Theta(n^r) \quad , \end{aligned} \tag{B.4}$$

as is a lower bound,

$$\begin{aligned}\binom{n}{r} &= \frac{n(n-1)\dots(n-r+1)}{r!}\\ &\geq \frac{(n-r+1)^r}{r!}\\ &= \Theta(n^r) \quad , \end{aligned}\tag{B.5}$$

and thus for fixed $r$,

$$\binom{n}{r} = \Theta(n^r). \tag{B.6}$$

Next, consider a binomial coefficient $\binom{n}{r}$, with $r$ permitted to vary with $n$. In order to bound $\binom{n}{r}$, we consider bounds for factorials. The familiar form of Stirling's formula states that [50]

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad . \tag{B.7}$$

However, as we require bounds on $n!$, we need to understand the rate of convergence in (B.7). By [99], we learn that

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}} \quad . \tag{B.8}$$

Using (B.8), the following upper bound on $\binom{n}{r}$, when $r$ varies with $n$, can be derived.

$$\begin{aligned}\binom{n}{r} &\leq \binom{n}{n/2}\\ &\leq \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}}{\sqrt{2\pi\frac{n}{2}}\left(\frac{n/2}{e}\right)^{n/2} e^{\frac{1}{12(n/2)+1}} \cdot \sqrt{2\pi\frac{n}{2}}\left(\frac{n/2}{e}\right)^{n/2} e^{\frac{1}{12(n/2)+1}}}\\ &= \sqrt{\frac{2}{\pi n}}2^n e^{\frac{1}{12n}-\frac{2}{6n+1}}\\ &= \sqrt{\frac{2}{\pi n}}2^n 2^{(\frac{1}{12n}-\frac{2}{6n+1})\log_2 e}\\ &= \sqrt{\frac{2}{\pi n}}2^{n+O(1/n)} \quad . \end{aligned}\tag{B.9}$$

In addition to bounds on binomial coefficients, we wish to derive a bound for binomial probabilities. Consider a random variable $\mathsf{z}_n$, which is a binomial random variable with parameters $\rho \in (0,1)$ and $n$. We wish to bound $\max_k \mathbb{P}(\mathsf{z}_n = k)$ as a function of $n$. As the distribution of $\mathsf{z}_n$ can be expressed as the convolution of the distribution of $\mathsf{z}_{n-1}$ with that of a Bernoulli random variable with parameter $\rho$, clearly $\max_k \mathbb{P}(\mathsf{z}_n = k)$ must be a decreasing function in $n$. Without loss of generality, we assume that $\rho n$ is an integer.

A fairly standard result is that the mode of a binomial random variable with parameters $\rho$ and $n$ will be $\lfloor \rho(n+1) \rfloor$ (in some cases the mode will not be unique, but this is irrelevant for our purposes) [46]. Consequently,

$$\max_k \mathbb{P}(\mathsf{z}_n = k) = \binom{n}{\lfloor \rho(n+1) \rfloor} \rho^{\lfloor \rho(n+1) \rfloor} (1-\rho)^{n - \lfloor \rho(n+1) \rfloor}$$
$$= \binom{n}{\rho n} \rho^{\rho n} (1-\rho)^{(1-\rho)n} \frac{\rho}{1-\rho} \quad . \tag{B.10}$$

Using (B.8), we can upper bound $\binom{n}{\rho n}$ as

$$\binom{n}{\rho n} < \frac{1}{\sqrt{2\pi n \rho(1-\rho)}} \rho^{-\rho n} (1-\rho)^{-(1-\rho)n} \quad , \tag{B.11}$$

and introducing this bound into (B.10), we obtain

$$\max_k \mathbb{P}(\mathsf{z}_n = k) < \frac{1}{\sqrt{2\pi n \rho(1-\rho)}} \rho^{-\rho n} (1-\rho)^{-(1-\rho)n} \rho^{\rho n} (1-\rho)^{(1-\rho)n}$$
$$= \frac{1}{\sqrt{2\pi n \rho(1-\rho)}} \quad , \tag{B.12}$$

meaning that $\max_k \mathbb{P}(\mathsf{z}_n = k) = O\left(\frac{1}{\sqrt{n}}\right)$. We can similarly upper bound $\binom{n}{\rho n}$ to show that

$$\max_k \mathbb{P}(\mathsf{z}_n = k) = \Theta\left(\frac{1}{\sqrt{n}}\right) \quad . \tag{B.13}$$

This result seems reasonable, considering that Chernoff bounds for binomial random variables [47] indicate that an interval of length $O(\sqrt{n})$ around the mean $\rho n$ can include any fraction $\alpha \in (0,1)$ of the mass of the random variable $\mathsf{z}_n$.

