UNIVERSITY OF CALGARY

The Application of Lie Theory to Markov Processes

Computation of the Maximum Likelihood Estimator of the Generator of Continuous Time Homogeneous Markov Processes from Stopped Random Variables

by

Aaron Geoffrey Sheldon

A PROJECT

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF ”Masters of Science”

GRADUATE PROGRAM IN ”Statistics”

CALGARY, ALBERTA

August, 2016

Abstract

Individually the probability theory of stochastic processes and Lie theory have been highly productive fields of investigation for more than a century; yet they remain ripe for cross pollination. In particular, the application of algebraic and analytic results from Lie theory can yield novel computational methods for the estimation of generators of continuous time homogeneous Markov processes on finite state spaces. In this project we derive the minimal Lie algebra that contains the generators of continuous time homogeneous Markov processes on finite state spaces, and then using the guarantees of algebraic and analytic closure construct a Newton-Raphson algorithm for maximum likelihood estimation of the generator of a continuous time homogeneous Markov process on a finite state space from stopped random variables using Páde approximations for Taylor series expressions of the first and second order derivatives of the exponential map.

Acknowledgements

First and foremost is my love and gratitude to my loving family; who withstood the long stretches of my physical presence, but mental absence while I developed, and explored the mathematical geography I describe in this report.

Second, I am deeply appreciative of my supervisor who has had the extraordinary patience to survive the many years of rumination it took for me to develop the meager ideas expressed in this report.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Algorithms

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
U of C	University of Calgary
$A, B, C, \ldots$	Matrices, except for I which we take as identity
$a, b, c, \ldots$	Constants
$A^\dagger, a^\dagger$	Matrix and vector transpose; *not the conjugate transpose!*
$\vec{a}, \vec{b}, \vec{c}, \ldots$	Vectors
$\hat{a}, \hat{b}, \hat{c}, \ldots$	Unit vectors
$\vec{\mathbb{1}}$	Row sum vector
$\hat{\mathbb{1}}$	Normalized row sum vector
$St(\hat{\mathbb{1}})$	Stochastic Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}(\hat{\mathbb{1}})$	Stochastic Lie algebra with respect to $\hat{\mathbb{1}}$
$St^+(\hat{\mathbb{1}})$	Stochastic contraction Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}^+(\hat{\mathbb{1}})$	Stochastic contraction Lie algebra with respect to $\hat{\mathbb{1}}$
$St^\dagger(\hat{\mathbb{1}})$	Dual stochastic Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}^\dagger(\hat{\mathbb{1}})$	Dual stochastic Lie algebra with respect to $\hat{\mathbb{1}}$
$St^{+\dagger}(\hat{\mathbb{1}})$	Dual stochastic contraction Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}^{+\dagger}(\hat{\mathbb{1}})$	Dual stochastic contraction Lie algebra with respect to $\hat{\mathbb{1}}$
$St(\hat{\mathbb{1}}, \hat{\mathbb{1}})$	Doubly stochastic Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}})$	Doubly stochastic Lie algebra with respect to $\hat{\mathbb{1}}$
$St^+(\hat{\mathbb{1}}, \hat{\mathbb{1}})$	Doubly stochastic contraction Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}^+(\hat{\mathbb{1}}, \hat{\mathbb{1}})$	Doubly stochastic contraction Lie algebra with respect to $\hat{\mathbb{1}}$
$\hat{a} \otimes \hat{b}$	Kronecker product of unit vectors $\hat{a}$ and $\hat{b}$
$\langle \hat{a}, \hat{b} \rangle$	Inner product of unit vectors $\hat{a}$ and $\hat{b}$
$\mathscr{F}_t$	Filtration of sigma algebras with respect to continuous parameter t

X_t	Stochastic process on the filtration $\mathscr{F}_t$
$\mathbb{P}[\Sigma]$	Probability measure of a measurable set Σ
$\mathbb{P}[\Sigma \parallel \mathscr{F}_t]$	Conditional probability with respect to $\mathscr{F}_t$
$\mathbb{I}[\Sigma]$	Indicator function of a measurable set Σ
$\mathbb{E}[X_t]$	Expectation of X_t
$\mathbb{E}[X_t \parallel \mathscr{F}_s]$	Expectation conditioned on $\mathscr{F}_s$, given $s \le t$
$\mathbb{V}ar[X_t]$	Variance of X_t
$\mathbb{V}ar[X_t \parallel \mathscr{F}_s]$	Variance conditioned on $\mathscr{F}_s$, given $s \le t$
$\mathbb{C}ov[X_t, Y_t]$	Covariance of X_t and Y_t
$\mathbb{C}ov[X_t, Y_t \parallel \mathscr{F}_s]$	Covariance conditioned on $\mathscr{F}_s$, given $s \le t$
$\mathbb{C}or[X_t, Y_t]$	Correlation of X_t and Y_t
$\mathbb{C}or[X_t, Y_t \parallel \mathscr{F}_s]$	Correlation conditioned on $\mathscr{F}_s$, given $s \le t$
$\mathrm{Ad}_X A$	Lie group Adjoint operator XAX^{-1}
$\mathrm{ad}_X A$	Lie algebra adjoint operator $[X, A]$
$[A, B]$	Commutator bracket $AB - BA$ of matrices A and B
δ_{ij}	Dirac delta operator
$\mathbb{N}$	Natural numbers
$\mathbb{Z}$	Integers numbers
$\mathbb{Q}$	Rational numbers
$\mathbb{R}$	Real numbers
$\mathbb{C}$	Complex numbers
$\mathbb{R}e(x)$	Real part of x
$\mathbb{I}m(x)$	Imaginary part of x
i	Imaginary unit $\sqrt{-1}$
$a \equiv b \mod c$	Modular equivalence $a = nc + b$ with $n \in \mathbb{Z}$

Chapter 1

Introduction

1.1 Motivation

Markov processes are a central subject of study in probability theory, and are a rich source of distributions for parameter estimation in statistics[5, 17, 18]. They have applications in diverse disciplines ranging through the physical and life sciences, including operations research, chemical process engineering, queuing theory, communications theory, natural language processing, finance, and machine learning. Under mild assumptions and constraints they offer tractable, and even closed form models; that can be reasoned about using physical heuristic analogies, and intuitive phenomenological interpretations. To varying degrees of rigor, methods for both simulation, and parameter estimation have been developed for many types of observed random and pseudo-random processes such as the syntax of sentences, disease states, cancer survival, epidemiology, and demographics. To apply Markov process a number of simplifying assumptions are made, including discretization of time and state spaces, homogeneity of the process, and restrictions of the allowed transitions. The simplifications have resulted in models such as phase type distributions, branching processes, birth-death processes, and hidden Markov models.

State of the art computational methods are focused on maximum likelihood parameter estimation by expectation maximization of hidden Markov Models; which assumes a finite state space obscured by random noise, with discrete homogeneous time steps, and all times of transitions being observed. The discretization of time allows for the time evolution of transition probabilities to be explicitly parameterized through matrix multiplication. The discrete time construction of hidden Markov models is successfully exploited by the Baum-Welch, Viterbi, and forward-backward algorithms to estimate parameters.

In contrast continuous time homogeneous Markov processes on a finite state space are more naturally parameterized through the generator, because the time evolution is represented through matrix exponentiation. Unfortunately parameterization of the generators of Markov processes, in more than four states, does not in general yield tractable closed from transition probabilities. This is because any explicit formulation of the transition probabilities from the generator would require solving the characteristic polynomial of the generator, which is not generally possible in dimensions greater than four. Yet computational approximations of the matrix exponential have been well developed, with methods to compute the gradient receiving recent attention. The focus of this recent research has been on computing the condition number of numerical problems, as a measure of convergence and stability of the numerical solutions[2].

Given a computational method to calculate the matrix exponential, and it's gradient, and Hessian, an application of the chain rule then allows for the computation of maximum likelihood estimates of any differentiable parameterization of the generators of a continuous time homogeneous Markov process on a finite state space. Of particular interest in such a method are stopped statistics, like the first hitting times of transitions from a fixed source state to a fixed target state. As such this work extends the current computational methods to include the Hessian of the matrix exponential; and further develops an alternate direct computation of the gradient of the matrix exponential.

Throughout this work we will attempt to conform to a simplified version of Lamport's guide to structuring, and presenting proofs[13].

1.2 Overview

From the perspective of Lie theory, classical parameter estimation of Markov processes has been a manifold first approach; starting with an explicit construction of an extrinsic smooth coordinate chart (parameterization) on a neighborhood of the sub-manifold to which the generators belong, and only then looking for computational simplifications and solutions. As hinted to in the previous section, we will proceed with an algebra first approach; developing the intrinsic algebraic structure of the generators of the Lie algebra, and then exploiting the implicit function theorem to carry out computations in specific parameterizations.

The second chapter establishes the algebraic and analytic closure properties necessary for chapter three, and establishes notation used throughout all the following chapters. Chapter two has a secondary role to help develop the physical intuition for the stochastic Lie group needed to work through the derivatives and approximations of chapter three. However the work on computing the logarithms of permutations can be set aside, as the material was developed to gain a fuller understanding of the structure of the stochastic Lie algebra.

The third chapter reviews the definition for the gradient of exponential map, and continues on to derives an analytic form for the Hessian of the exponential map. A Padé approximation of the gradient of the exponential map is then developed, followed by an algorithm that to calculate the Hessian of the exponential map. The Hessian algorithm combines the Padé approximation of the gradient with a novel recursive calculation of a bilinear non-commutative perturbation to the Hessian.

The fourth chapter seeks to illustrate the material developed in the preceding two chapters. First the cumulative distribution of the stopped statistic of first hitting times is recast in the stochastic Lie algebra developed in chapter two. These results are then put to use in developing a model for an aging process; which is a finite state reversible birth-death process. The chapter concludes by laying out the algorithms to calculate the gradient and Hessian of the log-likelihood of the aging process, and hence a Newton-Raphson method for maximization.

Finally, the fifth chapter concludes with summarizing remarks and a discussion of the direction for further investigation.

Chapter 2

The Lie Algebra of Markov Process Generators

2.1 Stochastic Matrices

2.1.1 Preliminaries

The classical Lie algebras of physics, like the infinitesimal symmetries of the special unitary algebra $\mathfrak{su}(n)$, are defined with respect to invariants of a Banach algebra, such as the matrix invariants of the determinant, trace, or norm. In contrast stochastic matrices are always characterized with respect to a specific unit vector, which we will denote by $\hat{\mathbb{1}}$, for reasons that will become clear later. In the next two sections we provide an explicit construction and characterization of the Lie algebra of stochastic matrices, building on the original the work of Johnson [12], and distilling the recent works of Sumner et al. [19, 8], Mourad [16], and Chruściński et al. [7].

The common approach to stochastic matrices begins with the restriction that the matrices have non-negative entries with respect to the standard orthonormal basis for the vector space on which it acts; namely $\left\langle \hat{e}_i, A\hat{e}_j \right\rangle \geq 0$ for all i, j. In addition to allowing for singular matrices, this poses an immediate obstacle to the necessary closure with respect to matrix inversion required for matrix groups; as the inverse of a stochastic matrix need not have non-negative entries with respect to the standard orthonormal basis.

For the moment we will set aside the restriction that the entries be non-negative, and instead begin with a generalization of fixed row sums to abstract linear operators. We will show that this generalization is preserved by operator inversion, and then develop an orthonormal basis from which matrix representations of the abstract linear operators can be constructed with non-negative entries. In essence tackling the problem from the reverse direction, starting with the an abstract generalization of the idea of fixed row sums, and then specifying matrices with non-negative entries with respect to a constructed orthonormal basis.

Definition 1. A bounded linear operator A on a finite dimensional Hilbert space is stochastic with respect to the unit vector $\hat{\mathbb{1}}$ if $A\hat{\mathbb{1}} = \hat{\mathbb{1}}$

Note that this definition does not stipulate any conditions on non-singularity, and thus includes, as representations of the linear operators, all the matrices in the convex polytope of stochastic matrices. For an n dimensional vector space the vector $\vec{\mathbb{1}} = \sqrt{n}\hat{\mathbb{1}}$ acts as the row sum operator on bounded linear operators stochastic with respect to $\hat{\mathbb{1}}$. We will make this claim more precise after we dispense with a few more foundational definitions.

Definition 2. Let $St(\hat{\mathbb{1}})$ denote the stochastic Lie group of invertible bounded linear operators stochastic with respect to $\hat{\mathbb{1}}$.

It is tempting to view the name stochastic Lie group as a bait and switch, or at least an abuse of the terminology, given we have removed the usual convex polytope of stochastic matrices and replaced it with a group of invertible bounded linear operators with a common eigenvector $\hat{\mathbb{1}}$. Previous authors have denoted the convex polytope of stochastic matrices as the stochastic semi-group, and the group of invertible matrices as the pseudo-stochastic Lie group. One could even consider incorporating Markov into the name, in reference to the fact that the transition matrices of a continuous time homogeneous Markov process on a finite state space are by definition invertible, and have common eigenvector $\hat{\mathbb{1}}$. However the suffix of Lie group in the name connotes both sufficient additional restrictions to make the name distinct, and still allows for an indication of a relationship with the original concept. Of course, this definition immediately necessitates a proof of the claim embedded in the definition.

Lemma 1. *$St(\hat{\mathbb{1}})$ is a Lie group*

Proof: We proceed by working mechanistically through the Lie group axioms [9].

1. The identity element I is in $St(\hat{\mathbb{1}})$. Clearly I is invertible and $I\hat{\mathbb{1}} = \hat{\mathbb{1}}$.

2. If $A, B \in St(\hat{\mathbb{1}})$ then $AB \in St(\hat{\mathbb{1}})$. This follows from the computation $AB\hat{\mathbb{1}} = A\hat{\mathbb{1}} = \hat{\mathbb{1}}$.

3. If $A \in St(\hat{\mathbb{1}})$ then $A^{-1} \in St(\hat{\mathbb{1}})$. Recognize that $A\hat{\mathbb{1}} = \hat{\mathbb{1}}$ implies $\hat{\mathbb{1}} = A^{-1}A\hat{\mathbb{1}} = A^{-1}\hat{\mathbb{1}}$.

4. Associativity follows from $St(\hat{\mathbb{1}})$ being a subgroup of $GL(n)$.

5. Finally, we need to prove that is $St(\hat{\mathbb{1}})$ closed within $GL(n)$. Consider a sequence $A_n \in St(\hat{\mathbb{1}})$ that converges to A, then $\hat{\mathbb{1}} = A_n\hat{\mathbb{1}} \to A\hat{\mathbb{1}}$. Now if A is invertible then we are done, and if A is not invertible then $A \notin GL(n)$, again satisfying closure within $GL(n)$. □

That $St(\hat{\mathbb{1}})$ is a proper Lie group implies that it must be infinitesimal generated by elements of a Lie algebra.

Definition 3. Let $\mathfrak{st}(\hat{\mathbb{1}})$ denote the stochastic Lie algebra of $St(\hat{\mathbb{1}})$

By infinitesimally generated we mean that every element of $St(\hat{\mathbb{1}})$ is the exponential map of at least one element in $\mathfrak{st}(\hat{\mathbb{1}})$. We can fully characterize this algebra as the set of bounded linear operators such that $\hat{\mathbb{1}}$ is in the kernel of each operator.

Lemma 2. *The algebra $\mathfrak{st}(\hat{\mathbb{1}})$ is exactly the set of all bounded linear operators with $\hat{\mathbb{1}}$ in their kernel.*

Proof: Working through the forward and backward inclusions we have

1. Suppose $A\hat{\mathbb{1}} = 0$ then from the definition of the exponential map we have:

$$\begin{aligned}\exp(A)\,\hat{\mathbb{1}} &= \sum_{n=0}^{\infty} \frac{1}{n!} A^n \hat{\mathbb{1}} \\ &= \hat{\mathbb{1}} + \sum_{n=1}^{\infty} \frac{1}{n!} 0 \\ &= \hat{\mathbb{1}}\end{aligned}$$

Thus $\exp(A) \in St(\hat{\mathbb{1}})$ implying that $A \in \mathfrak{st}(\hat{\mathbb{1}})$

2. Now proceeding with the reverse assumption, that $A \in \mathfrak{st}(\hat{\mathbb{1}})$. For all $t \in \mathbb{R}$ we have $\exp(tA)\,\hat{\mathbb{1}} = \hat{\mathbb{1}}$. Differentiation with respect to t and evaluation at $t = 0$ yields

$$\begin{aligned} 0 &= \frac{d}{dt}\hat{\mathbb{1}}\bigg|_{t=0} \\ &= \frac{d}{dt}\exp(tA)\,\hat{\mathbb{1}}\bigg|_{t=0} \\ &= \exp(tA)\,A\hat{\mathbb{1}}\big|_{t=0} \\ &= A\hat{\mathbb{1}} \end{aligned}$$

□

The following chapter will hinge on taking the derivatives of smooth parameterizations $X : \mathbb{R}^k \mapsto \mathfrak{st}(\hat{\mathbb{1}})$. The principle role of this chapter is to assure ourselves that we will not differentiate ourselves out of $\mathfrak{st}(\hat{\mathbb{1}}$. The next corollary nicely provides just such an assurance:

Corollary 1. *The tangent space* $T\mathfrak{st}(\hat{\mathbb{1}}) = \mathfrak{st}(\hat{\mathbb{1}})$

Proof: The proof is complementary to the preceding lemma and moves through each direction of inclusion:

1. To show that $\mathfrak{st}(\hat{\mathbb{1}}) \subseteq T\mathfrak{st}(\hat{\mathbb{1}})$ consider $X(x) = xX_0$ where x is a scalar parameter and $X_0 \in \mathfrak{st}(\hat{\mathbb{1}})$
2. Clearly $X(x) \in \mathfrak{st}(\hat{\mathbb{1}})$
3. Furthermore the tangent $\frac{\partial}{\partial x}X(x) = X_0 \in \mathfrak{st}(\hat{\mathbb{1}})$
4. Now to show that $T\mathfrak{st}(\hat{\mathbb{1}}) \subseteq \mathfrak{st}(\hat{\mathbb{1}})$ we start with an arbitrary smooth parameterization $X(x) : \mathbb{R}^k \mapsto \mathfrak{st}(\hat{\mathbb{1}})$
5. Using the same trick of differentiation we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial x}0 \\ &= \frac{\partial}{\partial x}\left(X(x)\,\hat{\mathbb{1}}\right) \\ &= \left(\frac{\partial}{\partial x}X(x)\right)\hat{\mathbb{1}} \end{aligned}$$

□

Clearly the tangent space to a normed vector space is the normed vector space. After all one can just choose a fixed basis and then differentiate the individual smoothly parameterized in projections. However, the proof of the preceding corollary was constructed to explicitly connect algebraic closure and differentiation, with the norm implicitly used in the differentiation. The proof further illustrates the intuition that any increase in a particular matrix element has to be compensated for by an equal decrease in some other matrix elements.

