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

The probability theory of stochastic processes and Lie theory have individually 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 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
$\vec{a}, \vec{b}, \vec{c}, \ldots$	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 $\hat{\mathbb{1}}$
$St^{T}(\hat{\mathbb{1}})$	Dual stochastic Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}^{T}(\hat{\mathbb{1}})$	Dual stochastic Lie algebra with respect to $\hat{\mathbb{1}}$
$St^{+T}(\hat{\mathbb{1}})$	Dual stochastic contraction Lie group with respect to $\hat{\mathbb{1}}$
$\mathfrak{st}^{+T}(\hat{\mathbb{1}})$	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 $s \leq t$
$\mathbb{V}ar[X_t]$	Variance of X_t
$\mathbb{V}ar[X_t \parallel \mathscr{F}_s]$	Variance conditioned on $\mathscr{F}_s$, given $s \leq t$
$\mathbb{C}ov[X_t, Y_t]$	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 $s \leq t$
$\mathbb{C}or[X_t, Y_t]$	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 $s \leq t$
$\mathrm{Ad}_X A$	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 $\sqrt{-1}$
$a \equiv b \bmod c$	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[3, 11, 12]. 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. Simplifying assumptions include discretization of time and state spaces, homogeneity of the process, and restrictions of the allowed transitions; resulting 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, with 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 can only 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; especially with regard to computing the condition number of numerical problems, as a measure of convergence and stability of the numerical solution[1].

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; from 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 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[8].

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 Lie algebra of the generators, 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. Chapter two has a secondary role to help develop the physical intuition for the stochastic Lie group necessary to work through the derivatives and approximations of chapter three. The third chapter derives the the first and second order derivatives of the exponential map and their Padé Approximation. The fourth chapter derives the maximum likelihood estimators from first hitting times, and the Newton-Raphson computation. The fifth chapter concludes with summarizing remarks and a discussion of the direction for further investigation.

1.3 Background

1.3.1 Continuous Time Homogeneous Markov Process

1.3.2 Maximum Likelihood Estimation

1.3.3 Lie Theory

1.3.4 Padé Approximation

1.3.5 Newton-Raphson Method

Chapter 2

The Lie Algebra of Markov Processes Generators

2.1 Stochastic Matrices

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 $\hat{\mathbb{1}}$. 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[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[4].

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 $GL(n)$. $\square$

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 begin 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

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

$$= \exp(tA) A \hat{\mathbb{1}} \big|_{t=0}$$
$$= A \hat{\mathbb{1}} \qquad \square$$

The following chapter will hinge on taking the derivatives of 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

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

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 parameterized in products. However the preceding corollary was presented in its form to illustrate the intuition that any increase in a particular element has to be compensated for by an equal decrease in some other elements.

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 $\hat{e}_{n+1}$ if 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 $i \leq n$

$$\frac{1}{\sqrt{n}} = \left\langle \hat{e}_i, \hat{\mathbb{1}}_n \right\rangle$$

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

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 $C_{ij} = \hat{e}_i \otimes \left(\hat{e}_j - \hat{e}_i\right)$ 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 $i = k$

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

$$= -\hat{e}_i \otimes (\hat{e}_l - \hat{e}_i)$$

$$= -C_{il}$$

3. When $j = k$ we have

$$C_{ij}C_{jl} = \left(\delta_{jj} - \delta_{ij}\right) \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_j\right)$$

$$= \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_j\right)$$

$$= \hat{e}_i \otimes \left(\hat{e}_l - \hat{e}_i + \hat{e}_i - \hat{e}_j\right)$$

$$= C_{il} - C_{ij}$$

4. Finally when none of the previous conditions apply

$$C_{ij}C_{kl} = \left(\delta_{jk} - \delta_{ik}\right) \hat{e}_i \otimes (\hat{e}_l - \hat{e}_k)$$

$$= 0 \cdot \hat{e}_i \otimes (\hat{e}_l - \hat{e}_k)$$

$$= 0$$ □

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.