## B.3 Asymptotic equalities for expressions with $n$ in the exponent

In Section 3.5, our concern is showing that various expressions are $m^{n+g(n)}$, where $g(n)$ is sufficiently small asymptotically for our purposes. We catalogue for quick reference several such equalities in the following theorem.

**Theorem 20.** *The following asymptotic equalities hold, where $m > 1$, $\underline{m} \geq 1$, $k$, and $\alpha \in [0,1)$ are constants, with $\underline{m} < m$:*

$$m^n + \Theta(n^k) = m^{n+O(1/n)} \tag{B.14}$$
$$m^n + m^{\alpha n} = m^{n+O(1/n)} \tag{B.15}$$
$$m^n + \underline{m}^n = m^{n+O(1/n)} \tag{B.16}$$
$$\Theta(n^k) m^n = m^{n+\Theta(\log n)} \quad \text{if } k > 0 \tag{B.17}$$
$$m^n + \Theta(n^k) m^n = m^{n+O(n^k)} \tag{B.18}$$

*Proof.* The techniques to prove the asymptotic equalities (B.14)–(B.18) are similar. One fundamental result that we will use repeatedly in our derivations is that

$$\log(1+x) \le x \quad . \tag{B.19}$$

For many of the derivations, it suffices to settle for weak bounds. Proof of (B.15):

$$\begin{aligned} m^n + m^{\alpha n} &= m^n(1+m^{-(1-\alpha)n}) \\ &= m^n m^{\log_m(1+m^{-(1-\alpha)n})} \\ &\le m^{n+m^{-(1-\alpha)n}} \\ &= m^{n+O(1/n)} \end{aligned} \tag{B.20}$$

It should be evident that (B.14) follows from (B.15). Moreover, as

$$m^n + \underline{m}^n = m^n + m^{n\log_m \underline{m}} \quad , \tag{B.21}$$

where $\log_m \underline{m} \in [0,1)$, it should also be clear that (B.16) also follows from (B.15).

Proof of (B.17):

$$\begin{aligned} \Theta(n^k)m^n &= m^n m^{\log_m \Theta(n^k)} \\ &= m^n m^{k\log_m(\Theta(n))} \\ &= m^{n+\Theta(\log n)} \end{aligned} \tag{B.22}$$

Proof of (B.18):

$$\begin{aligned} m^n + \Theta(n^k)m^n &= m^n(1+\Theta(n^k)) \\ &= m^n m^{\log_m(1+\Theta(n^k))} \\ &= m^{n+O(n^k)} \end{aligned} \tag{B.23}$$

□

# Appendix C

# Random Affine Subspaces

Before discussing the number of orthants that a random affine subspace may intersect, we first must define what it means to select an affine subspace uniformly at random. First, one draws a subspace uniformly at random, by independently drawing the orthogonal directions for the subspace from a rotationally symmetric distribution, e.g., choosing the orthogonal directions uniformly over a hypersphere. Then, to obtain an affine subspace chosen uniformly at random, the chosen subspace must be shifted according to a rotationally symmetric probability density function (so that any particular shift occurs with measure 0). A multivariate Gaussian composed of independent and identically distributed scalar Gaussian random variables is rotationally symmetric, and thus provides all of the necessary machinery to specify an affine subspace drawn uniformly at random. From this point onwards, it will be assumed that all affine subspaces are drawn uniformly at random.

To build some intuition into the question we pursue here, consider $\mathbb{R}^2$, and a 1-dimensional affine subspace (a line). Almost surely, a line will intersect three out of the four orthants. In $\mathbb{R}^3$, however, a 1-dimensional affine subspace will intersect 4 out of the 8 orthants almost surely. On the other hand, a 0-dimensional affine subspace (a point) will intersect 1 orthant, almost surely, regardless of the underlying dimension. Can this be generalized?