2.1.2 Canonical Generators

Over an n dimensional vector space, the condition on a matrix A that $A\hat{\mathbb{1}} = 0$ places n constraints on the n^2 dimensions of A. This leaves $n^2 - n$ free dimensions on $\mathfrak{st}(\hat{\mathbb{1}})$, when considered as a vector space. This hints that we can construct a generator of $\mathfrak{st}(\hat{\mathbb{1}})$ from order pairs of basis elements $\hat{e}_i$ for the vector space of $\hat{\mathbb{1}}$. To see how this is done we first construct a useful basis for the vector space to which $\hat{\mathbb{1}}$ is a member.

Lemma 3. *There exists an orthonormal basis $\hat{e}_i$ such that $\left\langle \hat{e}_i, \hat{\mathbb{1}} \right\rangle = \frac{1}{\sqrt{n}}$ for all i*

Proof: While a basis with the stipulated properties can be constructed through the Gram-Schmidt process, the proof of the existence proceeds by induction.

1. For $n = 1$ the desired basis is precisely the trivial set $\left\{\hat{\mathbb{1}}\right\}$ which satisfies the condition that $\left\langle \hat{\mathbb{1}}, \hat{\mathbb{1}} \right\rangle = 1$.

2. Assume the claim is true for n. For $n+1$ pick a unit vector $\hat{e}_\perp$ that is orthogonal to $\hat{\mathbb{1}}$ and construct the unit vector $\hat{e}_{n+1} = \frac{1}{\sqrt{n+1}}\hat{\mathbb{1}} + \sqrt{\frac{n}{n+1}}\hat{e}_\perp$. Clearly $\hat{e}_{n+1}$ satisfies the condition $\left\langle \hat{e}_{n+1}, \hat{\mathbb{1}} \right\rangle = \frac{1}{\sqrt{n+1}}$.

3. To use the the induction assumption we construct a new row sum unit vector $\hat{\mathbb{1}}_n = \sqrt{\frac{n+1}{n}}\hat{\mathbb{1}} - \frac{1}{\sqrt{n}}\hat{e}_{n+1}$ in one dimension lower by projecting onto the subspace orthogonal to $\hat{e}_{n+1}$.

4. By the induction assumption there exists a basis $\hat{e}_i$ with $i \leq n$, such that $\left\langle \hat{e}_i, \hat{\mathbb{1}}_n \right\rangle = \frac{1}{\sqrt{n}}$.

5. Because $\hat{e}_i$ with $i \leq n$ was constructed in the space orthogonal to $\hat{e}_{n+1}$ it follows that $\left\langle \hat{e}_i, \hat{e}_j \right\rangle = \delta_{ij}$ for all $i, j \leq n+1$.

6. Then using the definitions of $\hat{e}_{n+1}$ and $\hat{\mathbb{1}}_n$ we can calculate the inner product $\left\langle \hat{e}_i, \hat{\mathbb{1}}_n \right\rangle$ for $i \leq n$

$$\begin{aligned}\frac{1}{\sqrt{n}} &= \left\langle \hat{e}_i, \hat{\mathbb{1}}_n \right\rangle \\ &= \sqrt{\frac{n+1}{n}} \left\langle \hat{e}_j, \hat{\mathbb{1}} \right\rangle - \frac{1}{\sqrt{n}} \left\langle \hat{e}_i, \hat{e}_{n+1} \right\rangle \\ &= \sqrt{\frac{n+1}{n}} \left\langle \hat{e}_j, \hat{\mathbb{1}} \right\rangle\end{aligned}$$

   Inverting the fraction in the equality yields $\left\langle \hat{e}_i, \hat{\mathbb{1}} \right\rangle = \frac{1}{\sqrt{n+1}}$ for all $i \leq n+1$. □

As a direct result of the construction of the basis vectors $\hat{e}_i$ we see that $\vec{\mathbb{1}} = \sum_{i=1}^{n} \hat{e}_i$. Thus $\vec{\mathbb{1}}$ can be interpreted as the row sum vector in basis $\hat{e}_i$.

The constructed basis leads naturally to considering the minimal non-trivial matrices $C_{ij} = \hat{e}_i \otimes \left(\hat{e}_j - \hat{e}_i\right)$, with state transitions illustrated in figure 2.1, as holding significance in the structure of $\mathfrak{st}(\hat{\mathbb{1}})$. In fact this will be the central result of this chapter: that the algebraic closure of the matrices C_{ij} is the stochastic Lie algebra $\mathfrak{st}(\hat{\mathbb{1}})$. To establish this result we need a preliminary result that proves the commutators $\left[C_{ij}, C_{kl}\right]$ are linear combinations of matrices C_{ij}.

Lemma 4.

$$C_{ij}C_{kl} = \begin{cases} -C_{il} & i = k, \\ C_{il} - C_{ij} & j = k, \\ 0 & \textit{otherwise}. \end{cases}$$

Proof: We proceed in two steps; calculating the terms of the products, then simplifying the cases, always assuming $i \neq j$ and $k \neq l$.

1. Term wise computation of the Kronecker products yields

$$
\begin{aligned}
C_{ij}C_{kl} &= \hat{e}_i \otimes \left(\hat{e}_j - \hat{e}_i\right) \hat{e}_k \otimes \left(\hat{e}_l - \hat{e}_k\right) \\
&= \hat{e}_i \otimes \hat{e}_j \hat{e}_k \otimes \hat{e}_l + \hat{e}_i \otimes \hat{e}_i \hat{e}_k \otimes \hat{e}_k - \hat{e}_i \otimes \hat{e}_j \hat{e}_k \otimes \hat{e}_k - \hat{e}_i \otimes \hat{e}_i \hat{e}_k \otimes \hat{e}_l \\
&= \delta_{jk} \hat{e}_i \otimes \hat{e}_l + \delta_{ik} \hat{e}_i \otimes \hat{e}_k - \delta_{jk} \hat{e}_i \otimes \hat{e}_k - \delta_{ik} \hat{e}_i \otimes \hat{e}_l \\
&= \left(\delta_{jk} - \delta_{ik}\right) \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_k\right)
\end{aligned}
$$

2. We work through each case of $\delta_{jk} - \delta_{ik}$, starting with the case $i = k$

$$
\begin{aligned}
C_{ij}C_{il} &= \left(\delta_{jk} - \delta_{ii}\right) \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_i\right) \\
&= -\hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_i\right) \\
&= -C_{il}
\end{aligned}
$$

3. When $j = k$ we have

$$
\begin{aligned}
C_{ij}C_{jl} &= \left(\delta_{jj} - \delta_{ij}\right) \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_j\right) \\
&= \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_j\right) \\
&= \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_i + \hat{e}_i - \hat{e}_j\right) \\
&= C_{il} - C_{ij}
\end{aligned}
$$

4. Finally when none of the previous conditions apply

$$
\begin{aligned}
C_{ij}C_{kl} &= \left(\delta_{jk} - \delta_{ik}\right) \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_k\right) \\
&= 0 \cdot \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_k\right) \\
&= 0
\end{aligned}
\qquad \square
$$

While this result is sufficient to accomplish the central result, it is worth carrying through with the computation of the structure constants of the generators.

Corollary 2.

$$[C_{ij}, C_{kl}] = \begin{cases} C_{ij} - C_{il} & i = k, \\ C_{kj} - C_{ki} & i = l, \\ C_{il} - C_{ij} & j = k, \\ 0 & \textit{otherwise}. \end{cases}$$

Proof: As in the previous lemma we work case wise through the equalities.

1. Starting with $i = k$

$$\begin{aligned} [C_{ij}, C_{il}] &= C_{ij}C_{il} - C_{il}C_{ij} \\ &= C_{ij} - C_{il} \end{aligned}$$

2. For $i = l$

$$\begin{aligned} [C_{ij}, C_{ki}] &= C_{ij}C_{ki} - C_{ki}C_{ij} \\ &= C_{kj} - C_{ki} \end{aligned}$$

3. For $j = k$

$$\begin{aligned} [C_{ij}, C_{jl}] &= C_{ij}C_{jl} - C_{jl}C_{ij} \\ &= C_{il} - C_{ij} \end{aligned}$$

4. When none of the conditions apply

$$\begin{aligned} [C_{ij}, C_{kl}] &= C_{ij}C_{kl} - C_{kl}C_{ij} \\ &= 0 \end{aligned}$$

□

We can now proceed with the central result that motivates this chapter.

Theorem 1. *The canonical generators of* $\mathfrak{st}(\hat{\mathbb{1}})$ *are* C_{ij}

Proof: The previous lemma has established that the products, and thus the commutators, of C_{ij} are linear in C_{ij}. We then have to prove that the smallest algebra that contains C_{ij} is $\mathfrak{st}(\hat{\mathbb{1}})$. As thus, it is sufficient to prove that matrices C_{ij} from a basis for $\mathfrak{st}(\hat{\mathbb{1}})$. This is because a necessary condition for an algebra to contain the matrices C_{ij} is that it must contain all sums of the matrices C_{ij}. If one could sum their way out of the algebra then it would not be an algebra.

1. That $\mathfrak{st}(\hat{\mathbb{1}})$ is an $n^2 - n$ dimensional vector space should be clear from the previous discussion. A full formal proof of this claim is found through induction on the dimension n.

2. The matrices C_{ij} are in $\mathfrak{st}(\hat{\mathbb{1}})$. From the definition of the canonical generators

$$\begin{aligned} C_{ij}\hat{\mathbb{1}} &= \hat{e}_i \otimes \left(\hat{e}_j - \hat{e}_i\right)\hat{\mathbb{1}} \\ &= \hat{e}_i\left(\langle \hat{e}_j, \hat{\mathbb{1}}\rangle - \langle \hat{e}_i, \hat{\mathbb{1}}\rangle\right) \\ &= \hat{e}_i\left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n}}\right) \\ &= 0 \end{aligned}$$

3. C_{ij} is a set of $n^2 - n$ linear independent matrices and so must form a basis for all of $\mathfrak{st}(\hat{\mathbb{1}})$. That there are only $n^2 - n$ matrices is clear from the fact that $C_{ii} = 0$. While the formal proof of linear independence is again found through induction on the dimension n. □

The previous theorem serves as the definition of a set of canonical generators of $\mathfrak{st}(\hat{\mathbb{1}})$. It is important to note that neither the basis $\hat{e}_i$ nor the canonical generators C_{ij} are unique. They are defined only up to rotations orthogonal to the vector $\hat{\mathbb{1}}$. Regardless of the choice of basis $\hat{e}_i$ Jacobi's formula implies that if $G = \sum_{i,j} \alpha_{ij} C_{ij}$ then $\det G = \exp\left(\sum_{i,j} \alpha_{ij}\right)$.

2.1.3 Vertex Logarithms

Geometrically fixing a basis $\hat{e}_i$ defines the convex polytope of all matrices, M, stochastic with respect to $\hat{\mathbb{1}}$ that have nonnegative entries with respect to $\hat{e}_i$; also designated the convex polytope of singly stochastic matrices. The vertexes of the convex polytope of nonnegative stochastic matrices with respect to $\hat{e}_i$ can be formulated as linear combinations of the generators C_{ij} of $\mathfrak{st}(\hat{\mathbb{1}})$.

Definition 4. For a function $j(i):\{1,\ldots,n\}\mapsto S\subseteq\{1,\ldots,n\}$ the vertex matrix is given by $V_{j(\cdot)} = I+\sum_{i=1}^{n} C_{ij(i)}$

Every vertex matrix is uniquely dual to a matrix in $\mathfrak{st}(\hat{\mathbb{1}})$ through the relationship

$$\begin{aligned} V_{j(\cdot)} - I &= \sum_{i=1}^{n} \hat{e}_i \otimes \hat{e}_{j(i)} - \hat{e}_i \otimes \hat{e}_i \\ &= C_{j(\cdot)} \end{aligned}$$

It is important to note that in general the vertex duals $C_{j(\cdot)}$ in $\mathfrak{st}(\hat{\mathbb{1}})$ are not the generators, nor logarithms, of the vertexes $V_{j(\cdot)}$. In fact, if there exists $i_1 \neq i_2$ such that $j(i_i) = j(i_2)$ then $V_{j(\cdot)}$ does not have a generator, or logarithm. However, in this situation we will show that $V_{j(\cdot)}$ is a limit point of the boundary of $\mathfrak{st}(\hat{\mathbb{1}})$.

For now we need to establish the claim implicit in the definition of the vertex matrices.

Lemma 5. *The convex hull of the n^n vertex matrices $V_{j(\cdot)}$ is the convex polytope of nonnegative stochastic matrices with respect to $\hat{e}_i$.*

Proof: The proof proceeds by showing each direction of inclusion.

1. Starting with the forward inclusion, clearly every $V_{j(\cdot)}$ belongs to the convex polytope of singly stochastic matrices.

2. It follows that any convex sum of the vertexes $V_{j(\cdot)}$ also belongs to the convex polytope of singly stochastic matrices.

3. In the reverse inclusion, we work analogously to the Birkhoff-von Neumann theorem; where starting with an M in the convex polytope of singly stochastic matrices we eliminate nonzero entries from M with convex sums involving $V_{j(\cdot)}$.

4. Begin by picking an i,j such that $m_{ij} = \left\langle \hat{e}_i, M\hat{e}_j \right\rangle$ is the smallest $m_{ij} > 0$. Which we can always do because $M\vec{\mathbb{1}} = \vec{\mathbb{1}}$. If $m_{ij} = 1$ then by definition M is a vertex matrix and we are done.

5. By the same argument we can pick a function $f^{(1)}(k) : \{1,\ldots,n\} \mapsto S \subseteq \{1,\ldots,n\}$ such that $f^{(1)}(i) = j$ and $m_{kf^{(1)}(k)} \geq m_{ij}$ for all other $k \neq i$.

6. Letting $M^{(0)} = M$ we can decompose $M^{(1)}$ into the convex sum

$$M^{(0)} = \left(1 - m_{ij}\right) M^{(1)} + m_{ij} V_{f^{(1)}(\cdot)}$$

   It remains then to establish that $M^{(1)}$ is an element of the convex polytope of singly stochastic matrices.

7. Checking that $M^{(1)}$ has $\hat{\mathbb{1}}$ as an Eigen vector with Eigen value 1

$$\begin{aligned} M^{(1)}\hat{\mathbb{1}} &= \frac{M^{(0)}\hat{\mathbb{1}} - m_{ij} V_{f^{(1)}(\cdot)}\hat{\mathbb{1}}}{1 - m_{ij}} \\ &= \frac{1 - m_{ij}}{1 - m_{ij}}\hat{\mathbb{1}} \\ &= \hat{\mathbb{1}} \end{aligned}$$

8. We then check to ensure that $\left\langle \hat{e}_k, M^{(1)}\hat{e}_l \right\rangle > 0$ for every k,l

$$\begin{aligned} \left\langle \hat{e}_k, M^{(1)}\hat{e}_l \right\rangle &= \frac{\left\langle \hat{e}_k, M^{(0)}\hat{e}_l \right\rangle - m_{ij}\left\langle \hat{e}_k, V_{f^{(1)}(\cdot)}\hat{e}_l \right\rangle}{1 - m_{ij}} \\ &= \frac{m_{kl} - m_{ij}\delta_{f(k)l}}{1 - m_{ij}} \\ &\geq 0 \end{aligned}$$

   because $m_{kl} \geq m_{ij}$ for all k,l by construction, and $1 > m_{ij}$; or else we are done.

9. Next, observe that $\left\langle \hat{e}_i, M^{(1)}\hat{e}_j \right\rangle = 0$, so that we can carry out finite induction on this process, at most n^2 times.

10. Finally, the finite induction will generate matrices $M^{(t)}$ and $V_{f^{(t)}(\cdot)}$, proving that M is the convex sum of vertexes $V_{f^{(t)}(\cdot)}$. $\square$

The vertex dual matrices $C_{j(\cdot)}$ have a powerful commutator algebra. We start with a calculation of the products of vertex dual matrices.

Corollary 3. $C_{j(\cdot)}C_{k(\cdot)} = C_{k(j(\cdot))} - C_{j(\cdot)} - C_{k(\cdot)}$

Proof: Calculating directly from the definitions

$$\begin{aligned}
C_{j(\cdot)}C_{k(\cdot)} &= \left(\sum_{i=1}^{n} \hat{e}_i \otimes \hat{e}_{j(i)} - \hat{e}_i \otimes \hat{e}_i\right)\left(\sum_{i=1}^{n} \hat{e}_i \otimes \hat{e}_{k(i)} - \hat{e}_i \otimes \hat{e}_i\right) \\
&= \sum_{i=1}^{n} \hat{e}_i \otimes \hat{e}_{k(j(i))} - \hat{e}_i \otimes \hat{e}_{j(i)} - \hat{e}_i \otimes \hat{e}_{k(i)} + \hat{e}_i \otimes \hat{e}_i \\
&= C_{k(j(\cdot))} - C_{j(\cdot)} - C_{k(\cdot)} \\
&= C_{j\circ k(\cdot)} - C_{j(\cdot)} - C_{k(\cdot)}
\end{aligned}$$ $\square$

We can state immediately without proof, that the commutators of $C_{j(\cdot)}$ carry through with the commutation of $j(\cdot)$ and $k(\cdot)$.

Corollary 4. $\left[C_{j(\cdot)}, C_{k(\cdot)}\right] = C_{k(j(\cdot))} - C_{j(k(\cdot))}$

Furthermore, recognizing that $C_{j^0(\cdot)} = 0$, because $j^0_{(i)} = i$ for all i, we have that the powers of $C_{j(\cdot)}$ have the binomial form.

Corollary 5. *For* $n > 0$ *we have* $C^n_{j(\cdot)} = \sum_{m=0}^{n} \binom{n}{m} (-1)^{n-m} C_{j^m(\cdot)}$

Proof: Starting with $n = 1$ we carry through with induction.

