Swapping Methods for Fleming-Viot Estimators of Quasi-Stationary Distributions

Itamar Fiorino

The primary motivation of this thesis is to explore improvements to computing methods related to **quasi-stationary distributions**—limiting distributions conditioned on non-absorption to some absorbing subset that is reached almost-surely.

Contents

Acknowledgements

First and foremost, I would like to thank my parents for raising me to be curious and to expect the most out of myself. I am lucky to be passionate about my studies, and the attitude they imparted on me has been paramount in finding these passions and becoming someone I can be proud of. I want to thank professor Dupuis, my advisor for this thesis, for his seemingly boundless generosity in time, patience, knowledge, and ideas. Without him, this thesis would obviously not exist; much beyond that, my passion for applied math and probability would be less emphatic and poorly informed, at the very least.

I want to thank all of the professors at Brown who spent time teaching me the tricks of the trade, particularly professor Ramanan for helping review this thesis. Finally, I want to thank my friends at Brown (the Keene guys and Angel girls, Daksha and Joey, and many others) and my girlfriend Caroline for keeping me sane in the meantime. Thank you all.

1 Introduction

Include introduction to the problem, with narrative explaining importance in the context of MCMC methods.

2 Preliminaries

The central objects studied in this thesis are Markov processes.

Definition 2.1. *Given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in I}, \mathbb{P})$ and a totally ordered index set I, a **Markov process** taking values in a measurable space $(S, \mathcal{S})$ is a sequence of random variables $(X_t)_{t\in I}$ that satisfy the **Markov property***

$$\mathbb{P}(X_t \in A|\mathcal{F}_s) = \mathbb{P}(X_t \in A|X_s) \quad \forall A \in \mathcal{S}, t \geqslant s$$

In the literature, a **Markov chain** is usually a Markov process where either I or S is discrete. In this thesis, we will work in a setting where S is finite and I will be either $\mathbb{R}^+$ or $\mathbb{N}$; the distinction will be clear.

2.1 Discrete-Time Markov Chains

In this work, we will usually assume that our Markov chains look the same over different times, meaning that the conditional distribution defined by the Markov property is time-invariant, so that

$$\mathbb{P}(X_t \in A|X_s) = \mathbb{P}(X_{t-s} \in A|X_0) = P^{t-s}(X_0, A)$$

We call these Markov chains **homogeneous**, and the transition kernel $P^t(x, dy)$ is usually denoted the **semigroup** of X, as it is a semigroup homomorphic to I under matrix multiplication (or more generally, kernel integration). When we work in discrete-time, we write the matrix

$$P_{x,y} \coloneqq P^1(x, y) = \mathbb{P}(X_1 = y|X_0 = x)$$

Then by conditioning on the intermediate states, we have, (where we denote $\mathbb{P}_\mu(A) = \int \mathbb{P}_x(A)d\mu(x)$ and extend to expectations as usual)

$$\mathbb{E}_\mu[f(X_t)] = \mu P^t f$$

In the study of discrete-time Markov chains, a few conditions on a chain are essential for nice behavior. It is important to note that these conditions may be necessary to make the general statements relating to ergodicity which are found here and in §5.2. In particular, if we generalize to chains over possibly infinite countable state spaces, we need a positive recurrence condition for the Mean Ergodic Theorem to hold (see [7] Theorem 4.1.14).

Definition 2.2. *Given a Markov chain with transition kernel $P^t(x, y)$, a state i is said to be accessible from a state j if $\exists s \in I$ such that $P^s(i, j) > 0$; we write $i \to j$. If $i \to j$ and $i \leftarrow j$, we write $i \leftrightarrow j$ and say that i and j communicate. Communication is an equivalence relation, and the distinct classes of states that communicate are denoted **irreducible classes**. A chain with one irreducible class is said to be **irreducible**.*

Definition 2.3. *A probability measure π is a **stationary distribution** if $\pi P = \pi$*

It can be shown that every irreducible chain has a unique stationary distribution. A natural question to study is the sense in which samples of X converge to a stationary distribution. The first notion is the convergence of the marginals; for nicely-behaved Markov chains, we would like to know how $\mu P^n \to \pi$.

When $I = \mathbb{N}$, we may worry that states in the chain will cycle, in some suitable sense, so that the semigroup doesn't exhibit sufficiently well-behaved mixing properties. The **period** of a state x is $\gcd\{n | P^n(x,x) > 0\}$. The period of a state is a class property, and classes/chains with period 1 are said to be **aperiodic**. The period of an irreducible chain corresponds to the multiplicity of the eigenvalue 1 of the transition probability matrix.