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

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

1. Starting with $i = k$

$$\left[C_{ij}, C_{il}\right] = C_{ij}C_{il} - C_{il}C_{ij}$$

$$= C_{ij} - C_{il}$$

2. For $i = l$

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

3. For $j = k$

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

4. When none of the conditions apply

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

□

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

$$C_{ij}\hat{\mathbb{1}} = \hat{e}_i \otimes \left(\hat{e}_j - \hat{e}_i\right)\hat{\mathbb{1}}$$

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

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 uniquely defined only up to rotations orthogonal to the vector $\hat{\mathbb{1}}$. Nevertheless we can immediately see by Jacobi's formula that when $G = \sum_{i,j} \alpha_{ij} C_{ij}$ then $\det G = \exp\left(\sum_{i,j} \alpha_{ij}\right)$.

Geometrically fixing a basis $\hat{e}_i$ defines a convex polytope in the form of a unit hypercube $[0,1]^{n^2-n}$ in the parameter space isomorphic to $\mathfrak{st}(\hat{\mathbb{1}})$. For specificity, we refer to this as the convex polytope of matrices stochastic with respect to basis $\hat{e}_i$. Analytically $St(\hat{\mathbb{1}})$ is open, yet the boundary of the polytope is always a limit point of $St(\hat{\mathbb{1}})$. In particular every vertex of the polytope is a limit point of $St(\hat{\mathbb{1}})$. Probabilistically, the constructed basis elements $\hat{e}_i$ are a well defined enumeration of the states of a continuous time homogeneous Markov process on a finite state space. For example, we can Lie theoretically reformulate the classic two state continuous Markov process.

Corollary 3. $\exp\left(\alpha C_{ij} + \beta C_{ji}\right) = I + \frac{1-e^{-\alpha-\beta}}{\alpha+\beta}\left(\alpha C_{ij} + \beta C_{ji}\right)$

Proof: We first work out the powers of $\alpha C_{ij} + \beta C_{ji}$, and then apply the calculation to the definition of the matrix exponential.

1. Consider the square of $\alpha C_{ij} + \beta C_{ji}$

$$
\begin{aligned}
\left(\alpha C_{ij} + \beta C_{ji}\right)^2 &= \alpha^2 C_{ij}^2 + \alpha\beta C_{ij} C_{ji} + \alpha\beta C_{ji} C_{ij} + \beta^2 C_{ji}^2 \\
&= -\left(\alpha + \beta\right)\left(\alpha C_{ij} + \beta C_{ji}\right)
\end{aligned}
$$

2. It immediately follows that $\left(\alpha C_{ij}+\beta C_{ji}\right)^n = (-\alpha-\beta)^{n-1}\left(\alpha C_{ij}+\beta C_{ji}\right)$ and thus the exponential is

$$\begin{aligned}\exp\left(\alpha C_{ij}+\beta C_{ji}\right) &= I+\sum_{n=1}^{\infty}\frac{1}{n!}(-\alpha-\beta)^{n-1}\left(\alpha C_{ij}+\beta C_{ji}\right)\\ &= I+\frac{1-e^{-\alpha-\beta}}{\alpha+\beta}\left(\alpha C_{ij}+\beta C_{ji}\right)\end{aligned} \qquad \square$$

This last corollary admits a intuitive heuristic interpretation: that each canonical generator C_{ij} can be thought of as measuring the infinitesimal transition rate, or flow of probability, from the state represented by basis element $\hat{e}_i$ to the state represented by the basis element $\hat{e}_j$. This can be seen by considering the matrix representation of $\exp\left(\alpha C_{ij}\right)$ in the basis spanned by $\hat{e}_i$ and $\hat{e}_j$.

$$I+\left(1-e^{-\alpha}\right)C_{ij} = \begin{pmatrix} e^{-\alpha} & 1-e^{-\alpha} \\ 0 & 1 \end{pmatrix}$$