Indeed, we can generalize this idea to any $d$-dimensional subspace in $\mathbb{R}^n$. Let's begin by considering the 1-dimensional affine subspaces (lines). Let $z$ be a Cartesian coordinate in $\mathbb{R}^n$, and note that the orthants can be divided into $2^{n-1}$ pairs, with each pair of orthants being neighbors whose signatures (how an orthant is identified, i.e., the characteristic positive/negative sequence for coordinates of points in the orthant) are the same except differing at $z$. For example, in $\mathbb{R}^3$, with $z$ being the vertical coordinate direction, the orthants 'on top of one another' would correspond to a pair. Consider the slice of $\mathbb{R}^n$ corresponding to $z = 0$, which divides $\mathbb{R}^n$ into two half spaces, with each half space containing one orthant in each pair. Obviously, such a slice is a subspace isomorphic to $\mathbb{R}^{n-1}$. A line in $\mathbb{R}^n$ will intersect this slice in one point almost surely. Such a point will be a member of both orthants for a particular pair (if orthants are considered closed). On the other hand, for all other orthant pairs, the line will pass through only one orthant of the pair (almost surely). If we consider the projected line in $\mathbb{R}^{n-1}$ by simply dropping the $z$ coordinate, it should be evident that we have defined a 1-dimensional affine subspace chosen uniformly at random in $\mathbb{R}^{n-1}$. By projecting down into $\mathbb{R}^{n-1}$, each pair of orthants in $\mathbb{R}^n$ becomes a single orthant in $\mathbb{R}^{n-1}$. The number of orthants that the projected line passes through in $\mathbb{R}^{n-1}$ is exactly the number of orthnat pairs in $\mathbb{R}^n$ through which the unprojected line passes. Hence, the number orthants that the unprojected line in $\mathbb{R}^n$ passes through must equal the number of

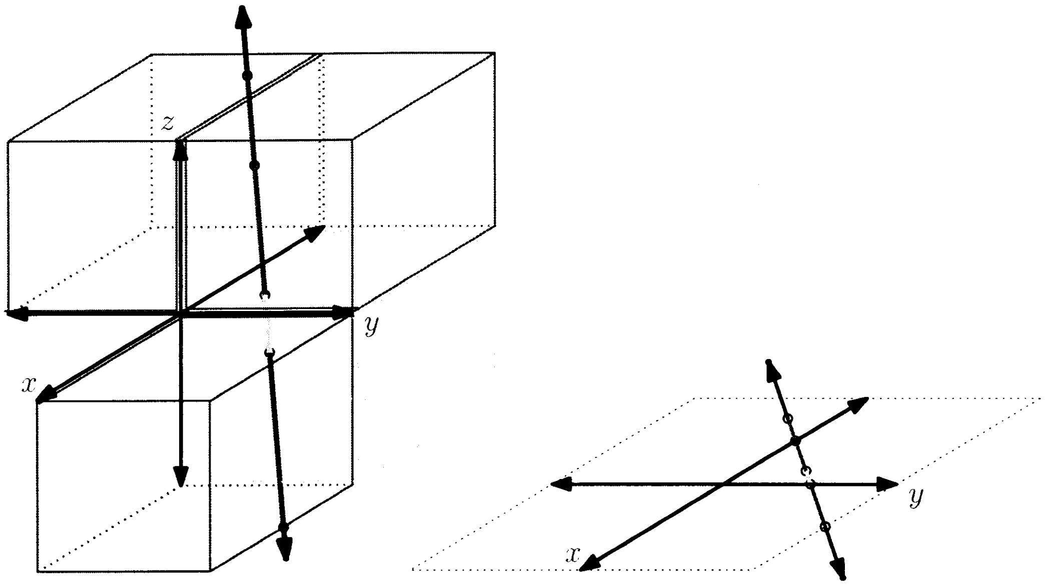


Figure C-1: Illustrating a 1-dimensional affine subspace in $\mathbb{R}^3$. The orthants that the affine subspace intersects are colored, and the line segments of the affine subspace within each orthant are colored appropriately. Note the two paired orthants through which the line passes as its $z$ coordinate goes from negative to positive. The projected 1-dimensional affine subspace in $\mathbb{R}^2$ is also shown.