1. Explicitly calculating for $n = 1$ we have

$$\begin{aligned}
C_{j(\cdot)} &= C_{j(\cdot)} - C_{j^0(\cdot)} \\
&= \sum_{m=0}^{1} \binom{1}{m} (-1)^{1-m} C_{j^m(\cdot)}
\end{aligned}$$

2. Now assume the claim is true for a fixed n, setting up the calculation for $n+1$ we have

$$\begin{aligned}
C_{j(\cdot)}^{n+1} &= C_{j(\cdot)} \sum_{m=0}^{n} \binom{n}{m} (-1)^{n-m} C_{j^m(\cdot)} \\
&= \sum_{m=1}^{n} \binom{n}{m} (-1)^{n-m} \left(C_{j^{m+1}(\cdot)} - C_{j^m(\cdot)} - C_{j(\cdot)} \right) \\
&= \binom{n+1}{1} (-1)^{(n+1)-1} C_{j(\cdot)} + \binom{n+1}{n+1} (-1)^{(n+1)-(n+1)} C_{j^{n+1}(\cdot)} \\
&\quad + \sum_{m=2}^{n} \left(\binom{n}{m-1} + \binom{n}{m} \right) (-1)^{(n+1)-m} C_{j^m(\cdot)} \\
&= \sum_{m=0}^{n+1} \binom{n+1}{m} (-1)^{(n+1)-m} C_{j^m(\cdot)} \qquad \square
\end{aligned}$$

As a special case, when $j(\cdot)$ is idempotent, $j^2(i) = j(i)$, the vertex dual matrix $C_{j(\cdot)}$ has odd parity, $C_{j(\cdot)}^2 = -C_{j(\cdot)}$. Furthermore in the case where $j(\cdot)$ is a permutation with period p we need only consider the vertex dual matrices for the first p compositions of $j(\cdot)$.

Corollary 6. *If $j(\cdot)$ is a permutation with period p then for $n > 0$ we have $\left(I + C_{j(\cdot)}\right)^n = I + C_{j^{n \bmod p}(\cdot)}$*

Proof: Calculating directly through the definition of the vertex $V_{j(\cdot)}$

$$\begin{aligned}
\left(I + C_{j(\cdot)}\right)^{n \bmod p} &= V_{j(\cdot)}^{n} \\
&= V_{j^{n \bmod p}(\cdot)} \\
&= I + C_{j^{n \bmod p}(\cdot)} \qquad \square
\end{aligned}$$

The result in the previous corollary hints at the importance of permutations in elucidating the structure of the vertex dual matrices. To develop this further for a given function $j(\cdot)$ we recursively define a partition of the set $\{1, \ldots, n\}$; composed of respectively the numbers for which $j(\cdot)$ is a permutation, and the numbers for which $j(\cdot)$ is transient, of each order.

Definition 5. The ergodic and transient subsets $E_{j(\cdot)}^{(n)} \subseteq \{1, \ldots, n\}$, partitioned by $j(\cdot)$, are recursively:

$$E_{j(\cdot)}^{(0)} = \{i : \exists p_i > 0 \text{ where } j^{p_i}(i) = i\}$$

$$E_{j(\cdot)}^{(n+1)} = \left\{ i : j(i) \in E_{j(\cdot)}^{(n)} \right\}$$

The partitioning of the integers into an ergodic subset and transient subsets allows us to construct restriction maps of $j(\cdot)$.

Definition 6. The ergodic and transient restrictions of $j(\cdot)$ are given by:

$$j_n(i) = \begin{cases} j(i) & i \in E_{j(\cdot)}^{(n)} \\ i & \text{otherwise} \end{cases}$$

If T is the highest order transient of $j(\cdot)$, so that T is the smallest t such that $j^t(i) \in E_{j(\cdot)}^{(0)}$, then nearly without any further proof we can state the following observations

Observation 1. *The following relationships for $j(\cdot)$, and $j_n(\cdot)$ hold*

1. $i \in E_{j(\cdot)}^{(0)} \Rightarrow j_0(i) = j(i) \in E_{j(\cdot)}^{(0)}$
2. $n > 0$ *and* $i \in E_{j(\cdot)}^{(n)} \Rightarrow j_n(i) = j(i) \in E_{j(\cdot)}^{(n-1)}$
3. $n > 0 \Rightarrow j_n^2(i) = j_n(i)$
4. $j(i) = j_0 \circ \cdots \circ j_T(i)$
5. $C_{j(\cdot)} = \sum_{n=1}^{T} C_{j_n(\cdot)}$
6. $m < n \Rightarrow \operatorname{im}\left(C_{j_n(\cdot)}\right) \leqslant \ker\left(C_{j_m(\cdot)}\right)$ *which follows from*

$$\begin{aligned} C_{j_m(\cdot)} C_{j_n(\cdot)} &= C_{j_n(j_m(\cdot))} - C_{j_m(\cdot)} - C_{j_n(\cdot)} \\ &= C_{j_m(\cdot)} + C_{j_n(\cdot)} - C_{j_m(\cdot)} - C_{j_n(\cdot)} \\ &= 0 \end{aligned}$$

The last observation admits a further generalization that will be of use in the following calculations

Observation 2. *If $m < n \Rightarrow [A_n, A_m] = A_n A_m$ then*

1. $\left(\sum_m A_m\right)^n = \sum_{m_1+m_2+\cdots=n} \cdots A_2^{m_2} A_1^{m_1}$

2. $\exp\left(\sum_m A_m\right) = \sum_{m_1,m_2,\cdots\geq 0} \frac{1}{(m_1+m_2+\cdots)!} \cdots A_2^{m_2} A_1^{m_1}$

By construction $j_0(\cdot)$ is a permutation, and at the very least has a period $P > 0$. As well, $j(\cdot)$ will have a highest transient order of at most $T > 0$. Taken together this implies that to find a limit in the stochastic Lie algebra $\mathfrak{st}(\hat{\mathbb{1}})$ that is equal to $V_{j_0(\cdot)}$ we need only consider a coefficient α_0 for the $C_{j_n(\cdot)}$ when $n > 0$, and $p-1$ coefficients α_n for the $p-1$ vertex dual matrices $C_{j_0^n(\cdot)}$. We propose a trial solution for $V_{j(\cdot)}$, to which we apply the previous observations to simply.

Proposition 1. *For a function $j(i) : \{1,\ldots,n\} \mapsto S \subseteq \{1,\ldots,n\}$*

$$V_{j(\cdot)} = \exp\left(\alpha_0 \sum_{l=1}^{T} C_{j_l(\cdot)} + \sum_{l=1}^{p-1} \alpha_l C_{j_0^l(\cdot)}\right)$$

With the vertex recovered by letting $\alpha_0 \to \infty$, and letting α_n take on the Pythagorean coefficients.

$$\alpha_n = \begin{cases} (-1)^{n+1} \frac{\pi}{p} \csc\left(\frac{n\pi}{p}\right) e^{\mathrm{i}\frac{n\pi}{p}} & p \text{ even} \\ (-1)^{n+1} \frac{\pi}{p} \csc\left(\frac{n\pi}{p}\right) & p \text{ odd} \end{cases}$$

The Pythagorean coefficients are so named because $\frac{\pi}{p}\csc\left(\frac{\pi}{p}\right)$ is ratio of π to the p inner Pythagorean approximation of π. The coefficients in the case of period of $p = 16$ are illustrated in figure 2.5, and in the case of period $p = 17$ in figure 2.6. For now the proposition remains unproven. However, a hint of the direction to take can be seen by expanding the power series.

$$V_{j(\cdot)} = e^{-\sum_{n=1}^{p-1}\alpha_n} \sum_{m_1,\ldots,m_{p-1}\geq 0} V_{j_0(\cdot)}^{\sum_{n=1}^{p-1} n m_n} \prod_{n=1}^{p-1} \frac{\alpha_n^{m_n}}{m_n!} + \left(1 - e^{-\alpha_0}\right) \sum_{l=1}^{T} C_{j_l(\cdot)} + \cdots$$
$$\cdots + (-1)^T C_{j_T(\cdot)} \cdots C_{j_1(\cdot)} \sum_{m_0,\ldots,m_T > 0} \frac{(-\alpha_1)^{m_1} \cdots (-\alpha_T)^{m_T}}{(m_0 + \cdots + m_T)!} \left(\sum_{l=1}^{p-1} \alpha_l C_{j_0^l(\cdot)}\right)^{m_0}$$

2.1.4 Contraction and Dual Algebras

It should be clear now that $St(\hat{\mathbb{1}})$ has a non-trivial structure; most importantly it is connected, but not simply connected. This can be seen because $St(\hat{\mathbb{1}})$ contains matrices with positive, negative, and complex determinants, while matrices with a determinant of zero are excluded, because they are not invertible. There is a simply connected normal sub-group of $St(\hat{\mathbb{1}})$ that has particular importance to continuous time homogeneous Markov processes on finite state spaces.

Definition 7. The stochastic contraction Lie group $St^{+}(\hat{\mathbb{1}})$ is the set of matrices such that $A \in St(\hat{\mathbb{1}})$ and $\det A \in \mathbb{R}^{+}$

This definition necessitates a proof of the claim in the previous paragraph.

Corollary 7. *$St^{+}(\hat{\mathbb{1}})$ is a simply connected normal sub-group of $St(\hat{\mathbb{1}})$*

Proof: It should be clear that $St^{+}(\hat{\mathbb{1}})$ is a Lie sub-group of $St(\hat{\mathbb{1}})$, thus we need only prove normality and simply connectedness.

1. Starting with normality, let $A \in St^{+}(\hat{\mathbb{1}})$ and $B \in St(\hat{\mathbb{1}})$, then

$$\begin{aligned}\det\left(BAB^{-1}\right) &= (\det B)(\det A)(\det B)^{-1} \\ &= \det A\end{aligned}$$

   thus $BAB^{-1} \in St^{+}(\hat{\mathbb{1}})$

2. It is sufficient to prove that $St^{+}(\hat{\mathbb{1}})$ is simply connected through the identity element.

3. Starting with a continuous path $A(t) \in St^{+}(\hat{\mathbb{1}})$ parameterized by $t \in [0,1]$ such that $A(0) = A(1) = I$, by definition of the Lie algebra there exists a continuous path $G(t) = \sum_{ij} \alpha_{ij}(t) C_{ij} \in \mathfrak{st}(\hat{\mathbb{1}})$ such that $A(t) = \exp G(t)$, and $\alpha_{ij}(0) = \alpha_{ij}(1) = 0$.

4. It follows that $\det A(t) = \exp\left(\sum_{ij} \alpha_{ij}(t)\right) \in \mathbb{R}^{+}$.

5. Now consider $s \in [0,1]$, and $A_s(t) = \exp(sG(t))$, then $A_1(t) = A(t)$ and $A_0(t) = I$, furthermore $\det A_s(t) = \exp\left(s\sum_{ij} \alpha_{ij}(t)\right) \in \mathbb{R}^{+}$.

6. Taking the limit as s goes to 0 provides the necessary simple connectedness. □

The use of the nomenclature of contraction is in deference to the equivalent definition used for the generators of continuous time homogeneous Markov processes on finite state spaces. Of course every good Lie group deserves a Lie algebra.

Definition 8. Let $\mathfrak{st}^+(\hat{\mathbb{1}})$ denote the stochastic contraction Lie algebra of $St^+(\hat{\mathbb{1}})$

This definition admits a similar characterization as before.

Corollary 8. *C_{ij} over $\mathbb{R}$ generates $\mathfrak{st}^+(\hat{\mathbb{1}})$.*

Proof: The result follows in much the same method as the central theorem of this chapter, except to show that $\mathfrak{st}^+(\hat{\mathbb{1}})$ is a real vector space over the basis C_{ij}. It is sufficient to check for closure with respect to linear sums and scalar multiplications:

1. For any $\alpha_{ij} \in \mathbb{C}$ such that $\exp\left(\sum_{ij} \alpha_{ij} C_{ij}\right) \in St^+(\hat{\mathbb{1}})$ Jacobi's formula requires that $\mathbb{I}m\left(\sum_{ij} \alpha_{ij}\right) \equiv 0 \bmod 2\pi$.

2. It follows that for $\exp\left(\sum_{ij} \alpha_{ij} C_{ij}\right), \exp\left(\sum_{ij} \beta_{ij} C_{ij}\right) \in St^+(\hat{\mathbb{1}})$ we have
$$\begin{aligned} 0 &\equiv \mathbb{I}m\left(\sum_{ij} \alpha_{ij}\right) + \mathbb{I}m\left(\sum_{ij} \beta_{ij}\right) \bmod 2\pi \\ &\equiv \mathbb{I}m\left(\sum_{ij} \alpha_{ij} + \beta_{ij}\right) \bmod 2\pi \end{aligned}$$

3. Thus $\exp\left(\sum_{ij} \left(\alpha_{ij} + \beta_{ij}\right) C_{ij}\right) \in St^+(\hat{\mathbb{1}})$.

4. Finally checking scalar multiplication, for fixed $a \in \mathbb{C}$, and any $\alpha_{ij} \in \mathbb{C}$ such that $\exp\left(\sum_{ij} \alpha_{ij} C_{ij}\right) \in St^+(\hat{\mathbb{1}})$ we have
$$\begin{aligned} 0 &\equiv \mathbb{I}m\left(a \sum_{ij} \alpha_{ij}\right) \bmod 2\pi \\ &\equiv \mathbb{R}e\,(a)\,\mathbb{I}m\left(\sum_{ij} \alpha_{ij}\right) + \mathbb{I}m\,(a)\,\mathbb{R}e\left(\sum_{ij} \alpha_{ij}\right) \bmod 2\pi \\ &\equiv \mathbb{I}m\,(a)\,\mathbb{R}e\left(\sum_{ij} \alpha_{ij}\right) \bmod 2\pi \\ &= \mathbb{I}m\,(a) \end{aligned}$$

5. It follows then that $\mathbb{I}m(a) = 0$ $\square$

Note that $\mathfrak{st}^{+}(\hat{\mathbb{1}})$ is a real vector space that is augmented with the group of integers $\mathbb{Z}$. This is because every real point has added to it every multiple of $\mathrm{i}2\pi$.

That $St^{+}(\hat{\mathbb{1}})$ is a simply connected normal Lie sub-group has an important consequence for the generator estimation methods developed in the next chapter. The methods are all constrained to algebraic operations, so that by closure of the Lie sub-algebra the algorithms will always result in generators from $\mathfrak{st}^{+}(\hat{\mathbb{1}})$. This can be seen because any continuous time homogeneous path through $St(\hat{\mathbb{1}})$ must always start at the identity matrix. Thus if the basis $\hat{e}_i$ enumerates a finite state space, then the generator estimated by algebraic operations with respect to C_{ij} will always have real (positive) expansion in the basis C_{ij}. This will occur even if a suitable complex generator is used to generate real (positive) transition probabilities. We can interpret this as meaning that $\mathfrak{st}^{+}(\hat{\mathbb{1}})$ is a closed branch of the matrix logarithm.

We have developed an interpretation of the Eigen equation $A\hat{\mathbb{1}} = \hat{\mathbb{1}}$ as a conservation of the row sums of A; likewise the Eigen equation $A^{\dagger}\hat{\mathbb{1}} = \hat{\mathbb{1}}$ can be interpreted as the conservation of the column sums of A. The dual definitions for the Lie group and algebra follow natural.

Definition 9. Let $St^{\dagger}(\hat{\mathbb{1}})$ denote the dual stochastic Lie group of invertible matrices whose transpose is stochastic with respect to $\hat{\mathbb{1}}$.

Definition 10. Let $\mathfrak{st}^{\dagger}(\hat{\mathbb{1}})$ denote the dual stochastic Lie algebra of $St^{\dagger}(\hat{\mathbb{1}})$.

Thus if C_{ij} are generators of $\mathfrak{st}(\hat{\mathbb{1}})$ then $C_{ij}^{\dagger} = \left(\hat{e}_j - \hat{e}_i\right) \otimes \hat{e}_i$ are generators of $\mathfrak{st}^{\dagger}(\hat{\mathbb{1}})$. The definitions of the dual stochastic contraction Lie group $St^{+\dagger}(\hat{\mathbb{1}})$ and Lie algebra $\mathfrak{st}^{+\dagger}(\hat{\mathbb{1}})$ follow analogously. That $St^{+}(\hat{\mathbb{1}}) \cap St^{+\dagger}(\hat{\mathbb{1}}) \subseteq St(\hat{\mathbb{1}}) \cap St^{\dagger}(\hat{\mathbb{1}}) \subseteq St(\hat{\mathbb{1}})$ are a Lie groups and $\mathfrak{st}^{+}(\hat{\mathbb{1}}) \cap \mathfrak{st}^{+\dagger}(\hat{\mathbb{1}}) \subseteq \mathfrak{st}(\hat{\mathbb{1}}) \cap \mathfrak{st}^{\dagger}(\hat{\mathbb{1}}) \subseteq \mathfrak{st}(\hat{\mathbb{1}})$ are a Lie algebras will be foundational for the next section.

2.2 Doubly Stochastic Matrices

2.2.1 Preliminaries

Doubly stochastic matrices require row and column conservation of the vector $\hat{\mathbb{1}}$, in the sense that both $A\hat{\mathbb{1}} = \hat{\mathbb{1}}$ and $A^{\dagger}\hat{\mathbb{1}} = \hat{\mathbb{1}}$ must hold. The group of invertible doubly stochastic matrices is then a subgroup of the group of stochastic matrices. The two constraints of row and column conservation leaves only $(n-1)^2$ linear degrees of freedom. This will be an important clue in the construction of canonical generators. In fact the canonical generators can be found by choosing one additional vector $\hat{e}_n$, from the basis constructed in the previous section, to center the combinatorial construction of the generators of the algebra around. This vector plays a similar role to the diagonal in the previous construction and is used to balance the row and column sums back to zero. As in the previous section we start with a foundational definition.

Definition 11. Let $St(\hat{\mathbb{1}},\hat{\mathbb{1}})$ denote the doubly stochastic Lie group of invertible matrices A such that both A and $A^{\dagger}$ are stochastic with respect to $\hat{\mathbb{1}}$

We can immediately observe without proof that $St(\hat{\mathbb{1}},\hat{\mathbb{1}}) = St(\hat{\mathbb{1}}) \cap St^{\dagger}(\hat{\mathbb{1}})$; leading to the next definition.

Definition 12. Let $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$ denote the doubly stochastic Lie algebra of $St(\hat{\mathbb{1}},\hat{\mathbb{1}})$.

It should be clear that $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}}) = \mathfrak{st}(\hat{\mathbb{1}}) \cap \mathfrak{st}^{\dagger}(\hat{\mathbb{1}})$. The implication being that $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$ is the algebra of all matrices A such that $\hat{\mathbb{1}}$ is in the kernel of both A and $A^{\dagger}$. As with $St(\hat{\mathbb{1}})$, the Lie group $St(\hat{\mathbb{1}},\hat{\mathbb{1}})$ is not simply connected. It contains the splitting contraction sub-group $St^{+}(\hat{\mathbb{1}},\hat{\mathbb{1}}) = St^{+}(\hat{\mathbb{1}}) \cap St^{+\dagger}(\hat{\mathbb{1}})$, and analogously defined contraction sub-algebra $\mathfrak{st}^{+}(\hat{\mathbb{1}},\hat{\mathbb{1}}) = \mathfrak{st}^{+}(\hat{\mathbb{1}}) \cap \mathfrak{st}^{+\dagger}(\hat{\mathbb{1}})$. $St^{+}(\hat{\mathbb{1}},\hat{\mathbb{1}})$ has the same properties of normality and simple connectedness within $St(\hat{\mathbb{1}},\hat{\mathbb{1}})$; of course it is not normal within $St(\hat{\mathbb{1}})$, because $St(\hat{\mathbb{1}},\hat{\mathbb{1}})$ is not normal within $St(\hat{\mathbb{1}})$.