Taking the limit as $\alpha \to \infty$ shows while the limits in the positive directions in the tangent space $\mathfrak{st}(\hat{\mathbb{1}})$ maybe finite, $St(\hat{\mathbb{1}})$ is not closed.

$$\begin{aligned}\lim_{\alpha\to\infty}\exp\left(\alpha C_{ij}\right) &= I+C_{ij}\\ &= \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\end{aligned}$$

The interpretation of C_{ij} as measuring a real flow is more nuanced than may first appear, as $e^{-\alpha} > 0$ for all $\alpha \in \mathbb{R}$. To realize $e^{-\alpha} < 0$ requires that $\mathbb{I}m(\alpha) \equiv \pi \bmod 2\pi$. The importance of imaginary numbers becomes clear when considering the exponential of the generator $\alpha\left(C_{ij}+C_{ji}\right)$.

$$\exp\left(\alpha\left(C_{ij}+C_{ji}\right)\right) = I+\frac{1-e^{-2\alpha}}{2}\left(C_{ij}+C_{ji}\right)$$

Taking the positive limit of α yields an open limit point on the boundary of $St(\hat{\mathbb{1}})$, but still in the the convex polytope of stochastic matrices with respect to basis $\hat{e}_i$.

$$\lim_{\alpha\to\infty} \exp\left(\alpha\left(C_{ij}+C_{ji}\right)\right) = I + \frac{1}{2}\left(C_{ij}+C_{ji}\right)$$
$$= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

However not all matrices inside the convex polytope are realized by positive values of α. For example to find the invertible vertex of the convex polytope on the line α, we need to consider $\alpha = \mathrm{i}\frac{\pi}{2}$.

$$\exp\left(\alpha\left(C_{ij}+C_{ji}\right)\right) = I + \frac{1-e^{-\mathrm{i}\pi}}{2}\left(C_{ij}+C_{ji}\right)$$
$$= I + \left(C_{ij}+C_{ji}\right)$$
$$= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

In the binary basis of $\hat{e}_i$ and $\hat{e}_j$ the vertexes can be summarized as

$$V_0 = \lim_{\substack{\alpha\to 0 \\ \beta\to\infty}} \exp\left(\alpha C_{ij} + \beta C_{ji}\right)$$
$$= I + C_{ji} \notin St(\hat{\mathbb{1}})$$

$$V_1 = \lim_{\substack{\alpha\to \mathrm{i}\frac{\pi}{2} \\ \beta\to \mathrm{i}\frac{\pi}{2}}} \exp\left(\alpha C_{ij} + \beta C_{ji}\right)$$
$$= I + C_{ij} + C_{ji} \in St(\hat{\mathbb{1}})$$

$$V_2 = \lim_{\substack{\alpha\to 0 \\ \beta\to 0}} \exp\left(\alpha C_{ij} + \beta C_{ji}\right)$$
$$= I \in St(\hat{\mathbb{1}})$$

$$V_3 = \lim_{\substack{\alpha\to\infty \\ \beta\to 0}} \exp\left(\alpha C_{ij} + \beta C_{ji}\right)$$
$$= I + C_{ij} \notin St(\hat{\mathbb{1}})$$

A general vertex V of the convex polytope of stochastic $n\times n$ matrices in is given by

$$V = I + \sum_{i=1}^{n} C_{ij(i)}$$

where $j(i) : \{1, \ldots, n\} \mapsto J \subseteq \{1, \ldots, n\}$ is any one of the n^n such functions. Given that the vertexes are linear combinations of C_{ij}, and it may seem plausible that the exponential of a linear combination of C_{ij} is always linear combination of C_{ij}. In fact, this is true in general for any matrix Lie group M, namely it is the Lie algebra offset by the identity element $M = I + \mathfrak{m}$. This can be seen from the Taylor series expansion of the matrix exponential. It follows that for a $G = \sum_{ij} g_{ij} C_{ij} \in \mathfrak{st}(\hat{\mathbb{1}})$ the matrix exponential is $\exp G = I + \sum_{ij} h_{ij} C_{ij}$.