orthants that the projected line passes through in $\mathbb{R}^{n-1}$ plus one (as the line passes through both orthants of a particular pair in $\mathbb{R}^n$). This gives us the following recursion

$$o_1^n = o_1^{n-1} + 1 \quad , \tag{C.1}$$

where $o_d^n$ is the number of orthants that a $d$-dimensional affine subspace chosen uniformly at random almost surely intersects. As $o_1^1 = 2$, evidently,

$$o_1^n = n + 1 \quad . \tag{C.2}$$

Fig. C attempts to illustrate these relationships in $\mathbb{R}^3$.

The idea for higher-dimensional affine subspaces is effectively the same. Consider a $d$-dimensional affine subspace in $\mathbb{R}^n$. Again, consider a coordinate $z$ and the induced orthant pairs. Taking the slice of $\mathbb{R}^n$ at $z = 0$ defines a $(n - 1)$-dimensional subspace, and the intersection of the $d$-dimensional affine subspace with the $(n-1)$-dimensional slice is almost surely a $(d-1)$-dimensional affine subspace, that can be thought to be chosen uniformly at random in $\mathbb{R}^{n-1}$. The only orthant pairs in $\mathbb{R}^n$ for which the $d$-dimensional affine subspace will intersect both, correspond to the orthants in $\mathbb{R}^{n-1}$ intersected by the uniformly chosen $(d - 1)$-dimensional affine subspace. We can now count the total number of orthants that the $d$-dimensional subspace in $\mathbb{R}^n$ intersects by simply counting the orthant pairs that it intersects. To accomplish this, we consider the projection of the $d$-dimensional affine subspace into $\mathbb{R}^{n-1}$ by eliminating the $z$-coordinate. Such a projection will itself be a $d$-dimensional subspace almost surely, and can be thought of as a $d$-dimensional subspace in $\mathbb{R}^{n-1}$ chosen uniformly at random. The number of orthants intersected by the projection of the $d$-dimensional subspace into $\mathbb{R}^{n-1}$ will be equal to, almost surely, the number of orthant pairs in $\mathbb{R}^n$ that the $d$-dimensional affine subspace intersects. Evidently,

$$o_d^n = o_{d-1}^{n-1} + o_d^{n-1} \quad . \tag{C.3}$$

Given the initial conditions to this recursion, $o_0^n = 1$ for all $n$, it can be verified (but probably is not immediately obvious) that

$$o_d^n = \sum_{k=0}^{d} \binom{n}{k} \quad . \tag{C.4}$$

# Appendix D

# Cone Theory

Many of the topics of this appendix, as well as its notation, are borrowed from [29]. The proofs of most theorems are omitted.

**Definition 10.** *A **cone** $\mathcal{K}$ is a nonempty subset of $\mathbb{R}^n$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$ and $\alpha_1, \alpha_2 \geq 0$, $\alpha_1 \boldsymbol{x} + \alpha_2 \boldsymbol{y} \in \mathcal{K}$.*

For any nonempty set $\mathcal{S} \subset \mathbb{R}^n$, we can define the cone $\mathcal{S}^G$ as the cone *generated* by $S$, consisting of all nonnegative linear combinations of the elements of $\mathcal{S}$. Note that by definition, cones are convex sets, and that extreme cases of cones include both $\{0\}$ and $\mathbb{R}^n$. A cone that we will repeatedly encounter is the nonnegative orthant in $n$-dimensional space, denoted as $\mathbb{R}^n_+$.

## D.1 Cone properties

We now cover some of the important properties for cones.

**Theorem 21.** *Cones are closed under intersections.*

**Definition 11** (reworded 2.4 from Ch. 1 of [29]). *A cone $\mathcal{K}$ is **polyhedral** if there exists a finite set $\mathcal{S}$ such that the cone generated by $\mathcal{S}$ is $\mathcal{K}$.*

Note that any polyhedral cone $\mathcal{K}$ must be closed.