Theorem 2.4. *If a Markov chain with transition probability matrix P is irreducible and aperiodic, then $\exists M \in \mathbb{N}, \varepsilon > 0$ such that for all $\mu \in \mathcal{P}(S)$*

$$\|\mu P^n - \pi\|_1 \leqslant 2(1-\epsilon)^{\lfloor \frac{n}{M} \rfloor}$$

Proof. This is a special case of Corollary 3.1.16 in [7]. ∎

Another way that samples of X can converge to the stationary distribution is in empirical measure, which is defined to be $m(X_n) = \frac{1}{N}\sum_1^N \delta_{\{X_n\}}$. Equivalently, if we choose a time M uniformly in $[N] := \{1, \ldots, N\}$, this is the law of X_M. The sense in which this convergence holds is in probability, under the (metrizable) topology of weak convergence (see §5.2 for a full discussion).

2.2 Continuous-Time Markov Chains

Continuous-time Markov chains (CTMC), the analogs of discrete time Markov chains, require some extra time-regularity conditions. There are several options for defining these processes, each of which is valuable in its own right, either for heuristics, modeling, or understanding the chains from a dynamical systems perspective, and have varying usefulness in defining nonlinear CTMCs. We start with the most general characterization.

Definition 2.5. *A Markov process $\{X_t\}_{t\in I}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **continuous-time Markov chain** if it takes values over a discrete space, and for almost every $\omega \in \Omega$ the map $t \mapsto X_t(\omega)$ is right continuous with left limit.*

The right continuous with left limit condition is often called càdlàg in the literature (from the French *"continue à droite limite à gauche"*).

Proposition 2.6. *The distributions of càdlàg processes are completely determined by its finite-dimensional distributions, so the Markov property can be restated as*

$$\mathbb{P}(X_{t_{n+1}} = x_{n+1} | X_{t_n} = x_n, \ldots, X_{t_1} = x_1) = P(X_{t_{n+1}} = i_{n+1} | X_{t_n} = x_n)$$

for any increasing $\{t_k\}_{k=1}^{n+1}$ and $\{x_k\}_{k=1}^{n+1}$.

As before, we denote $P_{x,y} = P^t(x, \{y\}) = \mathbb{P}(X_t = x | X_0 = y)$ and let $P(t)$ be the matrix $\{P_{x,y}(t)\}$. As before, the general characterization of a stationary distribution $\mathbb{P}_\mu(X_t \in A) = \mu(A)$ extends to

$$\pi P(t) = \pi$$

Definition 2.7. *A matrix Q is called a **transition rate matrix**, or just **rate matrix** if $Q\mathbf{1} = 0$ and $Q_{x,y} \geqslant 0$ when $x \neq y$. A Markov process X_t has rates $Q(t)$ if, for every $t, h > 0$, $x, y \in S$,*

$$\mathbb{P}(X_{t+h} = y | X_t = x) = \delta_{x,y} + Q_{x,y}(t)h + o(h)$$

Then it can be shown that, thinking of $P(t)$ as a Lie group action on the space $\mathbb{R}^k$, the evolution laws of homogeneous chains become homogeneous ODEs (and similarly for inhomogeneous chains):

Theorem 2.8. *If a homogeneous Markov process has rates $Q(t) = Q$, its marginal distributions $\mu(t)$ are just integral curves of the vector field Q starting at $\mu(0)$, so that $P(t) = e^{tQ}$, and P satisfies the relationship*

$$\frac{d}{dt}P(t) = P(t)Q$$

$$\frac{d}{dt}P(t) = QP(t)$$

These are called the Kolmogorov forward and Kolmogorov backward equations, respectively.

Given a discrete-time and discrete-state Markov chain, a basic construction is the **generator**, which measures expected increments at different states:

$$\mathcal{L}f(x) = \mathbb{E}[f(X_1) - f(X_0) | X_0 = x]$$

Obviously, the form of the generator leads to an alternate characterization of stationarity $\mathcal{L}^*\pi = \pi$. Generators are most useful in continuous-time, where they are defined as follows:

Definition 2.9. *Given a continuous-time Markov process X_t that takes values over a Polish space P (discrete, countable, or uncountable), the generator of X_t is an operator $\mathcal{L}$, defined on some subset of $C^2(P)$, as*

$$\mathcal{L}f(x) = \lim_{\delta \to 0} \frac{\mathbb{E}[f(X_\delta) - f(X_0) | X_0 = x]}{\delta}$$

The space of functions for which $\mathcal{L}$ is defined is of paramount importance. Typically, we want it to be defined over a subset of functions that is suitably large as to separate points; since we will be working primarily over finite spaces, we gloss over this detail. We also have two important statements, that are unproven here but which have some heuristic importance.