Unfortunately, because polynomials of degree greater than four are generally unsolvable, the relationship between the coefficients g_{ij} and h_{ij} is highly non-trivial in most circumstances. An exception to this difficulty can be found in the formulation of the of first exit times from a fixed initial state. The generator of the first exit from i to any state $j \neq i$ is given by $G_i = \sum_{i \neq j} \alpha_j C_{ij}$, [1] from which we have

$$\begin{aligned}
\exp G_i &= I + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\sum_{i \neq j} \alpha_j \right)^{n-1} \sum_{i \neq j} \alpha_j C_{ij} \\
&= I + \frac{1 - e^{-\sum_{i \neq k} \alpha_k}}{\sum_{i \neq l} \alpha_l} \sum_{i \neq j} \alpha_j C_{ij} \\
&= I + \frac{1 - e^{-\sum_{i \neq k} \alpha_k}}{\sum_{i \neq l} \alpha_l} G_i
\end{aligned}$$

Likewise the generator for the exclusive entrance to a state j, from any state $i \neq j$, is $G_j = \sum_{i \neq j} \alpha_i C_{ij}$, which yields

$$\begin{aligned}
\exp G_j &= I - \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{i \neq j} (-\alpha_i)^n C_{ij} \\
&= I + \sum_{i \neq j} \left(1 - e^{-\alpha_i} \right) C_{ij}
\end{aligned}$$

[1] A complete proof of the result will be given in chapter four.

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

Again this definition necessitates a proof of the claim in the previous paragraph.

Corollary 4. *$St^{+}(\hat{\mathbb{1}})$ 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 $\det A_s(t) = \exp\left(s\sum_{ij} \alpha_{ij}(t)\right) \in \mathbb{R}^{+}$. □

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

Definition 5. 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 5. *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}, whose main argument requires checking the condition of linear sum and scalar multiplication closure:

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

$$\begin{aligned} 0 &\equiv \mathbb{I}m\left(a \sum_{ij} \alpha_{ij}\right) \bmod 2\pi \\ &\equiv \mathbb{R}e\,(a)\,\mathbb{I}m\left(\sum_{ij} \alpha_{ij}\right) + \mathbb{I}m\,(a)\,\mathbb{R}e\left(\sum_{ij} \alpha_{ij}\right) \bmod 2\pi \\ &\equiv \mathbb{I}m\,(a)\,\mathbb{R}e\left(\sum_{ij} \alpha_{ij}\right) \bmod 2\pi \\ &= \mathbb{I}m\,(a) \end{aligned}$$

□

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 $St(\hat{\mathbb{1}})$ defined 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 $A^T\hat{\mathbb{1}} = \hat{\mathbb{1}}$ can be interpreted as the conservation of the column sums of A. The dual definitions for the Lie group and algebra follow natural.

Definition 6. Let $St^T(\hat{\mathbb{1}})$ denote the dual stochastic Lie group of invertible matrices whose transpose is stochastic with respect to $\hat{\mathbb{1}}$.

Definition 7. Let $\mathfrak{st}^T(\hat{\mathbb{1}})$ denote the dual stochastic Lie algebra of $St^T(\hat{\mathbb{1}})$.

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

2.2 Doubly Stochastic Matrices

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 $A^T\hat{\mathbb{1}} = \hat{\mathbb{1}}$ 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 is an important clue in the construction of a canonical representation. In fact the representation 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 8. Let $St(\hat{\mathbb{1}}, \hat{\mathbb{1}})$ denote the doubly stochastic Lie group of invertible matrices A such that both A and A^T are stochastic with respect to $\hat{\mathbb{1}}$

We can immediately observe with out proof that $St(\hat{\mathbb{1}}, \hat{\mathbb{1}}) = St(\hat{\mathbb{1}}) \cap St^T(\hat{\mathbb{1}})$; leading to the next definition.

Definition 9. Let $\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}})$ denote the doubly stochastic Lie algebra of $St(\hat{\mathbb{1}}, \hat{\mathbb{1}})$.