**Theorem 22** (2.5 from Ch. 1 of [29]). *A cone $\mathcal{K}$ is polyhedral if and only if it is the intersection of a finite number of closed half-spaces, each containing the origin on its boundary.*

**Corollary 20.** *Polyhedral cones are closed under intersections.*

**Definition 12** (2.6 from Ch. 1 of [29]). *A cone $\mathcal{K}$ is **pointed**, provided that $\mathcal{K} \cap -\mathcal{K} = \{0\}$.*

**Definition 13** (2.6, 2.7 from Ch. 1 of [29]). *A cone is **solid**, provided that the interior of $\mathcal{K}$ is nonempty, or equivalently, $\mathcal{K} - \mathcal{K} = \mathbb{R}^n$ (sometimes referred to as **reproducing**).*

**Definition 14** (2.10 from Ch. 1 of [29]). *A cone is **proper**, if it is pointed, solid, and closed.*

The cones $\mathbb{R}^n$, $\mathbb{R}^n_+$, and $\{0\}$ are all polyhedral. The ice cream cone, defined as $\{\boldsymbol{x} \in \mathbb{R}^n | (\boldsymbol{x}_2^2 + \boldsymbol{x}_3^2 + \ldots + \boldsymbol{x}_n^2)^{1/2} \leq \boldsymbol{x}_1\}$ is not polyhedral. In three dimensions, polyhedral cones can be thought of as the cones with flat boundary surfaces, while those which are not polyhedral have 'curved' boundary surfaces.

The cone $\mathbb{R}^n$ is obviously not pointed, while the nonnegative orthant $\mathbb{R}^n_+$ is pointed. A cone $\mathcal{K}$ is pointed if and only if $\mathcal{K}$ contains no subspace $\mathcal{L} \subset \mathbb{R}^n$, apart from the subspace $\mathcal{L} = \{0\}$.

**Theorem 23.** *A cone is solid if and only if for every subspace $\mathcal{L} \subset \mathbb{R}^n$ with $\mathcal{L} \neq \mathbb{R}^n$, the subspace $\mathcal{L}$ does not contain the cone $\mathcal{K}$.*

Thus a cone $\mathcal{K}$ defined with respect to the vector $\boldsymbol{x} \in \mathbb{R}^n$ as $\{\, \boldsymbol{y} \,|\, \boldsymbol{y} = \alpha\boldsymbol{x}, \alpha \geq 0\}$ is not solid for $n > 1$, as such a cone would be contained in the one-dimensional subspace $\mathcal{L} \neq \mathbb{R}^n$ with the basis $\{\boldsymbol{x}\}$.

The nonnegative orthant $\mathbb{R}^n_+$ and the ice cream cone are both proper, while $\mathbb{R}^n$ is not (it fails to be pointed).

## D.2 Extremal vectors and extremal rays

**Definition 15** (generalizes 2.11 from Ch. 1 of [29]). *Consider a closed cone $\mathcal{K}$. A vector $x \in \mathcal{K}$ is an extremal vector of $\mathcal{K}$ provided that when it is expressed as a linear combination of two vectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{K}$ it follows that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ must be linear combinations of one another.*

By Def. 12, a closed cone $\mathcal{K}$ has extremal vectors if and only if it is pointed.

**Theorem 24** (generalizes 2.12 from Ch. 1 of [29]). *A closed, pointed cone is generated by its extremal vectors.*

Naturally, there exists a finite set of extremal vectors that generate a pointed polyhedral cone $\mathcal{K}$.

# Bibliography

[1] P. Bremaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* New York, NY, USA: Springer-Verlag, 1998.

[2] R. G. Gallager, *Discrete Stochastic Processes.* Boston, MA, USA: Kluwer Academic Publishers, 1996.

[3] A. L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization." *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, February 2004.

[4] C. Lindemann and A. Thümmler, "Performance analysis of the general packet radio service," *Computer Networks*, vol. 41, pp. 1–17, 2003.

[5] C. E. Riddalls, S. Bennett, and N. S. Tipi, "Modelling the dynamics of supply chains," *International Journal of Systems Science*, vol. 31, no. 8, pp. 969–976, 2000.

[6] J. P. Serre, *Linear Representations of Finite Groups.* New York, NY, USA: Springer-Verlag, 1977, translated by L. L. Scott.

[7] A. Terras, *Fourier Analysis on Finite Groups and Applications.* Cambridge, UK: Cambridge University Press, 2000.