2.2.2 Canonical Generators

We can find canonical generators of $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$ by similar methods as in the previous section. Given a constructed basis $\hat{e}_i$ such that $\left\langle \hat{e}_i,\hat{\mathbb{1}}\right\rangle = \frac{1}{\sqrt{n}}$ we pick a single arbitrary element from the basis, say $\hat{e}_n$, the last element for example. We then balance a transition rate from i to j, with the reverse rates from j to n and n to i, yielding the matrix $C_{ijn} = C_{ij} + C_{ni} + C_{jn}$. The state transitions of the symmetric matrix C_{iin} are illustrated in figure 2.2, and the state transitions of the asymmetric matrix C_{ijn} are illustrated in figure 2.3.

The matrices C_{ijn} are elements of $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$. Furthermore they are a closed set with respect to matrix transposition, because $C^{\dagger}_{ijn} = C_{jin}$. The algebra $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$ is isomorphic to the space of $(n-1)\times(n-1)$ matrices; which can be seen by the from the improper commutators, where for any $i,j \leq n-1$.

$$\begin{aligned}
C_{iin} + C_{jjn} - C_{ijn} &= C_{in} + C_{nj} - C_{ij} \\
&= \hat{e}_i \otimes \hat{e}_n + \hat{e}_n \otimes \hat{e}_j - \hat{e}_i \otimes \hat{e}_j - \hat{e}_n \otimes \hat{e}_n
\end{aligned}$$

The state transitions of the improper commutators $C_{iin} + C_{jjn} - C_{ijn}$ are illustrated in figure 2.4.

The intuition being that any $n\times n$ matrix with fixed row and column sums can be created by starting with any $(n-1)\times(n-1)$ matrix and appending a compensating n row and n column. This relationship is implicitly used extensively in proving the following lemma and corollary on the products, commutators, and structure constants of C_{ijn}.

Lemma 6.

$$C_{ijn}C_{kln} = \begin{cases} C_{jin} - 2C_{ijn} & i = k \text{ and } j = l, \\ -\left(C_{ijn} + C_{jin}\right) & i = l \text{ and } j = k, \\ C_{jin} - C_{jjn} - C_{iin} + C_{lln} - C_{iln} & i = k, \\ C_{jkn} - C_{jjn} - C_{kkn} - C_{iin} & i = l, \\ C_{iln} - C_{ijn} - C_{jln} & j = k, \\ C_{jkn} - C_{jjn} - C_{kkn} + C_{iin} - C_{ijn} & j = l, \\ C_{jkn} - C_{jjn} - C_{kkn} & \text{otherwise}. \end{cases}$$

Proof: We proceed by calculating the terms of the products and then simplifying the cases; assuming $i \neq j$, $k \neq l$, and $i, j, k, l \neq n$.

1. Term wise computation of the Kronecker products yields

$$\begin{aligned} C_{ijn}C_{kln} &= \left(\hat{e}_i \otimes \hat{e}_j - \hat{e}_i \otimes \hat{e}_i + \hat{e}_n \otimes \hat{e}_i - \hat{e}_n \otimes \hat{e}_n + \hat{e}_j \otimes \hat{e}_n - \hat{e}_j \otimes \hat{e}_j\right) \\ &\quad \cdot \left(\hat{e}_k \otimes \hat{e}_l - \hat{e}_k \otimes \hat{e}_k + \hat{e}_n \otimes \hat{e}_k - \hat{e}_n \otimes \hat{e}_n + \hat{e}_l \otimes \hat{e}_n - \hat{e}_l \otimes \hat{e}_l\right) \\ &= -\hat{e}_n \otimes \hat{e}_k + \hat{e}_n \otimes \hat{e}_n + \hat{e}_j \otimes \hat{e}_k - \hat{e}_j \otimes \hat{e}_n \\ &\quad + \delta_{ik}\left(-\hat{e}_i \otimes \hat{e}_l + \hat{e}_i \otimes \hat{e}_k + \hat{e}_n \otimes \hat{e}_l - \hat{e}_n \otimes \hat{e}_k\right) \\ &\quad + \delta_{il}\left(-\hat{e}_i \otimes \hat{e}_n + \hat{e}_i \otimes \hat{e}_l + \hat{e}_n \otimes \hat{e}_n - \hat{e}_n \otimes \hat{e}_l\right) \\ &\quad + \delta_{jk}\left(\hat{e}_i \otimes \hat{e}_l - \hat{e}_i \otimes \hat{e}_k - \hat{e}_j \otimes \hat{e}_l + \hat{e}_j \otimes \hat{e}_k\right) \\ &\quad + \delta_{jl}\left(\hat{e}_i \otimes \hat{e}_n - \hat{e}_i \otimes \hat{e}_l - \hat{e}_j \otimes \hat{e}_n + \hat{e}_j \otimes \hat{e}_l\right) \\ &= C_{jkn} - C_{jjn} - C_{kkn} + \delta_{ik}\left(C_{lln} - C_{iln}\right) - \delta_{il}C_{iin} \\ &\quad + \delta_{jk}\left(C_{jjn} + C_{iln} - C_{ijn} - C_{jln}\right) + \delta_{jl}\left(C_{iin} - C_{ijn}\right) \end{aligned}$$

2. The cases follow from simplifying the δ functions; starting with $i = k$ and $j = l$

$$\begin{aligned} C_{ijn}C_{ijn} &= C_{jin} - C_{jjn} - C_{iin} + C_{jjn} - C_{ijn} + C_{iin} - C_{ijn} \\ &= C_{jin} - 2C_{ijn} \end{aligned}$$

3. When $i = l$ and $j = k$

$$C_{ijn}C_{jin} = C_{jjn} - C_{jjn} - C_{jjn} - C_{iin} + C_{jjn} + C_{iin} - C_{ijn} - C_{jin}$$
$$= -\left(C_{ijn} + C_{jin}\right)$$

4. When $i = k$

$$C_{ijn}C_{iln} = C_{jin} - C_{jjn} - C_{iin} + C_{lln} - C_{iln}$$

5. When $i = l$

$$C_{ijn}C_{kin} = C_{jkn} - C_{jjn} - C_{kkn} - C_{iin}$$

6. When $j = k$

$$C_{ijn}C_{jln} = C_{jjn} - C_{jjn} - C_{jjn} + C_{jjn} + C_{iln} - C_{ijn} - C_{jln}$$
$$= C_{iln} - C_{ijn} - C_{jln}$$

7. When $j = l$

$$C_{ijn}C_{kjn} = C_{jkn} - C_{jjn} - C_{kkn} + C_{iin} - C_{ijn}$$

8. When none of the conditions apply

$$C_{ijn}C_{kln} = C_{jkn} - C_{jjn} - C_{kkn} \qquad \square$$

Moving immediately to the commutators we have:

Corollary 9.

$$[C_{ijn}, C_{kln}] = \begin{cases} 0 & i = k \text{ and } j = l, \\ 0 & i = l \text{ and } j = k, \\ C_{jin} - 2C_{jjn} + 2C_{lln} - C_{iln} - C_{lin} + C_{ijn} & i = k, \\ C_{jkn} - C_{jjn} - C_{kkn} - C_{iin} - C_{kjn} + C_{kin} + C_{ijn} & i = l, \\ C_{iln} - C_{ijn} - C_{jln} - C_{lin} + C_{lln} + C_{iin} + C_{jjn} & j = k, \\ C_{jkn} - 2C_{kkn} + 2C_{iin} - C_{ijn} - C_{jin} + C_{kjn} & j = l, \\ C_{jkn} - C_{jjn} - C_{kkn} - C_{lin} + C_{lln} + C_{iin} & \text{otherwise}. \end{cases}$$

Proof: We work case wise through the equalities; assuming $i \neq j$, $k \neq l$, and $i, j, k, l \neq n$.

1. Starting with $i = k$ and $j = l$

$$\begin{aligned} [C_{ijn}, C_{ijn}] &= C_{ijn}C_{ijn} - C_{ijn}C_{ijn} \\ &= 0 \end{aligned}$$

2. When $i = l$ and $j = k$

$$\begin{aligned} [C_{ijn}, C_{jin}] &= C_{ijn}C_{jin} - C_{jin}C_{ijn} \\ &= -(C_{ijn} + C_{jin}) + (C_{jin} + C_{ijn}) \\ &= 0 \end{aligned}$$

3. When $i = k$

$$\begin{aligned} [C_{ijn}, C_{iln}] &= C_{ijn}C_{iln} - C_{iln}C_{ijn} \\ &= (C_{jin} - C_{jjn} - C_{iin} + C_{lln} - C_{iln}) \\ &\quad - (C_{lin} - C_{lln} - C_{iin} + C_{jjn} - C_{ijn}) \\ &= C_{jin} - 2C_{jjn} + 2C_{lln} - C_{iln} - C_{lin} + C_{ijn} \end{aligned}$$

4. When $i = l$

$$\begin{aligned}\left[C_{ijn}, C_{kin}\right] &= C_{ijn}C_{kin} - C_{kin}C_{ijn} \\ &= \left(C_{jkn} - C_{jjn} - C_{kkn} - C_{iin}\right) - \left(C_{kjn} - C_{kin} - C_{ijn}\right) \\ &= C_{jkn} - C_{jjn} - C_{kkn} - C_{iin} - C_{kjn} + C_{kin} + C_{ijn}\end{aligned}$$

5. When $j = k$

$$\begin{aligned}\left[C_{ijn}, C_{jln}\right] &= C_{ijn}C_{jln} - C_{jln}C_{ijn} \\ &= \left(C_{iln} - C_{ijn} - C_{jln}\right) - \left(C_{lin} - C_{lln} - C_{iin} - C_{jjn}\right) \\ &= C_{iln} - C_{ijn} - C_{jln} - C_{lin} + C_{lln} + C_{iin} + C_{jjn}\end{aligned}$$

6. When $j = l$

$$\begin{aligned}\left[C_{ijn}, C_{kjn}\right] &= C_{ijn}C_{kjn} - C_{kjn}C_{ijn} \\ &= \left(C_{jkn} - C_{jjn} - C_{kkn} + C_{iin} - C_{ijn}\right) \\ &\quad - \left(C_{jin} - C_{jjn} - C_{iin} + C_{kkn} - C_{kjn}\right) \\ &= C_{jkn} - 2C_{kkn} + 2C_{iin} - C_{ijn} - C_{jin} + C_{kjn}\end{aligned}$$

7. When none of the conditions apply

$$\begin{aligned}\left[C_{ijn}, C_{kln}\right] &= C_{ijn}C_{kln} - C_{kln}C_{ijn} \\ &= C_{jkn} - C_{jjn} - C_{kkn} - C_{lin} + C_{lln} + C_{iin}\end{aligned}$$ □

Restricting to the matrices where both the row and column sums are zero, which is equivalent to demanding the conservation of the transition rates, or infinitesimal flows of probability, introduces a significant degree of complexity to the algebra. In particular demanding that all the transition rates be balanced by transitions through $\hat{e}_n$ means that only the simplest two and three state processes have easily calculable algebras. Nevertheless, the result makes the sibling theorem accessible.

Theorem 2. *C_{ijn} are canonical generators of* $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$.

Proof: The proof proceeds in the same manner as the proof of the sibling theorem in the previous section.

1. As discussed before $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$ is an $(n-1)^2$ dimensional vector space.

2. By construction there are only $(n-1)^2$ matrices C_{ijn} for a fixed choice of $\hat{e}_n$.

3. Through induction the matrices C_{ijn} are linearly independent for a fixed choice of $\hat{e}_n$.

4. Thus the matrices C_{ijn}, for a fixed choice of $\hat{e}_n$, are a basis for $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$.

5. By the previous lemma the commutators of matrices C_{ijn} are linear combinations of themselves.

6. It follows then that the smallest algebra that contains the matrices C_{ijn} is $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$.□

As with the stochastic Lie algebra the generators of the doubly stochastic Lie algebra are not unique, not only do they depend on the choice of the basis $\hat{e}_i$ but also on the choice of the basis element $\hat{e}_n$ used to sum the rows and columns to zero.

2.2.3 Vertex Logarithms

In the case of the doubly stochastic Lie algebra $\mathfrak{st}(\hat{\mathbb{1}},\hat{\mathbb{1}})$, the basis elements $\hat{e}_i$ define a special convex polytope known as the Birkhoff polytope. This polytope is given by the matrices doubly stochastic with respect to $\hat{\mathbb{1}}$, and that have nonnegative entries with respect to $\hat{e}_i$. By the Birkhoff-von Neumann theorem there are $n!$ vertexes of the Birkhoff polytope, $V_{j(\cdot)}$, given by the $n!$ permutations $j(i) : \{1,\ldots,n\} \mapsto \{1,\ldots,n\}$.

From the work in the previous section on the vertexes of the convex polytope of singly stochastic matrices we can state a more specific version of the proposition from the previous section for the vertexes of the Birkhoff polytope are:

Proposition 2. *For a permutation $j(i) : \{1,\ldots,n\} \mapsto \{1,\ldots,n\}$, with period p*

$$V_{j(\cdot)} = \exp\left(\sum_{n=1}^{p-1} \alpha_n C_{j^n(\cdot)}\right)$$

The vertex recovered is by the Pythagorean coefficients.

$$\alpha_n = \begin{cases} (-1)^{n+1}\, \frac{\pi}{p} \csc\left(\frac{n\pi}{p}\right) e^{\mathrm{i}\frac{n\pi}{p}} & p \text{ even} \\ (-1)^{n+1}\, \frac{\pi}{p} \csc\left(\frac{n\pi}{p}\right) & p \text{ odd} \end{cases}$$

The logarithm is not unique, nor is it even the principle branch in general. Other logarithms can be quickly generated by further decomposing the permutation $j(\cdot)$ into sub-cycles, and applying the same formula for the coefficients α_n.

2.3 Figures and Illustrations

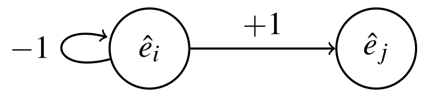

Figure 2.1: State transition circuit diagram of the canonical generators C_{ij} of $\mathfrak{st}(\hat{\mathbb{1}})$

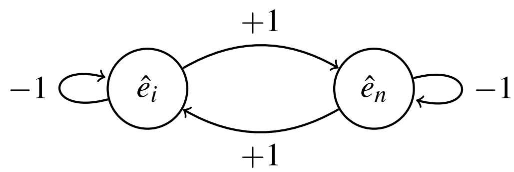

Figure 2.2: State transition circuit diagram of the canonical generators C_{iin} of $\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}})$

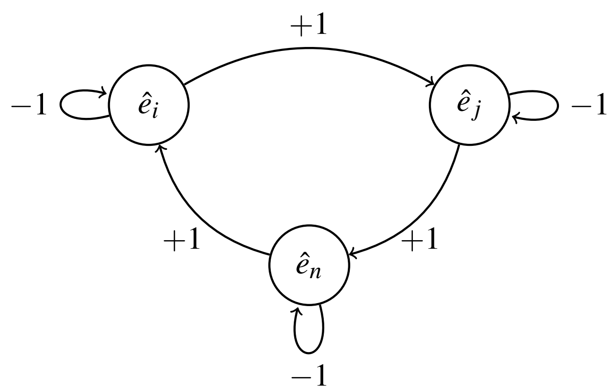

Figure 2.3: State transition circuit diagram of the canonical generators C_{ijn} of $\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}})$

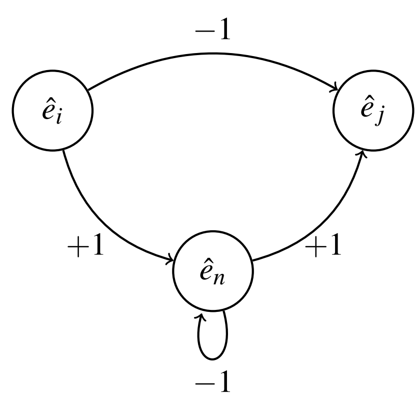

Figure 2.4: State transition circuit diagram of the improper generators $C_{iin} + C_{jjn} - C_{ijn}$ of $\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}})$

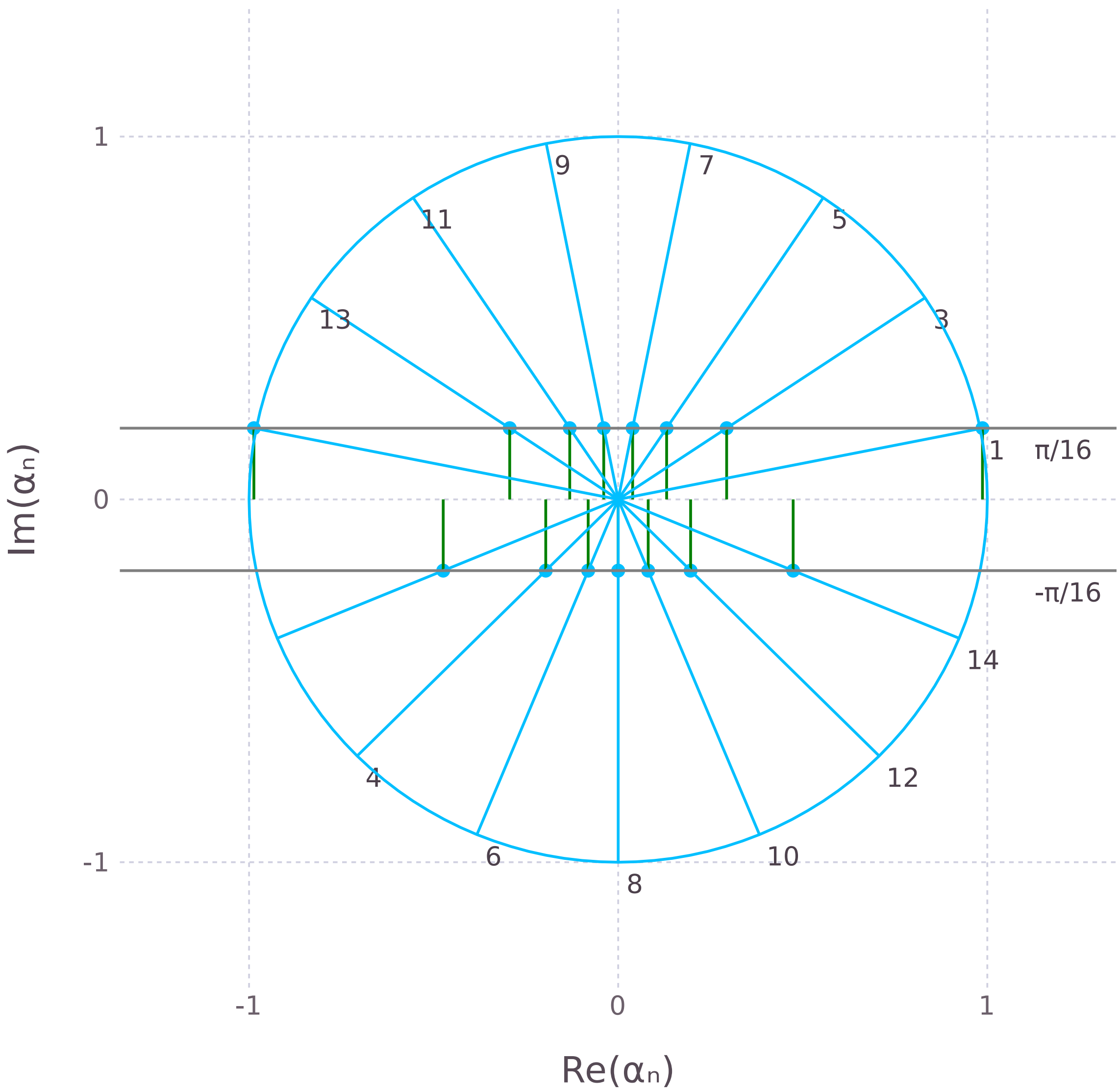

Figure 2.5: Complex plain geometry of the Pythagorean coefficients α_n of period $p = 16$. Blue points represent the location of the coefficients in the complex plane. Green lines illustrate the projection of the real component. Grey lines illustrate the imaginary component $\pm\frac{\pi}{16}$.