Again, it should be clear that $\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}}) = \mathfrak{st}(\hat{\mathbb{1}}) \cap \mathfrak{st}^T(\hat{\mathbb{1}})$. 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^T. As with $St(\hat{\mathbb{1}})$, the Lie group $St(\hat{\mathbb{1}}, \hat{\mathbb{1}})$ is not connected, and contains the splitting contraction sub-group $St^+(\hat{\mathbb{1}}, \hat{\mathbb{1}})$, with its similarly defined contraction sub-algebra $\mathfrak{st}^+(\hat{\mathbb{1}}, \hat{\mathbb{1}})$. $St^+(\hat{\mathbb{1}}, \hat{\mathbb{1}})$ has the same properties of normality and simple connectedness within $St(\hat{\mathbb{1}}, \hat{\mathbb{1}})$; of course it is not normal within $St(\hat{\mathbb{1}})$.

We can then 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 $\langle \hat{e}_i, \hat{\mathbb{1}} \rangle = \frac{1}{\sqrt{n}}$ 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 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 $C_{ijn}^T = C_{jin}$. 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 relationship, for any $i, j \leq n-1$.

$$C_{ijn} - C_{iin} - C_{jjn} = \hat{e}_i \otimes \hat{e}_j - \hat{e}_i \otimes \hat{e}_n - \hat{e}_n \otimes \hat{e}_j + \hat{e}_n \otimes \hat{e}_n$$

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 5.

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

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

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

$$+\delta_{jk}\left(C_{jjn}+C_{iln}-C_{ijn}-C_{jln}\right)+\delta_{jl}\left(C_{iin}-C_{ijn}\right)$$

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 $j=k$

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

4. When $i=k$

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

5. When $i=l$

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

6. When $j=k$

$$\begin{aligned} C_{ijn}C_{jln} &= C_{jjn}-C_{jjn}-C_{jjn}+C_{jjn}+C_{iln}-C_{ijn}-C_{jln} \\ &= C_{iln}-C_{ijn}-C_{jln} \end{aligned}$$

7. When $j=l$

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

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 6.

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

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 $j = l$

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

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

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

3. When $i = k$

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

4. When $i = l$

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

5. When $j = k$

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

6. When $j = l$

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

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 $\mathfrak{st}(\hat{\mathbb{1}}, \hat{\mathbb{1}})$.$\square$

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.

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[9, 10]. 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[5, 6]. The Higham implementation was further optimized for 64 bit architectures by Al-Mohy[2]. The same work developed an algorithm to approximate the derivative of the matrix exponential was for formulated, essentially 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[1].

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]

$$
\begin{aligned}
\frac{\partial e^X}{\partial x} &\left.\begin{aligned}
&= e^X \left[\int_0^1 e^{-s\,\mathrm{ad}_X \cdot} ds\right] \left(\frac{\partial X}{\partial x}\right) \\
&= e^X \left[\frac{1 - e^{-\mathrm{ad}_X \cdot}}{\mathrm{ad}_X \cdot}\right] \left(\frac{\partial X}{\partial x}\right) \\
&= e^X \left[\sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)!} \mathrm{ad}_X^n \cdot\right] \left(\frac{\partial X}{\partial x}\right)
\end{aligned}\right\} \text{left recursive} \\
&= \left[\frac{\mathrm{ad}_{e^X} \cdot}{\mathrm{ad}_X \cdot}\right] \left(\frac{\partial X}{\partial x}\right) \qquad \text{adjoint ratio} \\
&= e^{\frac{1}{2}X} \left[\frac{e^{\frac{1}{2}\mathrm{ad}_X \cdot} - e^{-\frac{1}{2}\mathrm{ad}_X \cdot}}{\mathrm{ad}_X \cdot}\right] \left(\frac{\partial X}{\partial x}\right) e^{\frac{1}{2}X} \qquad \text{hyperbolic} \\
&\left.\begin{aligned}
&= \left[\int_0^1 e^{s\,\mathrm{ad}_X \cdot} ds\right] \left(\frac{\partial X}{\partial x}\right) e^X \\
&= \left[\frac{e^{\mathrm{ad}_X \cdot} - 1}{\mathrm{ad}_X \cdot}\right] \left(\frac{\partial X}{\partial x}\right) e^X \\
&= \left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!} \mathrm{ad}_X^n \cdot\right] \left(\frac{\partial X}{\partial x}\right) e^X
\end{aligned}\right\} \text{right recursive}
\end{aligned}
$$

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

[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.