Lemma 2.10. *Given a homogeneous Markov process X_t over a finite space $\mathcal{S}$,*

(a) The generator $\mathcal{L}$ of X_t characterizes the process.

(b) For any $f \in C^2(S)$, $Y_t = f(X_t) - \int_0^t \mathcal{L}f(X_s)ds$ is a martingale

This result will extend to general Polish spaces under further conditions on $\mathcal{L}$, and the statement (b) on Y_t might be qualified to a local margingale property.

2.2.1 Nonlinear Continuous-Time Markov Chains

2.3 The Perron-Frobenius Theorem for ML Matrices

The next result in linear algebra is critical in characterizing the existence of stationary distributions (and, as we will see, Quasi-Stationary distributions) in the finite-state case. First, we define **Metzler-Leontief (ML)** matrices:

Definition 2.11. *A matrix $\mathbb{R}^{n\times m}$ is called an **ML** matrix if its off-diagonal elements are non-negative. If for any $i \neq j$, there exists $N > 0$ and $i_1, \ldots, i_{N-1}$ that serve as a path of strictly positive rates from i to j, the matrix is said to be irreducible. Rigorously, $q(i_k, i_{k+1}) > 0$ for any $k < N$.*

Then we have following theorem, which is crucial for existence and uniqueness statements:

Theorem 2.12 (The Perron-Frobenius Theorem for ML Matrices)**.** *Suppose A is an irreducible $n \times n$ ML matrix. Then there exists a real eigenvector r whose real part is maximal.*

(i) Left and right eigenvectors associated with r have entries strictly positive, and are unique up to constant multiples.

(ii) Any other eigenvector r_0 of A satisfies $Re(r_0) < r$.

(iii) r is a simple (multiplicity 1) root of the characteristic function of A.

(iv) $r < 0$ iff there exists some $\mathbb{R}^n \ni y \geqslant 0$ such that $Ay \leqslant 0$ with inequality in at least one entry.

(v) if we denote the right and left eigenvectors of (ii) as v^T and w and normalize such that $v^T w = 1$, then for any $t > 0$, there exists $\tau < r$ such that

$$e^{At} = e^{rt} w v^T + O(e^{\tau t})$$

Proof. For a proof of this result, see Page 46 of [6]. ∎

In the theory of ML matrices, r is typically denoted the **Perron-Frobenius** eigenvalue of A.

3 Quasi-Stationary Distributions

We will first explore the general characterization of a quasi-stationary distribution, which holds over general time and spatial constraints; many of the conclusions that hold for stationary distributions have discrete-time analogs. The issue in developing such extensions, underscored in §2.2.1 is that the dynamics associated with Quasi-Stationary Distributions are inherently nonlinear, so MCMC methods for computing them is difficult, and relies on the development of linear chains that approximate the nonlinear dynamics.

Consider a continuous-time Markov process $X = (X_t : t \geqslant 0)$ taking values in state space S endowed with σ-algebra $\mathcal{B}(S)$. We will be concerned with the behavior of our Markov chain within some measurable connected subset D, denoting our forbidden states $\partial = S \backslash D$. Let T denote the killing time $T = T_\partial = \inf\{t \geqslant 0 : X_t \in \partial\}$. The quantities of interest will depend only on X_T, so without loss of generality we set $X_t = X_T$ when $t \geqslant T$. We will assume that the process is almost surely killed so that $\mathbb{P}_x(T < \infty) = 1$.

3.1 Definitions

Definition 3.1. *A probability measure $\nu \in \mathcal{P}(D)$ is said to be a quasi-stationary distribution (QSD) of X on D if for any $t \in [0, \infty)$ and $A \in \mathcal{B}(D)$*

$$\mathbb{P}_\nu(X(t) \in A | T > t) = \nu(A) \tag{3.1.1}$$

or equivalently

$$\mathbb{P}_\nu(X(t) \in A, T > t) = \mathbb{P}_\nu(T > t)\nu(A) \tag{3.1.2}$$

Theorem 3.2. *If ν is a QSD, then $\exists \alpha(\nu) \geqslant 0$ such that*

$$\mathbb{P}_\nu(T > t) = e^{-\alpha(\nu)t}$$

Where we note that $\alpha = 0$ if and only if ν is a stationary distribution.