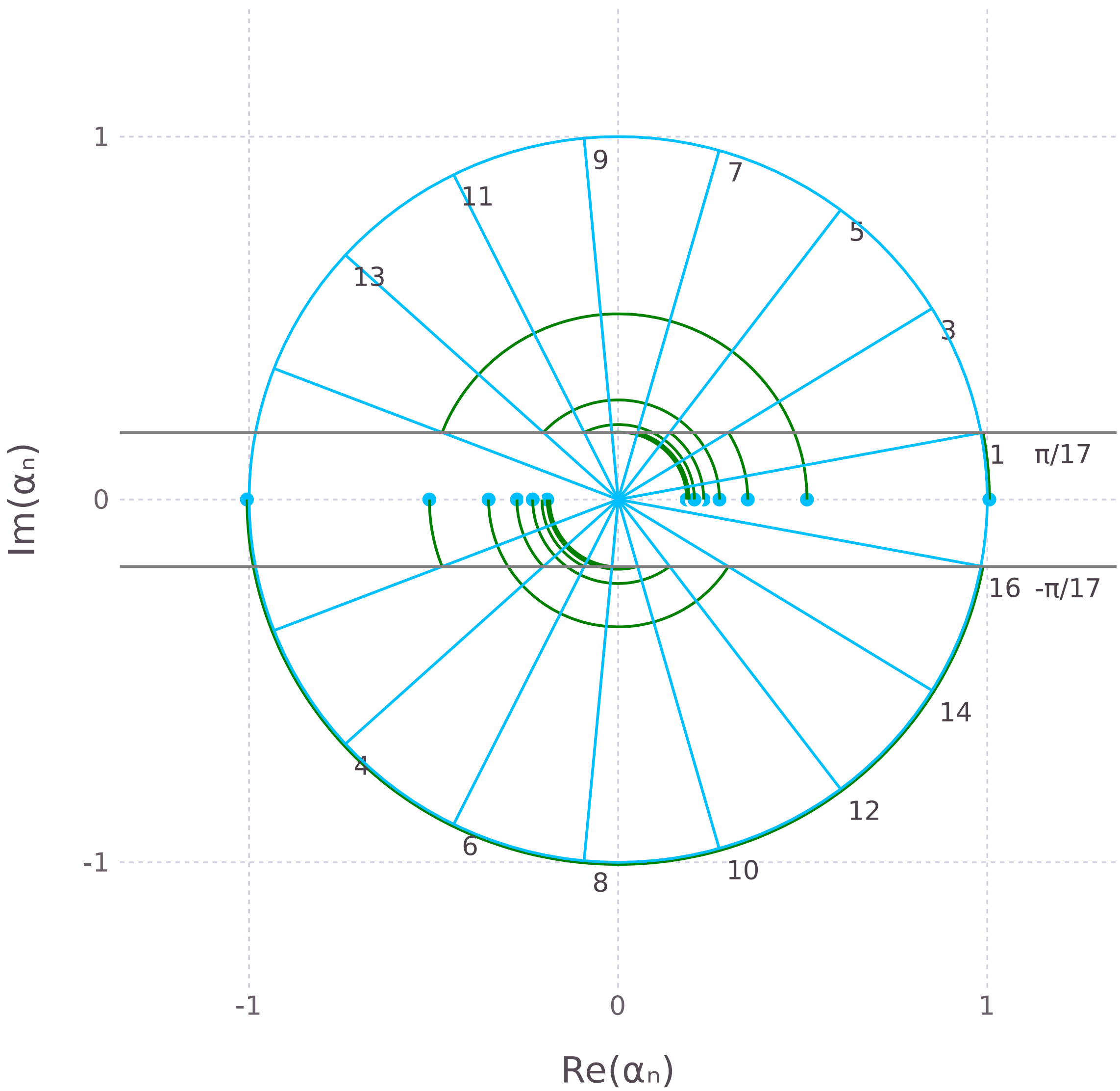

Figure 2.6: Complex plain geometry of the Pythagorean coefficients α_n of period $p = 17$. Blue points represent the location of the coefficients in the complex plane. Green lines carry the amplitude though to the intersection of the phase with the imaginary component $\pm\frac{\pi}{17}$.

Chapter 3

Padé Approximation of the Fréchet Derivatives of the Exponential Map

3.1 The Gradient

Moler and Van Loan seminally reviewed algorithms for calculating the matrix exponential in 1978, and revisited that review in 2003 [14, 15]. Building on the discussions of Moler and Van Loan, Higham established the standard implementation of the matrix exponential based on scaling and scaring and Padé approximation [10, 11]. The Higham implementation was further optimized for 64 bit architectures by Al-Mohy [3]. In the same work Al-Mohy developed an algorithm to approximate the derivative of the matrix exponential, formulated by taking the derivative of the Padé approximation of the matrix exponential, and then working out a recursive calculation for the derivatives of matrix powers [2].

While the derivative of the Padé approximation of an analytic function will converge to the derivative of the analytic function, it is not true that the derivative of the Padé approximation of an analytic function is the Padé approximation of the derivative of an analytic function. In the sense that Padé approximations of analytic functions are an optimal series of algebraic approximations the 2009 method proposed by Al-Mohy is not optimal.

In this chapter we will develop an approximation for the first, and second order Fréchet derivatives of the matrix exponential, by decomposing the derivatives into components that hold for the commutative condition, and components containing the perturbation due to non-commutativity. We will then derive the Padé approximation for the non-commutative perturbation. We begin by listing the eight forms of the Fréchet derivative of exponential map, in the direction $\frac{\partial X}{\partial x}$ at the point

X in the Lie algebra. [1] [2]

$$\left.\begin{aligned}\frac{\partial e^X}{\partial x} &= e^X\left[\int_0^1 e^{-s\,\mathrm{ad}_X\cdot}\,ds\right]\left(\frac{\partial X}{\partial x}\right)\\ &= e^X\left[\frac{1-e^{-\mathrm{ad}_X\cdot}}{\mathrm{ad}_X\cdot}\right]\left(\frac{\partial X}{\partial x}\right)\\ &= e^X\left[\sum_{n=0}^{\infty}\frac{(-1)^n}{(n+1)!}\,\mathrm{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)\end{aligned}\right\}\text{left recursive}$$

$$= \left[\frac{\mathrm{ad}_{e^X}\cdot}{\mathrm{ad}_X\cdot}\right]\left(\frac{\partial X}{\partial x}\right) \qquad \text{adjoint ratio}$$

$$= e^{\frac{1}{2}X}\left[\frac{e^{\frac{1}{2}\mathrm{ad}_X\cdot}-e^{-\frac{1}{2}\mathrm{ad}_X\cdot}}{\mathrm{ad}_X\cdot}\right]\left(\frac{\partial X}{\partial x}\right)e^{\frac{1}{2}X} \qquad \text{hyperbolic}$$

$$\left.\begin{aligned}&= \left[\int_0^1 e^{s\,\mathrm{ad}_X\cdot}\,ds\right]\left(\frac{\partial X}{\partial x}\right)e^X\\ &= \left[\frac{e^{\mathrm{ad}_X\cdot}-1}{\mathrm{ad}_X\cdot}\right]\left(\frac{\partial X}{\partial x}\right)e^X\\ &= \left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\,\mathrm{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)e^X\end{aligned}\right\}\text{right recursive}$$

The last equality demonstrates the non-commutative perturbation term most clearly; where the first multiplicative factor in the derivative accounts for the lack of commutativity between X and $\frac{\partial X}{\partial x}$, and the last term resembles the derivative in the commutative case. This can be seen clearly when considering the condition $\left[X, \frac{\partial X}{\partial x}\right] = 0$ in which case $\frac{\partial e^X}{\partial x} = \frac{\partial X}{\partial x}e^X$.

Even though the multiplicative factorization provides a transparent representation of the computational terms it is still far from optimal; because, when compared to matrix addition, matrix multiplication is both computationally more expensive, and less numerically stable. The numerical stability, and efficiency can be improved by decomposing the first multiplicative factor into a

[1] We have abused and confounded the notations for directional derivatives and partial derivatives here by assuming that X is parameterized by x so that $\frac{\partial e^X}{\partial x}$ is the derivative in the direction of change of x.

[2] With respect to the adjoint operator, we are using the currying partial application notation of $[Lf(\cdot)](y)$ to indicate the application of the operator L to $f(x)$ followed by evaluation of the result at y.

linear sum of the non-commutative perturbation term, which will reduce to 0 when $\left[X, \frac{\partial X}{\partial x}\right] = 0$, and an invariant term that contains the commutative relationship for all X.

$$
\begin{aligned}
\frac{\partial e^X}{\partial x} &= \left[\frac{e^{\mathrm{ad}_X \cdot} - 1 - \mathrm{ad}_X \cdot}{\mathrm{ad}_X^2 \cdot}\right] \left(\mathrm{ad}_X \frac{\partial X}{\partial x}\right) e^X + \frac{\partial X}{\partial x} e^X \\
&= \underbrace{\left[\sum_{n=0}^{\infty} \frac{1}{(n+2)!} \mathrm{ad}_X^n \cdot\right] \left(\mathrm{ad}_X \frac{\partial X}{\partial x}\right) e^X}_{\text{non-commutative anomaly}} + \underbrace{\frac{\partial X}{\partial x}}_{\text{invariant}} e^X
\end{aligned}
$$

Formally the infinite series in the non-commutative perturbation is related to the lower incomplete gamma function $\gamma(n,x)$. This can be seen by considering the general case when the offset of 2 in the factorial is allowed to be any natural number n, and then restating the sum in terms of a truncated exponential series.

$$
\begin{aligned}
\sum_{m=0}^{\infty} \frac{x^m}{(m+n)!} &= \frac{1}{x^n} \sum_{m=n}^{\infty} \frac{x^m}{m!} \\
&= \frac{1}{x^n} \left(e^x - \sum_{m=0}^{n-1} \frac{x^m}{m!}\right) \\
&= \frac{1}{x^n} \left(e^x - e^x \frac{\Gamma(n,x)}{\Gamma(n)}\right) \\
&= \frac{e^x}{(n-1)!x^n} \left(\int_0^{\infty} t^{n-1} e^{-t} dt - \int_x t^{n-1} e^{-t} dt\right) \\
&= \frac{e^x}{(n-1)!x^n} \int_0^x t^{n-1} e^{-t} dt \\
&= \frac{e^x}{(n-1)!x^n} \gamma(n,x)
\end{aligned}
$$

The non-commutative perturbation series is linear in $\frac{\partial X}{\partial x}$ and a Taylor series in the powers of $\mathrm{ad}_X \cdot$. Thus any computation of an approximation will be in the powers of $\mathrm{ad}_X \cdot$. As was discussed in the Moler and Van Loan [14, 15], naive computation of the Taylor series itself results in an approximation that will converge slowly, requiring a larger number of powers to be computed before the threshold of floating point error is reached. Padé approximation by rational functions remedy this problem, by offering convergence to the threshold of floating point error in smaller powers, and fewer computational steps.

However the question remains, given that $\frac{e^x - 1 - x}{x^2}$ is a rational perturbation of e^x, why not simply reuse the polynomials of the Padé approximation of the exponential function to compute new polynomials for a rational approximation of the non-commutative perturbation Taylor series. This method has two shortcomings: first, the approximation found in this manner is not itself a Padé approximation of the anomaly Taylor series, and so is not bound by the same theoretical asymptotic results as Padé approximations; second, computation by $\frac{e^x-1-x}{x^2}$ suffers from the same floating point errors near 0 as naive computation of e^x-1 by first computing e^x and then subtracting 1.

While it is clear that $\frac{e^x}{x^2}\gamma(2,x) : x \mapsto \frac{e^x-1-x}{x^2}$ is analytic for $x \in \mathbb{R}$ or $x \in \mathbb{C}$, and thus can be approximated by a Padé series with coefficients in $\mathbb{C}$; that the rational approximation with the same coefficients can be extended to $\mathrm{ad}_X\cdot$ requires more careful consideration.

When X is in $\mathfrak{st}(\hat{\mathbb{1}})$ the adjoint operator $\mathrm{ad}_X\cdot$ is a linear endomorphism on the vector space $\mathfrak{st}(\hat{\mathbb{1}})$; and so belongs to the algebra of general linear operators $GL\left(\mathfrak{st}(\hat{\mathbb{1}})\right)$. For a Padé approximation with numerator polynomial $P(X)$ and denominator polynomial $Q(X)$, that $\mathrm{ad}_X\cdot \in GL\left(\mathfrak{st}(\hat{\mathbb{1}})\right)$ implies that $P(\mathrm{ad}_X\cdot), Q(\mathrm{ad}_X\cdot) \in GL\left(\mathfrak{st}(\hat{\mathbb{1}})\right)$. It follows that when a solution Y to $[P(\mathrm{ad}_X\cdot)]\left(\frac{\partial X}{\partial x}\right) = [Q(\mathrm{ad}_X\cdot)](Y)$ exists it is guaranteed to belong to $\mathfrak{st}(\hat{\mathbb{1}})$, because $\frac{\partial X}{\partial x} \in \mathfrak{st}(\hat{\mathbb{1}})$.

The remaining question is whether or not Y, a solution to $[P(\mathrm{ad}_X\cdot)]\left(\frac{\partial X}{\partial x}\right) = [Q(\mathrm{ad}_X\cdot)](Y)$, is an approximation of $\left[\sum_{n=0}^{\infty} \frac{1}{(n+2)!}\mathrm{ad}_X^n\cdot\right]\left(\mathrm{ad}_X \frac{\partial X}{\partial x}\right)$? Loosely, $\mathfrak{st}(\hat{\mathbb{1}})$ is a normed vector space in the usual sense, so the convergence of the Padé approximations that apply in scalar spaces carry over. Thus given a sequence of Padé approximations Y_{mn} that solve $[P_n(\mathrm{ad}_X\cdot)]\left(\frac{\partial X}{\partial x}\right) = [Q_m(\mathrm{ad}_X\cdot)](Y_{mn})$, convergence $Y_{mn} \to \left[\sum_{n=0}^{\infty} \frac{1}{(n+2)!}\mathrm{ad}_X^n\cdot\right]\left(\mathrm{ad}_X \frac{\partial X}{\partial x}\right)$ is assured.

Recapitulating $[n/m]_f(x)$ Padé approximations, we seek a rational polynomial approximation to the Taylor series.

$$\begin{aligned}\frac{P_n(x)}{Q_m(x)} &= \frac{p_0 + p_1 x + \cdots + p_n x^n}{1 + q_1 x + \cdots + q_m x^m} \\ &\approx \sum_{n=0}^{\infty} \frac{1}{(n+2)!} x^n\end{aligned}$$

Such that the first $k \leq n+m$ derivatives of the rational polynomial approximation evaluated at $x=0$ equal the first $k \leq n+m$ coefficients of the Taylor series.

$$\left.\frac{d^k}{dx^k}\frac{P_n(x)}{Q_m(x)}\right|_{x=0} = \frac{1}{(k+2)!}$$

The symbolic computation of the exact rational coefficients of the $[13/14]_f(x)$ Padé approximation was carried out in Julia programming language [4] using big integer mathematics, and the Polynomials package. The results of the computation are displayed in table 3.1. The order of the Padé approximation was chosen so that the largest denominator in the rational coefficients of the numerator polynomial (7244400176133120000), and the largest denominator in the rational coefficients of the denominator polynomial (6761440164390912000) were the largest integers less than the largest 64 bit signed integer (9223372036854775807). This choice of approximation will need further research to better optimize.

To make use of the Padé approximation we need to be able to compute powers of $\mathrm{ad}_X\cdot$. This can be accomplished through the Kronecker representation of $\mathrm{ad}_X\cdot$, which requires representing the matrices of $\mathfrak{sl}(\hat{\mathbb{1}})$ as vectors. The vector representation of a matrix is achieved by the matrix reshaping operator $\mathrm{vec}(Y) = \vec{y}$, which forms a vector $\vec{y}$ by concatenation of the columns of Y, called the vectorization of the matrix. We denote the inverse operator to vectorization $\mathrm{mat}(\vec{y}) = \mathrm{vec}^{-1}(\vec{y}) = Y$, which reshapes a vector, of n^2 entries, into an $n \times n$ matrix.