[8] P. Diaconis and M. Shahshahani, "Generating a random permutation with random transpositions," *Probabilty Theory and Related Fields*, vol. 57, no. n, pp. 159–179, 1981.

[9] P. Diaconis, *Group Representations in Probability and Statistics*, ser. IMS Lecture Notes–Monograph Series. Hayward, CA: Institute of Mathematical Statistics, 1988.

[10] J. Huang, C. Guestrin, and L. Guibas, "Fourier theoretic probabilistic inference over permutations," *Journal of Machine Learning Research*, vol. 10, pp. 997–1070, 2009.

[11] G. James and A. Kerber, *The Representation Theory of the Symmetric Group*, ser. Encyclopedia of Mathematics and its Applications. Reading, MA, USA: Addison-Wesley, 1981.

[12] C. Asavathiratham, "The influence model: A tractable representation for the dynamics of networked Markov chains," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2000.

[13] C. Asavathiratham, S. Roy, B. Lesieutre, and G. Verghese, "The influence model," *IEEE Control Systems Magazine*, pp. 52–64, December 2001.

[14] S. Roy, "Moment-linear stochastic systems and their applications," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2003.

[15] S. Roy, G. Verghese, and B. Lesieutre, "Moment-linear stochastic systems," in *Informatics in Control, Automation and Robotics I*, J. Braz, H. Araújo, A. Vieira, and B. Encarnação, Eds. Cordrecht, The Netherlands: Springer, 2006, pp. 263–271.

[16] C. A. Gómez-Uribe and G. C. Verghese, "Mass fluctuation kinetics: Capturing stochastic effects in systems of chemical reactions through coupled mean-variance computations," *The Journal of Chemical Physics*, vol. 126, no. 2, p. 024109, 2007.

[17] B. Barzel, O. Biham, and R. Kupferman, "Analysis of the multiplane method for stochastic simulations of reaction networks with fluctuations," *Multiscale Modeling Simulation*, vol. 6, no. 3, pp. 963–982, 2007.

[18] C. S. Gillespie, "Moment-closure approximations for mass-action models," *IET Systems Biology*, vol. 3, no. 1, pp. 52–58, 2009.

[19] A. Singh and J. P. Hespanha, "Approximate moment dynamics for chemically reacting systems," *IEEE Transactions on Automatic Control*, 2010, to appear.

[20] G. Basile and G. Marro, "Controlled and conditioned invariant subspaces in linear system theory," *Journal of Optimization Theory and Applications*, vol. 3, no. 5, pp. 306–315, 1969.

[21] W. M. Wonham, *Linear Multivariable Control : A Geometric Approach.* Berlin, Germany: Springer-Verlag, 1974.

[22] B. Plateau and K. Atif, "Stochastic automata network for modeling parallel systems," *IEEE Transactions on Software Engineering*, vol. 17, pp. 1093–1108, 1991.

[23] B. Plateau and J. M. Fourneau, "A methodology for solving Markov models of parallel systems," *Journal of Parallel and Distributed Computing*, vol. 12, pp. 370–387, 1991.

[24] A. Benoit, P. Fernandes, B. Plateau, and W. J. Stewart, "On the benefits of using functional transitions and Kronecker algebra," *Performance Evaluation*, vol. 58, pp. 367–390, 2004.

[25] A. N. Langville and W. J. Stewart, "The Kronecker product and stochastic automata networks," *Journal of Computational and Applied Mathematics*, vol. 167, pp. 429–447, 2004.

[26] W. J. Stewart, K. Atif, and B. Plateau, "The numerical solution of stochastic automata networks," *European Journal of Operational Research*, vol. 86, pp. 503–525, 1995.

[27] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London*, vol. A222, pp. 309–368, 1922.

[28] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis.* Cambridge, UK: Cambridge University Press, 1991.

[29] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences.* Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1994.

[30] A. Pfeffer, “Sufficiency, separability and temporal probabilistic models,” in *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01).* San Francisco, CA: Morgan Kaufmann, 2001, pp. 421–42.

[31] C. Asavathiratham, “Linear algebra approach to separable Bayesian networks,” in *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06).* Arlington, Virginia: AUAI Press, 2006.