Proof. First, we notice that for any $g \in \mathcal{M}^+(D)$, monotone convergence and (3.1.2) yields

$$\mathbb{E}_\nu[g(X(t))\mathbb{1}_{T>t}] = \nu(g)\mathbb{P}_\nu(T > t)$$

Setting $g(x) = \mathbb{P}_x(T > s)$ we have $\mathbb{E}_\nu\left[\mathbb{E}_{X(t)}[\mathbb{1}_{T>s}]\mathbb{1}_{T>t}\right] = \mathbb{P}_\nu(T > s)\mathbb{P}_\nu(T > t)$. Finally, we use the Markov property to see

$$\begin{aligned}\mathbb{P}(T > s + t) &= \mathbb{E}_\nu\left[\mathbb{1}_{T>t+s}\right] = \mathbb{E}_\nu\left[\mathbb{1}_{T>t}\mathbb{E}\left[\mathbb{1}_{T>t+s}|\mathcal{F}_t\right]\right] \\ &= \mathbb{E}_\nu\left[\mathbb{E}_{X(t)}[\mathbb{1}_{T>s}]\mathbb{1}_{T>t}\right] = \mathbb{P}_\nu(T > s)\mathbb{P}_\nu(T > t)\end{aligned}$$

Since T has the **memoryless** property, it must be exponentially distributed. Of course, if $\alpha = 0$, the process is never killed, so the quasi-stationary distribution and stationary distribution coincide. ∎

We denote α the **decay rate** or **exponential decay rate**.

3.1.1 Hard-Killing and Soft-Killing Regimes

While the preceding section identifies specific absorbing states, these states and their structure is irrelevant to the distributional properties of a QSD. This identifies two distinct ways of characterizing the QSD, which are in some sense the same but have heuristic and algorithmic implications. In the **hard killing** regime, chains evolve and upon entrance to an absorbing state, never exit; they are in a loose way "dead" to us, hence the term "killing", and their sample paths are irrelevant to the QSD. In the **soft killing** regime, we imagine that a Markov chain evolves and is independently killed according to a state-dependent rate $c(x)$. At the hitting time of $c(x)$, we decide to kill the sample.

Of course, the hard killing regime can be turned into an equivalent soft killing regime by restricting the state space and evolution of particles, and taking $c(x)$ to be $q(x, \partial)$, or by setting $c = \infty \cdot \delta_\partial$ (we are glossing over some details for the extension of the killing rate to infinite values here). Alternatively, a soft-killed regime can be turned into a hard-killed chain by identifying an absorbing state. Hard-killing is typically more useful from a pedagogical perspective when the states that we are conditioning on are explicitly defined, and soft-killing is useful for notation, calculations, and generalizing to other problems that are not QSD problems (or to diffusions).

3.2 Characterization of the QSD

We remember a quasi-stationary distribution is the limiting distribution of a Markov chain conditioned on staying within a region, again call it $D \subset S$. Let,

$$c(x) = \begin{cases} \infty & x \notin D \\ 0 & x \in D \end{cases}$$

Then, we have

$$\mathbb{P}_x(X_t \in A, T > t) = \mathbb{P}_x(X_t \in A, X_s \in D \; \forall \; 0 \leqslant s \leqslant t) = \mathbb{E}_x \left[e^{-\int_0^t c(X_s) ds} \mathbb{1}_A(X_t) \right]$$

since the first expression will be 1 when X_s remains in D before t and zero otherwise (although X is allowed to leave the boundary in some negligible time, this is also negligible over path-space). Therefore, we can find a limiting distribution by taking

$$\theta_t(x, A) = \frac{\mathbb{E}_x \left[e^{-\int_0^t c(X_s) ds} \mathbb{1}_A(X_t) \right]}{\mathbb{E}_x \left[e^{-\int_0^t c(X_s) ds} \right]} \longrightarrow \theta(x, A)$$

That quasi-limiting distributions are quasi-stationary—and a discussion on $c(x)$, which in stochastic control representations is called the cost function—can be found in [5]. Interpreting both θ_t and θ as measures/kernels, it can be shown that $\int \theta_t(y, A)\theta(x, dy) \equiv \theta(x, A)$, which illustrates the quasi-stationarity of θ.

We will start by looking at stationary distributions. From the general expression for the infinitesimal generator $\mathcal{L}$, we can reach a characterization of stationarity. A weak characterization for a stationary distribution is

$$\int_S f(x)\pi(dx) = \int_S \int_S f(y)P_\Delta(x,dy)\pi(dx)$$

where P_Δ is the transition kernel for the process over some time interval Δ. Then,

$$\begin{aligned} 0 &= \int \left[\int f(y)P_\Delta(x,dy) - f(x) \right] \pi(dx) \\ &= \int \int \left[f(y) - f(x) \right] P_\Delta(x,dy)\pi(dx) \end{aligned}$$

Dividing by Δ and taking the limit as $\Delta \to \infty$.

$$\begin{aligned} 0 &= \int \mathcal{L}f(x)\pi(dx) \\ &\overset{\star}{=} \int f(x)\mathcal{L}^*\pi(x)dx \end{aligned}$$

and we get the result $\mathcal{L}^*\pi(x) = 0$. We note that $\overset{\star}{=}$ relies on a change of variables to some reference measure over the space and then use of the L^2 adjoint.