After juggling the indexes of $\mathrm{vec}\left(\frac{\partial X}{\partial x}\right)$, the Kronecker representation of $\mathrm{ad}_X \frac{\partial X}{\partial x}$ follows as

$$\mathrm{ad}_X \frac{\partial X}{\partial x} = \mathrm{mat}\left(\left(I \otimes X - X^\dagger \otimes I\right) \mathrm{vec}\left(\frac{\partial X}{\partial x}\right)\right)$$

Proceeding by induction we find that

$$\mathrm{ad}_X^n \frac{\partial X}{\partial x} = \mathrm{mat}\left(\left(I \otimes X - X^\dagger \otimes I\right)^n \mathrm{vec}\left(\frac{\partial X}{\partial x}\right)\right)$$

It follows that $\frac{\partial e^X}{\partial x}$ can be computed by

$$\frac{\partial e^X}{\partial x} = \mathrm{mat}\left(\sum_{n=0}^{\infty} \frac{1}{(n+2)!} \left(I \otimes X - X^\dagger \otimes I\right)^{n+1} \mathrm{vec}\left(\frac{\partial X}{\partial x}\right)\right) e^X + \frac{\partial X}{\partial x} e^X$$

And thus can be approximated by

$$\frac{\partial e^X}{\partial x} \approx \mathrm{mat}\left(\frac{P_n\left(I \otimes X - X^\dagger \otimes I\right)}{Q_m\left(I \otimes X - X^\dagger \otimes I\right)} \left(I \otimes X - X^\dagger \otimes I\right) \mathrm{vec}\left(\frac{\partial X}{\partial x}\right)\right) e^X + \frac{\partial X}{\partial x} e^X$$

We can summarize this work in an algorithm to compute the non-commutative perturbation 1. This algorithm servers as a sketch only, highlighting only the novel elements developed in this section.

Many additional optimizations could be implemented including minimizing memory assignments by carrying out in place computations, conditioning matrices to improve numerical stability, and using recursive squaring and summing methods to efficiently compute the matrix polynomials. A point of concern is the first assignment of the Kronecker product, which results in a quadratic increase in the amount of memory used. Finally, an algorithm to compute the gradient of the matrix exponential can be formulated 2.

3.2 The Hessian

Assuming X parameterized by $x, y \in \mathbb{R}$ is analytic, or at least twice differentiable the Hessian of e^X exists and depends on three tangent matrices, $\frac{\partial X}{\partial x}$, $\frac{\partial X}{\partial y}$, and $\frac{\partial^2 X}{\partial x \partial y}$. In general none of these tangent matrices need commute with each other. Even the second derivatives $\frac{\partial^2 X}{\partial x^2}$, will not generally commute with the first derivate $\frac{\partial X}{\partial x}$. While the additional terms complicate the computations, they do not lead to intractable results in the same way that finding a closed form for e^X in dimensions greater than 4 becomes intractable. Unfortunately in the most general case when X neither commutes with $\frac{\partial X}{\partial x}$, nor $\frac{\partial X}{\partial y}$ we will find that we have to compute the Taylor expansion of a bilinear form, which is not susceptible to Padé approximation.

To proceed we need a pair of results focused on the binomial combinatorics of adjoints. We will use these results in developing the bilinear non-commutative perturbation of the Hessian.

Lemma 7. *For differentiable matrix function X parameterized by $x \in \mathbb{R}$, any matrices A, B, and integer $n \geq 0$*

$$\left[\frac{\partial}{\partial x} \mathrm{ad}_X^n \cdot\right](A) = \sum_{k=1}^{n} \left(\mathrm{ad}_X^{k-1} A\right) \left(\mathrm{ad}_{\frac{\partial X}{\partial x}} A\right) \left(\mathrm{ad}_X^{n-k} A\right)$$
$$\mathrm{ad}_X^n AB = \sum_{k=0}^{n} \binom{n}{k} \left(\mathrm{ad}_X^k A\right) \left(\mathrm{ad}_X^{n-k} B\right)$$
$$\mathrm{ad}_X^n [A, B] = \sum_{k=0}^{n} \binom{n}{k} \left[\mathrm{ad}_X^k A, \mathrm{ad}_X^{n-k} B\right]$$

Proof: For each of the equalities we have:

1. Proceed by induction on n.

2. Proceed by induction on n.

3. Take the antisymmetric difference of previous equality. □

We will also find use of the following corollary to the binomial theorem.

Corollary 10. *For any $n, m \geq 0$*

$$\sum_{k=m}^{n+m} \binom{k}{m} = \binom{n+m+1}{n}$$

Proof: Proceed by induction on n. □

In the next step of developing the Hessian we derive the Taylor series for the bilinear non-commutative perturbation. Unfortunately because this form is a bilinear map it does not admit the formulation of Padé approximation.

Corollary 11. *For any differentiable matrix function X parameterized by $x, y \in \mathbb{R}$*

$$\begin{aligned}
&\left[\frac{\partial}{\partial x}\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n \cdot\right]\left(\frac{\partial X}{\partial y}\right)+\left[\frac{\partial}{\partial y}\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n \cdot\right]\left(\frac{\partial Y}{\partial x}\right)\\
&\quad=\sum_{n\geq m\geq 0}\frac{1}{(n+2)!}\left(\binom{n+1}{m+1}-\binom{n+1}{m}\right)\left[\operatorname{ad}_X^m\frac{\partial X}{\partial x},\operatorname{ad}_X^{n-m}\frac{\partial X}{\partial y}\right]
\end{aligned}$$

Proof: Apply the previous results in the order they were stated to the symmetric sum of the two terms.

$$\begin{aligned}
&\left[\frac{\partial}{\partial x}\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n \cdot\right]\left(\frac{\partial X}{\partial y}\right)+\left[\frac{\partial}{\partial y}\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n \cdot\right]\left(\frac{\partial Y}{\partial x}\right)\\
&\quad=\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\sum_{k=1}^{n}\operatorname{ad}_X^{k-1}\operatorname{ad}_{\frac{\partial X}{\partial x}}\operatorname{ad}_X^{n-k}\frac{\partial X}{\partial y}\\
&\qquad+\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\sum_{k=1}^{n}\operatorname{ad}_X^{k-1}\operatorname{ad}_{\frac{\partial X}{\partial y}}\operatorname{ad}_X^{n-k}\frac{\partial X}{\partial x}\\
&\quad=\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\sum_{k=1}^{n}\operatorname{ad}_X^{k-1}\left[\frac{\partial X}{\partial x},\operatorname{ad}_X^{n-k}\frac{\partial X}{\partial y}\right]\\
&\qquad+\sum_{n=1}^{\infty}\frac{1}{(n+1)!}\sum_{k=1}^{n}\operatorname{ad}_X^{k-1}\left[\frac{\partial X}{\partial y},\operatorname{ad}_X^{n-k}\frac{\partial X}{\partial x}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{\infty} \frac{1}{(n+1)!} \sum_{k=1}^{n} \sum_{m=0}^{k-1} \binom{k-1}{m} \left[\mathrm{ad}_X^m \frac{\partial X}{\partial x}, \mathrm{ad}_X^{n-m-1} \frac{\partial X}{\partial y}\right] \\
&\quad + \sum_{n=1}^{\infty} \frac{1}{(n+1)!} \sum_{k=1}^{n} \sum_{m=0}^{k-1} \binom{k-1}{m} \left[\mathrm{ad}_X^m \frac{\partial X}{\partial y}, \mathrm{ad}_X^{n-m-1} \frac{\partial X}{\partial x}\right] \\
&= \sum_{n=0}^{\infty} \frac{1}{(n+2)!} \sum_{k=0}^{n} \sum_{m=0}^{k} \binom{k}{m} \left[\mathrm{ad}_X^m \frac{\partial X}{\partial x}, \mathrm{ad}_X^{n-m} \frac{\partial X}{\partial y}\right] \\
&\quad + \sum_{n=0}^{\infty} \frac{1}{(n+2)!} \sum_{k=0}^{n} \sum_{m=0}^{k} \binom{k}{m} \left[\mathrm{ad}_X^m \frac{\partial X}{\partial y}, \mathrm{ad}_X^{n-m} \frac{\partial X}{\partial x}\right] \\
&= \sum_{n,m\geq 0}^{\infty} \frac{1}{(n+m+2)!} \left[\mathrm{ad}_X^m \frac{\partial X}{\partial x}, \mathrm{ad}_X^n \frac{\partial X}{\partial y}\right] \sum_{k=m}^{n+m} \binom{k}{m} \\
&\quad + \sum_{n,m\geq 0}^{\infty} \frac{1}{(n+m+2)!} \left[\mathrm{ad}_X^n \frac{\partial X}{\partial y}, \mathrm{ad}_X^m \frac{\partial X}{\partial x}\right] \sum_{k=n}^{n+m} \binom{k}{n} \\
&= \sum_{n,m\geq 0}^{\infty} \frac{1}{(n+m+2)!} \left(\binom{n+m+1}{n} - \binom{n+m+1}{m}\right) \left[\mathrm{ad}_X^m \frac{\partial X}{\partial x}, \mathrm{ad}_X^n \frac{\partial X}{\partial y}\right] \\
&= \sum_{n\geq m\geq 0} \frac{1}{(n+2)!} \left(\binom{n+1}{m+1} - \binom{n+1}{m}\right) \left[\mathrm{ad}_X^m \frac{\partial X}{\partial x}, \mathrm{ad}_X^{n-m} \frac{\partial X}{\partial y}\right] \\
&= \sum_{n=0}^{\infty} \frac{1}{(n+2)!} F_n
\end{aligned}
$$

Where we have defined F_n recursively as:

$$
\begin{aligned}
F_0 &= 0 \\
F_{n+1} &= \left[\frac{\partial X}{\partial x}, \mathrm{ad}_X^{n+1} \frac{\partial X}{\partial y}\right] + \mathrm{ad}_X F_n - \left[\mathrm{ad}_X^{n+1} \frac{\partial X}{\partial x}, \frac{\partial X}{\partial y}\right]
\end{aligned}
$$

This recursive calculation is illustrate in figure 3.1. The proof of which is found by carrying out induction on n. □

There are three special cases that simplify the calculation of Taylor series considerably.

$$
\sum_{n=0}^{\infty} \frac{F_n}{(n+2)!} = \begin{cases} 0 & \mathrm{ad}_X \frac{\partial X}{\partial x} = 0 \text{ and } \mathrm{ad}_X \frac{\partial X}{\partial y} = 0 \\ \left[\frac{\partial X}{\partial x}, \sum_{n=0}^{\infty} \frac{n}{(n+2)!} \mathrm{ad}_X^n \frac{\partial X}{\partial y}\right] & \mathrm{ad}_X \frac{\partial X}{\partial x} = 0 \\ \left[\frac{\partial X}{\partial y}, \sum_{n=0}^{\infty} \frac{n}{(n+2)!} \mathrm{ad}_X^n \frac{\partial X}{\partial x}\right] & \mathrm{ad}_X \frac{\partial X}{\partial y} = 0 \end{cases}
$$

The Taylor series in the last two cases does admit a Padé approximation, by the same reasoning as presented in the previous section. In particular we can further reduce the Taylor series.

$$\begin{aligned}\sum_{n=0}^{\infty}\frac{n}{(n+2)!}x^n &= \sum_{n=1}^{\infty}\frac{n}{(n+2)!}x^n \\ &= \left(\sum_{n=0}^{\infty}\frac{n+1}{(n+3)!}x^n\right)x\end{aligned}$$

Using the same criteria as in the previous section yields a $[12/14]_f(x)$ Padé approximation for $f(x)=\sum_{n=0}^{\infty}\frac{n+1}{(n+3)!}x^n$. The coefficients of the Padé approximation are summarized in table 3.2. This approximation is then used in a pair of branches in the calculation of the bilinear non-commutative perturbation to the Hessian of the matrix exponential 3.

Assuming that $\frac{\partial^2 X}{\partial x\partial y},\frac{\partial^2 X}{\partial y\partial x}$ are continuous we have, by corollary to Clairaut's theorem, that $\frac{\partial^2 e^X}{\partial x\partial y}=\frac{\partial^2 e^X}{\partial y\partial x}$. We can then compute the Hessian by symmetrizing the partial differential so that antisymmetric terms cancel out.

$$\begin{aligned}\frac{1}{2}\left(\frac{\partial^2 e^X}{\partial x\partial y}+\frac{\partial^2 e^X}{\partial y\partial x}\right) &= \frac{1}{2}\frac{\partial}{\partial x}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial y}\right)e^X \\ &\quad+\frac{1}{2}\frac{\partial}{\partial y}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)e^X \\ &= \frac{1}{2}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial y}\right)\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)e^X \\ &\quad+\frac{1}{2}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial^2 X}{\partial x\partial y}\right)e^X \\ &\quad+\frac{1}{2}\left[\frac{\partial}{\partial x}\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial y}\right)e^X \\ &\quad+\frac{1}{2}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial y}\right)e^X \\ &\quad+\frac{1}{2}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial^2 X}{\partial y\partial x}\right)e^X \\ &\quad+\frac{1}{2}\left[\frac{\partial}{\partial y}\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)e^X \\ &= \left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial^2 X}{\partial x\partial y}\right)e^X \\ &\quad+\frac{1}{2}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial y}\right)\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\operatorname{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)e^X\end{aligned}$$

$$
+\frac{1}{2}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\,\mathrm{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial x}\right)\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\,\mathrm{ad}_X^n\cdot\right]\left(\frac{\partial X}{\partial y}\right)e^X
$$

$$
+\frac{1}{2}\sum_{n\geq m\geq 0}\frac{1}{(n+2)!}\left(\binom{n+1}{m+1}-\binom{n+1}{m}\right)\left[\mathrm{ad}_X^m\frac{\partial X}{\partial x},\mathrm{ad}_X^{n-m}\frac{\partial X}{\partial y}\right]
$$

To flush out the final algorithm for the Hessian lets consider each summand in the last equality individually. The first term involving $\frac{\partial^2 X}{\partial x \partial y}$ can be calculated using the non-commutative perturbation algorithm developed in the preceding section

$$
\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}\,\mathrm{ad}_X^n\cdot\right]\left(\frac{\partial^2 X}{\partial x\partial y}\right)=\left[\sum_{n=0}^{\infty}\frac{1}{(n+2)!}\,\mathrm{ad}_X^n\cdot\right]\left(\mathrm{ad}_X\frac{\partial^2 X}{\partial x\partial y}\right)+\frac{\partial^2 X}{\partial x\partial y}
$$

The next two summands together are the Poisson bracket of the non-commutative perturbations the gradients $\frac{\partial e^X}{\partial x}$, and $\frac{\partial e^X}{\partial y}$. The final summand is the bilinear non-commutative perturbation. Taken together the algorithm for the Hessian is then a sequence of calls to the non-commutative perturbation and the bilinear non-commutative perturbation; as outlined in algorithm 4.

As in the previous section the algorithm we have developed for the Hessian of the matrix exponential is merely a starting point for further optimizations. In particular the bilinear non-commutative perturbation needs attention to see if the Taylor series i susceptible to further efficiency gains. As well the Padé approximation needs further refinement. Nevertheless both of the algorithms for the non-commutative perturbation and the bilinear non-commutative perturbation are stable with respect to the the stochastic contraction Lie algebra $\mathfrak{st}^+(\hat{\mathbb{1}})$; in the sense that if $X, \frac{\partial X}{\partial x}, \frac{\partial X}{\partial y}, \frac{\partial^2 X}{\partial x \partial y} \in \mathfrak{st}^+(\hat{\mathbb{1}})$, then the result of the algorithms will be in $\mathfrak{st}^+(\hat{\mathbb{1}})$.

3.3 Figures and Illustrations

Table 3.1: Padé Approximation of $\sum_{n=0}^{\infty} \frac{1}{(n+2)!} x^n$

Degree	Numerator	Denominator
x^0	$\frac{1}{2}$	$\frac{1}{1}$
x^1	$\frac{-13}{174}$	$\frac{-14}{29}$
x^2	$\frac{1}{58}$	$\frac{13}{116}$
x^3	$\frac{-11}{7830}$	$\frac{-13}{783}$
x^4	$\frac{11}{75168}$	$\frac{11}{6264}$
x^5	$\frac{-11}{1461600}$	$\frac{-11}{78300}$
x^6	$\frac{1}{2192400}$	$\frac{11}{1252800}$
x^7	$\frac{-1}{64832400}$	$\frac{-11}{25212600}$
x^8	$\frac{1}{1728864000}$	$\frac{1}{57628800}$
x^9	$\frac{-1}{79873516800}$	$\frac{-1}{1815307200}$
x^{10}	$\frac{1}{3594308256000}$	$\frac{1}{72612288000}$
x^{11}	$\frac{-1}{295931379744000}$	$\frac{-1}{3793992048000}$
x^{12}	$\frac{1}{28409412455424000}$	$\frac{1}{273167427456000}$
x^{13}	$\frac{-1}{7244400176133120000}$	$\frac{-1}{30185000733888000}$
x^{14}		$\frac{1}{6761440164390912000}$

The exact rational coefficients of the numerator and denominator polynomials of the $[13/14]_f(x)$ Padé approximation of $f(x) = \sum_{n=0}^{\infty} \frac{1}{(n+2)!} x^n$; symbolically computed.

Algorithm 1 Numerical calculation of the non-commutative perturbation of the gradient of the matrix exponential $\frac{\partial e^X}{\partial x}$, using the $[13/14]_f(x)$ Padé approximation of $f(x) = \sum_{n=0}^{\infty} \frac{1}{(n+2)!} x^n$

function PEX(X, $\frac{\partial X}{\partial x}$)
  $A_x \leftarrow \left[X, \frac{\partial X}{\partial x}\right]$ ▷ Allocates memory
  if $A_x = 0$ **then**
    return 0
  end if
  $A_X \leftarrow I \otimes X - X^{\dagger} \otimes I$ ▷ if X is $n \times n$ the result is $n^2 \times n^2$
  $\vec{a}_x \leftarrow \mathrm{VEC}(A_x)$ ▷ Change of indexing
  $P \leftarrow P_{13}(A_X)$ ▷ Padé numerator by recursive summing and squaring
  $Q \leftarrow Q_{14}(A_X)$ ▷ Padé denominator by recursive summing and squaring
  Solve for $\vec{y}$: $P\vec{a}_x = Q\vec{y}$ ▷ Call to linear solver
  return $\mathrm{MAT}(\vec{y})$
end function

Algorithm 2 Numerical calculation of the gradient of the matrix exponential $\frac{\partial e^X}{\partial x}$.

function GEX(X, $\frac{\partial X}{\partial x}$)
  return $\left(\frac{\partial X}{\partial x} + \mathrm{PEX}\left(X, \frac{\partial X}{\partial x}\right)\right)\mathrm{EXP}(X)$
end function

Table 3.2: Padé Approximation of $\sum_{n=0}^{\infty} \frac{n+1}{(n+3)!} x^n$

Degree	Numerator	Denominator
x^0	$\frac{1}{6}$	$\frac{1}{1}$
x^1	$\frac{-237571687770}{2235524846677109}$	$\frac{-2238375706930349}{4471049693354218}$
x^2	$\frac{75899205239671}{22355248466771090}$	$\frac{7}{58}$
x^3	$\frac{-6158055716087}{289724020129353264}$	$\frac{-172676164807703}{9286026286197222}$
x^4	$\frac{14904653897243}{751136348483508624}$	$\frac{989302956681193}{482873366882255544}$
x^5	$\frac{-2677622343899}{225340904545052587200}$	$\frac{-317642104732513}{1857205257239444400}$
x^6	$\frac{15554305889797}{347668824155223991680}$	$\frac{1079293136317583}{96574673376451108800}$
x^7	$\frac{-1086326787569}{44424349753167510048000}$	$\frac{-18091761597171823}{31097044827217257033600}$
x^8	$\frac{23665068673453}{586401416741811132633600}$	$\frac{215879922742633}{8884869950633502009600}$

Continued on next page.