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 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 background material, 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.

Blah, blah, blah the next paragraphs are non-sense and needing replacing. Instead show that the adjoint belongs to an algebra, so that solutions Y to rational equations $P(X) = Q(X)Y$ can be found, when they exist, should make it clear $[P(X), Q(X)] = 0$, as part of justifying rational approximation vis-a-vi can be computed using standard linear algebra. After that return to developing how to carry out of the computation using Kronecker representation of adjoint. Need to point out that the adjoint in a linear operator itself so has a representation as matrix.

The operator $\mathrm{ad}_X\cdot$ has tensorial relationships to the underlying vector space on which X is a linear operator. Specifically $\mathrm{ad}_X\cdot$ is a bounded, and thus continuous, linear operator, such that if the underlying vector space is n dimensional then the vector space of linear operators is n^2 dimensional, and $\mathrm{ad}_X\cdot$ by tensor extension can be represented as an $n^2 \times n^2$ matrix.

Computational the powers of $\mathrm{ad}_X\cdot$ can be calculated through the Kronecker representation, which requires representing the matrices 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 vectorization. We denote the inverse operator to vectorization $\mathrm{mat}(\vec{y}) = \mathrm{vec}^{-1}(\vec{y}) = Y$, which reshapes a vector, n^2, into an $n \times n$ matrix.

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

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

Proceeding by induction we find that

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

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

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

3.2 The Hessian

Need preamble on combinatorics of powers of adjoint.

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.

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

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

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

The only terms requiring approximation are precisely the approximation we derived in the previous section. Reusing the components in the algorithm in the previous section we can present an algorithm for the Hessian.

3.3 General Fréchet Derivatives

Chapter 4

Maximum Likelihood Estimation from First Hitting Times

4.1 Distribution of First Hitting Times

4.2 The Likelihood and Its Maximization

4.3 Newton-Raphson Maximization

4.3.1 Formulation

4.3.2 Algorithm

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 tangent to the Lie group. We denoted these the $St(\hat{\mathbb{1}})$ stochastic Lie group with respect to $\hat{\mathbb{1}}$ and $\mathfrak{st}(\hat{\mathbb{1}})$ the stochastic Lie group with respect to $\hat{\mathbb{1}}$. We further characterized the doubly stochastic Lie group and algebra with respect to $\hat{\mathbb{1}}$, 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 $\hat{\mathbb{1}}$, by generalizing the stochastic Lie group and algebra to the dual stochastic Lie group and algebra, $St^T(\hat{\mathbb{1}})$ and $\mathfrak{st}^T(\hat{\mathbb{1}})$, of invertible matrices such that $A^T\hat{\mathbb{1}} = \hat{\mathbb{1}}$.

5.2 Discussion

Application of Lie Theory to the embedding problem for first hitting times... Differentiate between problem of choosing a branch of the matrix logarithm and multiple Markov models having the same first hitting time distribution. Once a principle branch of the logarithm is fixed stochastic Lie algebra can give meaning to the idea of a simplest model, the one expressed in the fewest canonical generators.

Padé approximation needs optimization. The Kronecker vectorization means only low to moderate dimensional models can be handled, due to the quadratic scaling.

Bibliography

[1] 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.

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

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

[4] 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).

[5] 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.

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

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

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

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

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

[11] 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).

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

Appendix A

Julia Implementations

Code dumps of implementations of the algorithms in Julia.