[32] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann, 1988.

[33] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber, “Automatic symbolic traffic scene analysis using belief networks,” in *In AAAI.* AAAI Press, 1994, pp. 966–972.

[34] A. E. Nicholson and J. M. Brady, “Dynamic belief networks for discrete monitoring,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, pp. 1593–1610, 1994.

[35] Z. Ghahramani, “Learning dynamic Bayesian networks,” in *Adaptive Processing of Sequences and Data Structures.* Springer-Verlag, 1998, pp. 168–197.

[36] W. J. Richoux and G. C. Verghese, “A generalized influence model for networked stochastic automata,” *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 41, 2011.

[37] D. Williams, *Probability with Martingales.* Cambridge, UK: Cambridge University Press, 1991.

[38] V. Chellaboina, S. Bhat, W. M. Haddad, and D. S. Bernstein, “Modeling and analysis of mass-action kinetics,” *IEEE Control Systems Magazine*, vol. 29, no. 4, pp. 60–78, 2009.

[39] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge, UK: Cambridge University Press, 1995.

[40] D. W. Stroock, *An Introduction to Markov Processes.* Berlin, Germany: Springer-Verlag, 2005.

[41] S. L. Lauritzen, *Graphical Models.* Oxford, UK: Oxford University Press, 1996.

[42] N. Vorob’ev, “Consistent families of measures and their extensions,” *Theory of Probability and its Applications*, vol. 7, no. 2, pp. 147–163, 1962.

[43] ——, “Markov measures and markov extensions,” *Theory of Probability and its Applications*, vol. 8, no. 4, pp. 420–429, 1963.

[44] H. G. Kellerer, “Matheoretische marginalprobleme,” *Mathematische Annalen*, vol. 153, pp. 168–198, 1964.

[45] D. P. Bertsekas, *Convex Analysis and Optimization.* Belmont, MA, USA: Athena Scientific, 2003.

[46] R. Kaas and J. M. Buhrman, “Mean, median, and mode in binomial distributions,” *Statistica Neerlandica*, vol. 34, no. 1, pp. 13–18, 1980.

[47] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Hoboken, NJ, USA: John Wiley & Sons, 2006.

[48] D. P. Bertsekas and J. N. Tsitskilis, *Introduction to Probability.* Belmont, MA, USA: Athena Scientific, 2008.

[49] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[50] G. Grimmett and D. Stirzaker, *Probability and Random Processes.* New York, NY, USA: Oxford University Press, 2001.

[51] L. Devroye, *Non-Uniform Random Variate Generation.* New York, NY, USA: Springer-Verlag, 1986.

[52] G. DeRise, "Some $n$-dimensional geometry," *International Journal of Mathematical Education in Science and Technology*, vol. 23, no. 3, pp. 371–379, 1992.

[53] J. Bochnak, M. Coste, and M. Roy, *Real Algebraic Geometry.* Berlin, Germany: Springer, 1998.

[54] G. Pistone, E. Riccomagno, and H. P. Wynn, *Algebraic Statistics: Computational Commutative Algebra in Statistics.* Boca Raton, FL, USA: Chapman and Hall/CRC, 2001.

[55] L. Pachter and B. Sturmfels, *Algebraic Statistics for Computational Biology.* New York, NY, USA: Cambridge University Press, 2005.

[56] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition." in *Proceedings of the IEEE*, 1989, pp. 257–286.

[57] A. Jammalamadaka, "Aspects of inference for the influence model and related graphical models." Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2004.

[58] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control.* Oxford, UK: Oxford University Press, 1992.

[59] L. Zager and G. Verghese, "Epidemic thresholds for infections in uncertain networks," *Complexity*, vol. 14, no. 4, pp. 12–25, 2009.

[60] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Learning human interactions with the influence model," *MIT Media Laboratory Technical Note*, no. 539, 2001.

[61] W. Dong and A. Pentland, "Modeling influence between experts," in *Lecture Notes on Artificial Intelligence, Special Volume on Human Computing.* Springer-Verlag, 2007, vol. 4451, pp. 170–189.

[62] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *ICMI '07: Proceedings of the 9th International Conference on Multimodal Interfaces.* New York, NY, USA: ACM, 2007, pp. 271–278.