Continued from previous page.

Degree	Numerator	Denominator
x^9	$\frac{-579062082583}{31665676504057801162214400}$	$\frac{-129862571470787}{15992765911403036172800}$
x^{10}	$\frac{1006334079353}{8209619834385355856870400 0}$	$\frac{239909667097237}{1119493613779821253209600 0}$
x^{11}	$\frac{-129504852863}{3649996978367729213964579840 0}$	$\frac{-290350506989849}{6684976150856646912023040 00}$
x^{12}	$\frac{18809879890171}{32849972805309562925681218560000}$	$\frac{539968479224557}{842306995007937510914903040 00}$
x^{13}		$\frac{-359560490231}{5809013758675431109757952000}$
x^{14}		$\frac{77182092353629}{2606097842554558658770710005760 00}$

The exact rational coefficients of the numerator and denominator polynomials of the $[12/14]_f(x)$ Padé approximation of $f(x) = \sum_{n=0}^{\infty} \frac{n+1}{(n+3)!} x^n$; symbolically computed.

$$
\begin{array}{c}
0 \\
1 \quad -1 \\
1 \quad 0 \quad -1 \\
1 \quad 1 \quad -1 \quad -1 \\
1 \quad 2 \quad 0 \quad -2 \quad -1 \\
1 \quad 3 \quad 2 \quad -2 \quad -3 \quad -1 \\
1 \quad 4 \quad 5 \quad 0 \quad -5 \quad -4 \quad -1 \\
1 \quad 5 \quad 9 \quad 5 \quad -5 \quad -9 \quad -5 \quad -1 \\
1 \quad 6 \quad 14 \quad 14 \quad 0 \quad -14 \quad -14 \quad -6 \quad -1 \\
1 \quad 7 \quad 20 \quad 28 \quad 14 \quad -14 \quad -28 \quad -20 \quad -7 \quad -1 \\
\left[\frac{\partial X}{\partial x}, \mathrm{ad}_X^{n+1} \frac{\partial X}{\partial y}\right] \qquad\qquad\qquad\qquad -\left[\mathrm{ad}_X^{n+1} \frac{\partial X}{\partial x}, \frac{\partial X}{\partial y}\right]
\end{array}
$$

Figure 3.1: Illustration of the calculation of the first 8 rows of coefficients of the divergence of Pascal's triangle; emphasizing the coefficients in F_n.

Algorithm 3 Numerical calculation of the bilinear perturbation of the Hessian of the matrix exponential $\frac{\partial^2 e^X}{\partial x \partial y}$, using the $[12/14]_f(x)$ Padé approximation of $f(x) = \sum_{n=0}^{\infty} \frac{n+1}{(n+3)!} x^n$.

function BEX(X, $\frac{\partial X}{\partial x}$, $\frac{\partial X}{\partial y}$, $\varepsilon =$ machine float precision)
  $A_x \leftarrow \left[X, \frac{\partial X}{\partial x}\right]$ ▷ Allocates memory
  $A_y \leftarrow \left[X, \frac{\partial X}{\partial y}\right]$ ▷ Allocates memory
  if $A_x = 0$ **and** $A_y = 0$ **then**
    return 0
  end if
  if $A_x = 0$ **then**
    $A_X \leftarrow I \otimes X - X^\dagger \otimes I$ ▷ if X is $n \times n$ the result is $n^2 \times n^2$
    $\vec{a}_y \leftarrow \mathrm{VEC}(A_y)$ ▷ Change of indexing
    $P \leftarrow P_{12}(A_X)$ ▷ Padé numerator by recursive summing and squaring
    $Q \leftarrow Q_{14}(A_X)$ ▷ Padé denominator by recursive summing and squaring
    Solve for $\vec{z}$: $P\vec{a}_y = Q\vec{z}$ ▷ Call to linear solver
    return $\left[\frac{\partial X}{\partial x}, \mathrm{MAT}(\vec{z})\right]$
  end if
  if $A_y = 0$ **then**
    $A_X \leftarrow I \otimes X - X^\dagger \otimes I$ ▷ if X is $n \times n$ the result is $n^2 \times n^2$
    $\vec{a}_x \leftarrow \mathrm{VEC}(A_x)$ ▷ Change of indexing
    $P \leftarrow P_{12}(A_X)$ ▷ Padé numerator by recursive summing and squaring
    $Q \leftarrow Q_{14}(A_X)$ ▷ Padé denominator by recursive summing and squaring
    Solve for $\vec{z}$: $P\vec{a}_x = Q\vec{z}$ ▷ Call to linear solver
    return $\left[\frac{\partial X}{\partial y}, \mathrm{MAT}(\vec{z})\right]$
  end if
  $R \leftarrow 0$ ▷ Allocates memory
  $F \leftarrow \left[\frac{\partial X}{\partial x}, A_y\right] + \left[\frac{\partial X}{\partial y}, A_x\right]$ ▷ Allocates memory
  $m \leftarrow 3$ ▷ Factorial scalars
  $n \leftarrow 6$ ▷ Factorial scalars
  while $\|F\| > n\varepsilon$ **do**
    $R \leftarrow R + \frac{F}{n}$ ▷ In place computation
    $A_x \leftarrow [X, A_x]$ ▷ In place computation
    $A_y \leftarrow [X, A_y]$ ▷ In place computation
    $F \leftarrow \left[\frac{\partial X}{\partial x}, A_y\right] + [X, F] + \left[\frac{\partial X}{\partial y}, A_x\right]$ ▷ In place computation
    $m \leftarrow m + 1$ ▷ Increment factorial
    $n \leftarrow mn$ ▷ Increment factorial
  end while
  return R
end function

Algorithm 4 Numerical calculation of the Hessian of the matrix exponential $\frac{\partial^2 e^X}{\partial x \partial y}$.

function $\mathrm{HEX}(X, \frac{\partial X}{\partial x}, \frac{\partial X}{\partial y}, \frac{\partial^2 X}{\partial x \partial y})$

$\quad P_x \leftarrow \frac{\partial X}{\partial x} + \mathrm{PEX}\left(X, \frac{\partial X}{\partial x}\right)$ ▷ Call to non-commutative perturbation

$\quad P_y \leftarrow \frac{\partial X}{\partial y} + \mathrm{PEX}\left(X, \frac{\partial X}{\partial y}\right)$ ▷ Call to non-commutative perturbation

$\quad$ **return** $\left(\frac{\partial^2 X}{\partial x \partial y} + \mathrm{PEX}\left(X, \frac{\partial^2 X}{\partial x \partial y}\right) + \frac{1}{2}\,\mathrm{BEX}\left(X, \frac{\partial X}{\partial x}, \frac{\partial X}{\partial y}\right) + \frac{1}{2}\left(P_x P_y + P_y P_x\right)\right)\mathrm{EXP}\left(X\right)$

end function

Chapter 4

Maximum Likelihood Estimation from First Hitting Times

4.1 Distribution of First Hitting Times

Contemporary methods for fitting time homogeneous Markov processes on a finite state space require directly parameterizing the transition probability matrix $\mathbb{P}[X_n = j \| X_0 = i] = \langle \hat{e}_i, P^n \hat{e}_j \rangle$, as they depend on realizing the process through discrete time steps n. While this formulation has many powerful applications, there are analyses where the parameterization of the generator of the time homogeneous Markov process, $\mathbb{P}[X_t = j \| X_0 = i] = \langle \hat{e}_i, \exp(tG)\, \hat{e}_j \rangle$, is of greater meaning, or importance. In particular when the observed process is, at least in principle continuous, or when the desired parameterization is in units of rates per time parameterization of the generator is the more natural choice.

The natural experimental design for continuous time homogeneous Markov process on a finite state space is the observation of a stopped process, such as the statistic of first hitting time of a state, or the first exit time from a state statistic. Surprisingly it is possible to explicitly formulate the distribution of these two statistics in terms of the generator of the process.

To start, recall the definition of the projection operator on a finite dimensional vector space $P_i = \hat{e}_i \otimes \hat{e}_i$; which projects each vector in the space onto a fixed unit vector $\hat{e}_i$. Projection operators hold a special purpose in analyzing continuous time homogeneous Markov processes on a finite state spaces. The projection operator $I - P_k$, when left multiplied to the generator $G = \sum_{ij} x_{ij} C_{ij}$, yields a new process $(I - P_k)\, G$, where state k is an absorbing state.

The project operators P_i are useful in formulating the distribution of first hitting times of a process generated by G in terms of the transition probabilities of a process generated by $(I - P_i)\, G$. We can apply this to restate the first hitting time results in the exercises of Rogers and Williams [17].

Theorem 3. *If $T_j = \inf\{t : X_t = j\}$ is the first hitting time statistic of the transition to j of a process generated by $G = \sum_{ij} x_{ij} C_{ij}$ then*

$$\mathbb{P}_G\left[T_j \leq t \,\|X_0 = i\right] = 1 - \left\langle \hat{e}_i, \exp\left(t\left(I - P_j\right)G\right)\hat{e}_j\right\rangle$$

Proof: The proof hinges on formalizing the intuition that once we know a continuous time homogeneous Markov process on a finite state space X_t has first touched the state j then we need no further information, and so we can work with the simpler process where j is an absorbing state.

$$\begin{aligned}
\mathbb{P}_G\left[T_j \leq t \,\|X_0 = i\right] &= 1 - \mathbb{P}_G\left[T_j > t \,\|X_0 = i\right] \\
&= 1 - \mathbb{P}_G\left[\forall s \leq t \;\; X_s \neq j, \;\; \exists u > t \;\; X_u = j \,\|X_0 = i\right] \\
&= 1 - \mathbb{P}_G\left[\forall s \leq t \;\; X_s \neq j \,\|X_0 = i\right]\mathbb{P}_G\left[\exists u > t \;\; X_u = j \,\|\forall s \leq t \;\; X_s \neq j\right] \\
&= 1 - \mathbb{P}_G\left[\forall s \leq t \;\; X_s \neq j \,\|X_0 = i\right]\mathbb{P}_G\left[X_u = j, u > t \,\|X_t \neq j\right] \\
&= 1 - \mathbb{P}_{\left(I-P_j\right)G}\left[\forall s \leq t \;\; X_s \neq j \,\|X_0 = i\right]\mathbb{P}_{\left(I-P_j\right)G}\left[X_u = j, \;\; u > t \,\|X_t \neq j\right] \\
&= 1 - \mathbb{P}_{\left(I-P_j\right)G}\left[\forall s \leq t \;\; X_s \neq j, \;\; X_u = j, \;\; u > t \,\|X_0 = i\right] \\
&= 1 - \mathbb{P}_{\left(I-P_j\right)G}\left[X_t = j \,\|X_0 = i\right] \\
&= 1 - \left\langle \hat{e}_i, \exp\left(t\left(I - P_j\right)G\right)\hat{e}_j\right\rangle \qquad \square
\end{aligned}$$

This result generalizes in the obvious manner; where if we have a set of first hitting states J then the transition probabilities of the process $(I - P_J)\,G$ gives the cumulative distributions of the hitting times; where $P_J = P_{j_1} + P_{j_2} + \cdots$. This implies that if we can design our experiment to observe as many of the first hitting times as possible we will greatly simplify our statistical estimators.

In light of this, we can reformulate the standard textbook result, for example in Buchholz et. al [6], of first hitting time statistics in the context of the stochastic contraction Lie algebra $\mathfrak{st}^+(\hat{\mathbb{1}})$. As is standard we start with an the experiment designed to observe the first exit time $T_{i \to j}$ from $\hat{e}_i$ to every other state $\hat{e}_j$, where $i \neq j$. Assuming the process is generated by $G = \sum_{ij} x_{ij} C_{ij}$, and keeping $i \neq j$ fixed, the density of the distribution of $T_{i \to j}$ is

$$p_G\left(T_{i \to j} = t \,\|X_0 = i\right) = \frac{d}{dt}\mathbb{P}_G\left[T_j \leq t \,\|X_0 = i\right]$$

$$
\begin{aligned}
&= \frac{d}{dt}\mathbb{P}_{P_iG}\left[T_j \le t \,\|X_0 = i\right] \\
&= \frac{d}{dt}\left(1 - \left\langle \hat{e}_i, \exp\left(tP_iG\right)\hat{e}_j\right)\right\rangle \\
&= \frac{d}{dt}\left(1 - \left\langle \hat{e}_i, \exp\left(t\sum_{l\neq i} x_{il}C_{il}\right)\hat{e}_j\right)\right\rangle \\
&= \frac{d}{dt}\left(1 - \left\langle \hat{e}_i, e^{-tx_{ij}}\hat{e}_i\right)\right\rangle \\
&= \frac{d}{dt}\left(1 - e^{-tx_{ij}}\right) \\
&= x_{ij}e^{-tx_{ij}}
\end{aligned}
$$

Intuitively if we design our experiment to observer the durations $t_n = T_{i\to j}$ between N_{ij} replicated transitions $i \to j$, the maximum likelihood estimate of each rate x_{ij} is then the simple average

$$
\tilde{x}_{ij} = \frac{N_{ij}}{\sum_{n=1}^{N_{ij}} t_n}
$$

However for experiments that involve opportunistic sampling, surveys, or population monitoring it is generally not possible to observe every distinct transition. Typically the initial state of the transition is known or can be inferred, but only a subset of exit states are observed. In this situation the projection operator P_i onto a single dimensional subspace is replaced with a projection $I - P_A = P_{i_1} + P_{i_2} + \cdots$ onto a multidimensional subspace; where P_A is the projection onto the observed absorbing states in set A.

4.2 The Likelihood and Its Maximization

With a method to derive the density in hand we can proceed to formulate the log-likelihood of the first hitting times. To do so we must carefully formulate the experimental design to which the log-likelihood will apply. Rather than attempt to formulate the most general likelihood model possible, which would be notationally laborious given the infinite permutations and combinations of models available, we will illustrate the formulation of the likelihood through a specific application to an aging process.

An aging process is a continuous time homogeneous finite birth death-process, where all the sequential transitions between states are reversible except for transitions to the final state, which is an absorbing state representing death. In the context first hitting time statistics, a finite subset of the states act as sentinel states, where the first hitting time statistic for the transition between any pair of, possible non-adjacent, sentinel states is observed. An example of this process is illustrated in figure 4.1, which displays a seven state aging process, with three sentinel states. The transitions between states that are not sentinel are not directly observed; but rather acts as a type of memory register that broadens the centrality of the distribution of first hitting times.

Given an U state aging process, the generator takes on the simple sequential form:

$$\begin{aligned} G &= \sum_{i=1}^{U-1} x_{i(i+1)} C_{n(i+1)} \\ &= \sum_{i=1}^{U-1} x_{i(i+1)} \left(\hat{e}_i \otimes \hat{e}_{i+1} - \hat{e}_i \otimes \hat{e}_i\right) \end{aligned}$$

The U state aging process as $2U-3$ unknown parameters, $x_{i(i+1)}$ for $1 \le i \le U-1$ and $x_{i(i-1)}$ for $2 \le i \le U-1$, to be fitted by likelihood maximization. Of the U states, a subset of $V \le U$ states are sentinel states, $1 \le i_1 < \cdots i_V \le U$, for which we observe the first hitting time statistics for the transitions between the sentinel states.

Generalizing the theorem in the previous section the generators $G_v^{\pm}$ for the observed first hitting time statistics, $T_{i_v \to i_{v\pm1}}$, are given by

$$\begin{aligned} G_v^{\pm} &= \sum_{j=i_v}^{i_{v\pm1}\mp1} P_j G \\ &= \sum_{j=i_v}^{i_{v\pm1}\mp1} x_{j(j-1)} C_{j(j-1)} + x_{j(j+1)} C_{j(j+1)} \end{aligned}$$

The distribution density of $T_{i_v \to i_{v\pm1}}$ can be concisely stated as

$$p_G\left(T_{i_v \to i_{v\pm1}} = t\right) = \left\langle \hat{e}_{i_v}, G_v^{\pm} \exp\left(t G_v^{\pm}\right) \hat{e}_{i_{v\pm1}} \right\rangle$$

The next step is to formulate the log-likelihood. This requires establishing the observed data. For each of the first hitting time statistics $T_{i_v \to i_{v\pm1}}$ we observe $N_v^{\pm}$ replications of durations $t_v^{\pm(n)}$, enumerated by $1 \le n \le N_v^{\pm}$. Using the convention that summing to $N_v^{\pm}$ means a sum to N_v^{+} and a sum to N_v^{-} the log-likelihood follows as

$$\begin{aligned}
\Lambda &= \sum_{v=1}^{V} \sum_{n=1}^{N_v^{\pm}} \ln p_G \left(T_{i_v \to i_{v\pm1}} = t_v^{\pm(n)} \right) \\
&= \sum_{v=1}^{V} \sum_{n=1}^{N_v^{\pm}} \ln \left\langle \hat{e}_{i_v}, G_v^{\pm} \exp\left(t_v^{\pm(n)} G_v^{\pm} \right) \hat{e}_{i_{v\pm1}} \right\rangle \\
&= \sum_{v=1}^{V} \sum_{n=1}^{N_v^{\pm}} \ln \left\langle x_{i_v(i_v+1)} \hat{e}_{i_v+1} + x_{i_v(i_v-1)} \hat{e}_{i_v-1} - \left(x_{i_v(i_v+1)} + x_{i_v(i_v-1)} \right) \hat{e}_{i_v}, \exp\left(t_v^{\pm(n)} G_v^{\pm} \right) \hat{e}_{i_{v\pm1}} \right\rangle
\end{aligned}$$

Maximization requires differentiation by the $2U-3$ parameters $x_{k(k\pm1)}$. By the linearity of the generator this will result in algebraically replacing the $x_{k(k\pm1)}$ with indicator functions of the form $\mathbb{I}[i_v = k]$, and $\mathbb{I}[k \in i_v, \ldots, i_{v\pm1} \mp 1]$, where k is the index of partial differentiation. To simplify the notation let $\mathbb{I}_v[k] = \mathbb{I}[i_v = k]$, $\mathbb{I}_v^{+}[k] = \mathbb{I}[i_v \le k \le i_{v+1} - 1]$, and $\mathbb{I}_v^{-}[k] = \mathbb{I}[i_{v-1} + 1 \le k \le i_v]$

$$\begin{aligned}
\frac{\partial \Lambda}{\partial x_{k(k\pm1)}} &= \sum_{v=1}^{V} \mathbb{I}_v[k] \sum_{n=1}^{N_v^{\pm}} \frac{\left\langle \hat{e}_{i_v\pm1} - \hat{e}_{i_v}, \exp\left(t_v^{\pm(n)} G_v^{\pm} \right) \right\rangle}{\left\langle \hat{e}_{i_v}, G_v^{\pm} \exp\left(t_v^{\pm(n)} G_v^{\pm} \right) \hat{e}_{i_{v\pm1}} \right\rangle} \\
&\quad + \sum_{v=1}^{V} \mathbb{I}_v^{+}[k] \sum_{n=1}^{N_v^{+}} \frac{\left\langle \hat{e}_{i_v}, G_v^{+} \frac{\partial \exp\left(t_v^{+(n)} G_v^{+} \right)}{\partial x_{k(k\pm1)}} \hat{e}_{i_{v+1}} \right\rangle}{\left\langle \hat{e}_{i_v}, G_v^{+} \exp\left(t_v^{+(n)} G_v^{+} \right) \hat{e}_{i_{v+1}} \right\rangle} \\
&\quad + \sum_{v=1}^{V} \mathbb{I}_v^{-}[k] \sum_{n=1}^{N_v^{-}} \frac{\left\langle \hat{e}_{i_v}, G_v^{-} \frac{\partial \exp\left(t_v^{-(n)} G_v^{-} \right)}{\partial x_{k(k\pm1)}} \hat{e}_{i_{v-1}} \right\rangle}{\left\langle \hat{e}_{i_v}, G_v^{-} \exp\left(t_v^{-(n)} G_v^{-} \right) \hat{e}_{i_{v-1}} \right\rangle} \\
&= \sum_{v=1}^{V} \mathbb{I}_v[k] \sum_{n=1}^{N_v^{\pm}} \frac{\left\langle \hat{e}_{i_v\pm1} - \hat{e}_{i_v}, \exp\left(t_v^{\pm(n)} G_v^{\pm} \right) \right\rangle}{\left\langle \hat{e}_{i_v}, G_v^{\pm} \exp\left(t_v^{\pm(n)} G_v^{\pm} \right) \hat{e}_{i_{v\pm1}} \right\rangle} \\
&\quad + \sum_{v=1}^{V} \mathbb{I}_v^{+}[k] \sum_{n=1}^{N_v^{+}} t_v^{+(n)} \frac{\left\langle \hat{e}_{i_v}, G_v^{+} \,\mathrm{DEX}\left(t_v^{+(n)} G_v^{+}, C_{k(k\pm1)} \right) \hat{e}_{i_{v+1}} \right\rangle}{\left\langle \hat{e}_{i_v}, G_v^{+} \exp\left(t_v^{+(n)} G_v^{+} \right) \hat{e}_{i_{v+1}} \right\rangle} \\
&\quad + \sum_{v=1}^{V} \mathbb{I}_v^{-}[k] \sum_{n=1}^{N_v^{-}} t_v^{-(n)} \frac{\left\langle \hat{e}_{i_v}, G_v^{-} \,\mathrm{DEX}\left(t_v^{-(n)} G_v^{-}, C_{k(k\pm1)} \right) \hat{e}_{i_{v-1}} \right\rangle}{\left\langle \hat{e}_{i_v}, G_v^{-} \exp\left(t_v^{-(n)} G_v^{-} \right) \hat{e}_{i_{v-1}} \right\rangle}
\end{aligned}$$

$$= 0$$

Factoring the summands by both the transition between sentinel states and the observations makes the algorithm to calculate the terms of the partial differential equation more transparent 5. For a given generator G and differential parameter $x_{k(k\pm 1)}$ we first determine the canonical generator $C_{k(k+1)}$ and sub-generator G_v. After that one time computation, the data $t_v^{\pm(n)}$ is looped through to produce the full partial derivative of the log-likelihood6.

While the gradient alone is sufficient for gradient decent searches for local maximum, maximization of the log-likelihood by the Newton-Raphson method requires calculation of the Hessian of the log-likelihood. By necessity the Hessian of the log-likelihood will be complicated. However in this example we have two simplifications at our disposal. First, the generator G is linear in x_{ij} and so the second derivatives vanish. Second, the partial derivative by the transition rates $x_{l(l\pm 1)}$ and $x_{k(k\pm 1)}$ will only be non-trivial when the states fall between the same sentinel states.

$$\frac{\partial^2 \Lambda}{\partial x_{k(k\pm 1)} \partial x_{l(l\pm 1)}} = 0$$

A full implementation of both the Hessian of the log-likelihood and the Newton-Raphson method for likelihood maximization is complex undertaking, but fundamentally is not intractable. For the most part, successful implementation requires patience and diligence on the part of the developer.

4.3 Figures and Illustrations

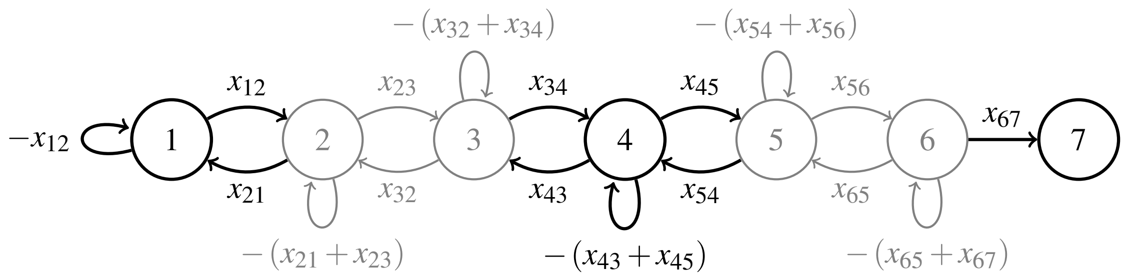

Figure 4.1: A representation of an aging process by a reversible 7 state birth-death process, with 3 sentinel states: healthy (1), care placement (4), and death (7). Each pair of intermediate states represents either a state of improving (2, 5) or worsening (3, 6) health.

Algorithm 5 Numerical calculation of a single summand in the gradient of the log-likelihood $\frac{\partial \Lambda}{\partial x_{k(k\pm 1)}}$. The inner products do not need full evaluation, rather by definition of the basis $\hat{e}_i$ they select entries from the matrix by index.

---

function GLL(G, t, i, j, k, l)
  if $i = j$ **and not** $i \wedge j+1 \leq k \leq i \vee j-1$ **then**
    return 0
  end if
  $E \leftarrow \mathrm{EXP}(tG)$ ▷ Call to matrix exponential
  $C \leftarrow \hat{e}_k \otimes \hat{e}_l - \hat{e}_k \otimes \hat{e}_k$ ▷ Canonical generator
  $D \leftarrow (C + \mathrm{PER}(tG, C))E$ ▷ Call to gradient perturbation
  $x_+ \leftarrow \langle \hat{e}_i, G\hat{e}_{i+1} \rangle$ ▷ Right transition rate
  $x_- \leftarrow \langle \hat{e}_i, G\hat{e}_{i-1} \rangle$ ▷ Left transition rate
  if $k = i$ **then**
    $e_+ \leftarrow \left\langle \hat{e}_{i+1}, E\hat{e}_j \right\rangle$ ▷ Right probability
    $e_0 \leftarrow \left\langle \hat{e}_i, E\hat{e}_j \right\rangle$ ▷ Central probability
    $e_- \leftarrow \left\langle \hat{e}_{i-1}, E\hat{e}_j \right\rangle$ ▷ Left probability
    return $\frac{e_\pm - e_0 + t\left(x_+\left\langle \hat{e}_{i+1}, D\hat{e}_j \right\rangle + x_-\left\langle \hat{e}_{i-1}, D\hat{e}_j \right\rangle - (x_+ + x_-)\left\langle \hat{e}_i, D\hat{e}_j \right\rangle\right)}{x_+e_+ + x_-e_- - (x_+ + x_-)e_0}$
  end if
  return $t\frac{x_+\left\langle \hat{e}_{i+1}, D\hat{e}_j \right\rangle + x_-\left\langle \hat{e}_{i-1}, D\hat{e}_j \right\rangle - (x_+ + x_-)\left\langle \hat{e}_i, D\hat{e}_j \right\rangle}{x_+\left\langle \hat{e}_{i+1}, E\hat{e}_j \right\rangle + x_-\left\langle \hat{e}_{i-1}, E\hat{e}_j \right\rangle - (x_+ + x_-)\left\langle \hat{e}_i, E\hat{e}_j \right\rangle}$
end function

Algorithm 6 Numerical calculation of the gradient of the complete sum log-likelihood $\frac{\partial \Lambda}{\partial x_{k(k\pm 1)}}$. Implemented as a straight forward single instruction, multiple data loop (SIMD).

function GCL(G, $\{(t,i,j),\dots\}$, k, l)
    if $|k-l| \neq 1$ **then**
        return 0
    end if
    $r \leftarrow 0$ ▷ Loop initialization
    for all $t,i,j \in \{(t,i,j),\dots\}$ **do**
        if $i < j$ **then**
            $G_v \leftarrow \left(\sum_{n=i}^{j-1} \hat{e}_n \otimes \hat{e}_n\right) G$ ▷ First hitting time generator
            $r \leftarrow r + \mathrm{GLL}\left(G_v,t,i,j,k,l\right)$ ▷ SIMD computation
        else if $j < i$ **then**
            $G_v \leftarrow \left(\sum_{n=j+1}^{i} \hat{e}_n \otimes \hat{e}_n\right) G$ ▷ First hitting time generator
            $r \leftarrow r + \mathrm{GLL}\left(G_v,t,i,j,k,l\right)$ ▷ SIMD computation
        end if
    end for
    return r
end function

Chapter 5

Conclusion

5.1 Summary of Results

In chapter two we reversed the normal development of stochastic matrices; which usually starts with characterizing matrices as having non-negative entries with fixed row sums in a standard orthonormal basis $\hat{e}_i$. The line of typical development then notices that the vector $\vec{\mathbb{1}} = \sum_{i=1}^{n} \hat{e}_i$ is an Eigenvector. Instead we began by characterizing all invertible matrices A such that $A\hat{\mathbb{1}} = \hat{\mathbb{1}}$ with respect to a fixed unit vector $\hat{\mathbb{1}}$. We showed that there is always a basis $\hat{e}_i$ such that $\vec{\mathbb{1}} = \sqrt{n}\hat{\mathbb{1}}$ can be interpreted as the row sum vector, and found that this allowed us to characterize both the Lie group in which the matrices $A\hat{\mathbb{1}} = \hat{\mathbb{1}}$ reside, and the Lie algebra of tangents to the Lie group. We denoted the $St(\hat{\mathbb{1}})$ stochastic Lie group with respect to $\hat{\mathbb{1}}$, and $\mathfrak{st}(\hat{\mathbb{1}})$ the stochastic Lie algebra with respect to $\hat{\mathbb{1}}$. We further characterized the doubly stochastic Lie group and algebra with respect to $\hat{\mathbb{1}}$, respectively denoted $St(\hat{\mathbb{1}}, \hat{\mathbb{1}})$ and $\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}})$, of invertible matrices of fixed row and column sums with respect to $\hat{\mathbb{1}}$. This accomplished by generalizing the stochastic Lie group and algebra to the dual stochastic Lie group and algebra, $St^{\dagger}(\hat{\mathbb{1}})$ and $\mathfrak{st}^{\dagger}(\hat{\mathbb{1}})$, of invertible matrices such that $A^{\dagger}\hat{\mathbb{1}} = \hat{\mathbb{1}}$.

In chapter three we used analytic and closure properties of the stochastic Lie algebras to argue that the algorithms developed for calculating the gradient and Hessian of the matrix exponential will result in matrices that belong to the stochastic Lie algebra. An initial computation of the coefficients for a Padé approximation was presented, and an sketch of algorithms to calculate the gradient and the Hessian were outlined.

In chapter four we combined the results of the previous two chapters, illustrated through the example of the aging process. The algorithms to calculate the log-likelihood, its gradient, and Hessian were developed. Finally the Newton-Raphson maximization was briefly introduced.

5.2 Discussion

Any attempt to fit the generator of a Markov process to observations is in essence an exercise in calculating the logarithm of a matrix. Like the scalar complex logarithm the matrix logarithm is not unique, and has many branches. To demonstrate this point, in chapter two a calculation of one of the branches of the logarithm of permutations was developed, in the context of the stochastic Lie algebra. However, much as there is a unique real logarithm of a positive real number, working within the normal subgroup, and real the sub-algebra, of the stochastic Lie group does provide for certain guarantees of algebraic and analytic closure. Furthermore any time homogeneous Markov process can never escape the normal subgroup because the identity element is a member of the normal subgroup of matrices with positive determinant.

The fitting of generators of Markov processes to observations is bedeviled by a second source degeneracy beyond branches of the matrix logarithm. As reviewed in the beginning of chapter four, unless the experiment can be designed to observe every possible transition between states there will always be indeterminacy in the model. In particular when in complete number of stopping statistics are used many Markov models will generate the exact same distributions for the stopping statistics. For example one could always add ghost states and transitions between them that are never observed. This is were the intuition of the scientist is of critical importance. Where in they apply Occam's razor to select the most parsimonious explanation for the observations. Even an application of the likelihood ratio test or the Akiake information criteria [1] will not be of assistance as the extra parameters washout when the distribution of specific stopped statistics are formulated. A full resolution of this indeterminacy waits a comprehensive classification of the generators of continuous time homogeneous Markov processes on discrete state spaces by the distributions of their stopped statistics. This undertaking should be well within the reach of contemporary techniques.

This should be of no discouragement however as there is much fertile ground that still needs covering. The Padé approximation presented were only the first pass are derivation, and much work remains to be done in numerically optimizing the choice of approximation orders for various implementations are architectures. Furthermore the calculation of the adjoint through the Kronecker products will present a serious bottleneck to scaling the current algorithms up to the terabyte and pentabyte scales data of numerical fitting. As well the branch of the bilinear non-commutative perturbation that calculates using a recursive Taylor series approximation needs deep investigation to determine if there are additional factorizations that can numerically stabilize the loop and speed up the computation. On top of that the algorithms presented are merely sketches for the actual implementation, which when done will need carefully consideration of memory management, assignment, and logic branching. Finally the propositions in chapter two deserve to be give proper treatment and have the proofs of the claims completed.

Bibliography

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control **19** (1974), no. 6, 716–723.

[2] A. Al-Mohy and N. Higham, *Computing the Fréchet Derivative of the Matrix Exponential, with an Application to Condition Number Estimation*, SIAM Journal on Matrix Analysis and Applications **30** (2009), no. 4, 1639–1657.

[3] ______, *A New Scaling and Squaring Algorithm for the Matrix Exponential*, SIAM Journal on Matrix Analysis and Applications **31** (2009), no. 3, 970–989.

[4] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah, *Julia: A Fresh Approach to Numerical Computing*, arXiv:1411.1607 [cs] (2014), arXiv: 1411.1607.

[5] Patrick Billingsley, *Probability and Measure*, 3 edition ed., Wiley-Interscience, New York, May 1995 (English).

[6] Peter Buchholz, Jan Kriege, and Iryna Felko, *Input Modeling with Phase-Type Distributions and Markov Models*, SpringerBriefs in Mathematics, Springer International Publishing, Cham, 2014.

[7] D. Chruściński, V. I. Man'ko, G. Marmo, and F. Ventriglia, *On pseudo-stochastic matrices and pseudo-positive maps*, Physica Scripta **90** (2015), no. 11, 115202, arXiv: 1504.05221.

[8] Jesús Fernández-Sánchez, Jeremy G. Sumner, Peter D. Jarvis, and Michael D. Woodhams, *Lie Markov models with purine/pyrimidine symmetry*, arXiv:1206.1401 [math, q-bio, stat] (2012), arXiv: 1206.1401.

[9] Brian Hall, *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, 1st ed. 2003. corr. 2nd printing 2004 edition ed., Springer, New York, August 2004 (English).

[10] N. Higham, *The Scaling and Squaring Method for the Matrix Exponential Revisited*, SIAM Journal on Matrix Analysis and Applications **26** (2005), no. 4, 1179–1193.

[11] Nicholas J. Higham, *Functions of Matrices: Theory and Computation*, Society for Industrial & Applied Mathematics,U.S., Philadelphia, March 2008 (English).

[12] Joseph E. Johnson, *Markov-type Lie groups in GL(n,R)*, Journal of Mathematical Physics **26** (1985), no. 2, 252–257.

[13] Leslie Lamport, *How to write a 21st century proof*, Journal of Fixed Point Theory and Applications **11** (2012), no. 1, 43–63 (en).

[14] C. Moler and C. Van Loan, *Nineteen Dubious Ways to Compute the Exponential of a Matrix*, SIAM Review **20** (1978), no. 4, 801–836.

[15] ———, *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*, SIAM Review **45** (2003), no. 1, 3–49.

[16] Bassam Mourad, *On a Lie-theoretic approach to generalized doubly stochastic matrices and applications*, Linear and Multilinear Algebra **52** (2004), no. 2, 99–113.

[17] L. C. G. Rogers and David Williams, *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*, 2 edition ed., Cambridge University Press, Cambridge, U.K. ; New York, May 2000 (English).

[18] ———, *Diffusions, Markov Processes and Martingales: Volume 2, Itô Calculus*, 2 edition ed., Cambridge University Press, Cambridge, U.K. ; New York, September 2000 (English).

[19] J. G. Sumner, J. Fernández-Sánchez, and P. D. Jarvis, *Lie Markov models*, Journal of Theoretical Biology **298** (2012), 16–31.

Appendix A

Julia Implementations

Basic implementations of the presented algorithms.

```
# The polynomials package provides both the polynomial type and
# the Pade approximation algorithm. If it is not installed
# uncomment the next line
# Pkg.add('Polynomials')
using('Polynomials')

# General function to compute the [n/m] approximation of
# 1 / (a k + b)!
f(n,m,a,b) = Pade
(
  Poly
  (
    [Rational(1,factorial(BigInt(a*k+b))) for k=0:(n+m+1)]
  ),
  n,
  m
)

# Compute the desired coefficients
c = f(16,16,1,2)

# Print the numerator coefficients
c.p.a

# Print the denominator coefficients
c.q.a
```

Listing A.1: Poor man's symbolic computation of the Pade coefficients of the generalized hyperbolic functions in Julia