[63] S. Singh, H. Tu, W. Donat, K. Pattipati, and P. Willett, "Anomaly detection via feature-aided tracking and hidden Markov models," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 39, no. 1, pp. 144 –159, 2009.

[64] U. N. Bhat and G. K. Miller, *Elements of Applied Stochastic Processes.* Hoboken, NJ, USA: John Wiley & Sons, 2002.

[65] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial.* Hanover, MA, USA: Now Publishers Inc., 2004.

[66] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory.* Chichester, UK: John Wiley & Sons, 1999.

[67] I. Csiszár, "Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problmes," *The Annals of Statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.

[68] ——, "Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten," *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, vol. 8, pp. 85–108, 1963.

[69] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[70] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[71] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions*, no. 1, pp. 204–237, 1984.

[72] I. Csiszár, "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling," *The Annals of Statistics*, vol. 17, no. 3, pp. 1409–1413, 1989.

[73] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Sociated: Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[74] G. H. Golub and C. F. V. Loan, *Matrix Computations.* Baltimore, MD, USA: The Johns Hopkins University Press, 1996.

[75] T. Ottmann, S. Schuierer, and S. Soundaralakshmi, "Enumerating extreme points in higher dimensions," in *STACS 95*, ser. Lecture Notes in Computer Science, E. Mayr and C. Puech, Eds. Berlin, Germany: Springer, 1995, vol. 900, pp. 562–570.

[76] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 50, no. 2, pp. 157–224, 1988.

[77] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[78] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of the American Mathematical Society*, vol. 73, pp. 360–363, 1967.

[79] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds," *Pacific Journal of Mathematics*, vol. 27, no. 2, pp. 211–227, 1968.

[80] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[81] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[82] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260 – 269, 1967.

[83] G. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268 – 278, 1973.

[84] R. E. Neopolitan, *Learning Bayesian Networks.* Upper Saddle River, NJ, USA: Prentice Hall, 2004.

[85] D. Geiger and J. Pearl, "Logical and algorithmic properties of conditional independence and graphical models," *The Annals of Statistics*, vol. 21, no. 4, pp. 2001–2021, 1993.

[86] A. P. Dawid, "Conditional independence in statistical theory," *Journal of the Royal Statistical Society, Series B*, vol. 41, no. 1, pp. 1–31, 1979.

[87] F. V. Jensen, "Junction trees and decomposable hypergraphs," *Judex Datasystemer, Aalborg, Denmark*, 1988.

[88] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in boltzmann machines," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations.* Cambridge, MA, USA: MIT Press, 1986, pp. 282–317.

[89] R. Kindermann and J. L. Snell, *Markov random fields and their applications.* Providence, RI, USA: American Mathematical Society, 2000.

[90] M. I. Jordan, "An introduction to variational methods for graphical models," in *Machine Learning.* MIT Press, 1999, pp. 183–233.

[91] T. S. Jaakkola and M. I. Jordan, "Variational methods for inference and estimation in graphical models," *Uncertainty and artificial intelligence: Proceedings of the twelfth conference*, 1996.

[92] ——, "Improving the mean field approximation via the use of mixture distributions," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA, USA: MIT Press, 2006, pp. 163–173.

[93] G. Parisi, *Statistical Field Theory.* Redwood City, CA, USA: Addison-Wesley, 1988.

[94] B. D. O. Anderson, M. Deistler, L. Farina, and L. Benvenuti, "Nonnegative realization of a linear system with nonnegative impulse response," *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications*, vol. 43, no. 2, pp. 134–142, 1996.

[95] B. D. O. Anderson, "The realization problem for hidden Markov models," *Mathematics of Control, Signals, and Systems*, vol. 12, no. 1, pp. 80–120, 1999.

[96] M. Vidyasagar, "The realization problem for hidden Markov models: A survey and some new results," *Mathematics of Control, Signals, and Systems*, to appear.

[97] E. Seneta, *Non-negative Matrices and Markov chains.* New York, NY, USA: Springer, 2006.

[98] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms.* Cambridge, MA, USA: MIT Presss, 2009.

[99] S. Roman, *Coding and Information Theory.* New York, NY, USA: Springer, 1992.