We now find a similar result for quasi-stationary distributions, where we need to worry about losing mass. In particular, we have seen that when a process is evolved starting from a quasi-stationary distribution ν, without conditioning, it will lose mass exponentially [3], so that $\mathbb{P}_\nu(T > t) = e^{-\alpha t}$, where T is the hitting time, $\mathbb{P}_\nu(A) = \int \mathbb{P}_x(A)d\nu$, and α may depend on ν.

Lemma 3.3. *A quasi-stationary distribution $\theta(dx) = \theta(x)dx$, up to normalization, will be characterized by $-\mathcal{L}^*\theta(x) = \alpha\theta(x)$, where α is the decay rate and $\mathcal{L}^*$ is the L^2 adjoint of the infinitesimal generator.*

Proof. First, we use exponential decay to attain the expression

$$\mathbb{E}_\theta \left[e^{-\int_0^\Delta c(X_s)ds} \mathbb{1}_A(X_\Delta) \right] = e^{-\alpha\Delta}\theta(A)$$

which we reached from our earlier characterization of the QSD. By evaluating over C^2 test functions we have

$$\mathbb{E}_\theta \left[e^{-\int_0^\Delta c(X_s)ds} f(X_\Delta) \right] = \mathbb{E}_\theta \left[e^{-\alpha\Delta} f(X_0) \right]$$

and the same process from the regular stationary distribution yields

$$\begin{aligned} 0 &= \mathbb{E}_\theta \left[e^{-\int_0^\Delta c(X_s)ds} [f(X_\Delta) - f(X_0)] - f(X_0) \left[e^{-\int_0^\Delta c(X_s)ds} - e^{-\alpha\Delta} \right] \right] \\ \times -\frac{1}{\Delta}\lim \Longrightarrow &= \int \mathcal{L}f(x)\theta(x)dx - \mathbb{E}_\theta \left[f(X_0) \left(1 - \lim_{\Delta\to 0} [1 + \alpha\frac{\Delta}{\Delta} - \alpha^2\frac{\Delta^2}{2\Delta} + \cdots] \right) \right] \\ &= \int f(x)\mathcal{L}^*\theta(x)dx + \alpha \int f(x)\theta(x)dx \end{aligned}$$

which yields the desired expression. We justify moving the limit inside by dominated convergence. ■

3.3 Existence and Uniqueness over a Finite State-Space

Consider now that the continuous-time Markov chain X_t is over a finite state space. In the language of [5], we assume some transition rate matrix $\widetilde{Q}$ with the following structure:

$$\widetilde{Q} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{a} & \mathbf{Q} \end{bmatrix}$$

where the first row and column represent ∂, the absorbing state. Our forbidden states do not communicate, and our allowed states are absorbed with rates $\mathbf{a}$. Our internal rate matrix Q, is assumed to be irreducible (otherwise we have issues with the uniqueness of our QSD). Then we have the following compound rate matrix:

$$\widetilde{Q}^n = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{Q}^{n-1}\mathbf{a} & \mathbf{Q}^n \end{bmatrix}$$

Seen inductively below.

$$\widetilde{Q}^{n-1}\widetilde{Q} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{Q}^{n-2}\mathbf{a} & \mathbf{Q}^{n-1} \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{a} & \mathbf{Q} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{Q}^{n-1}\mathbf{a} & \mathbf{Q}^n \end{bmatrix}$$

Then with $\widetilde{P}(t)$ as the transition probability matrix of the absorbing chain at time $t > 0$,

$$\widetilde{P}(t) = \sum_{n=0}^{\infty} \frac{\widetilde{Q}^n t^n}{n!}$$

The following lemma is clear for diagonalizable rate matrices, but holds for all matrices

Lemma 3.4. *If λ is an eigenvalue of matrix A associated with an eigenvector v, then e^λ is an eigenvalue of matrix e^A associated with the eigenvector v.*

Now we can characterize the Quasi-Stationary Distribution as a solution to a Linear Algebra problem. We start by finding the forward generator for X_t.

Proposition 3.5. *The generator for a discrete-state process with transition rates $q(x,y)$ is*

$$\mathcal{L}f(x) = \sum_{y \in \mathcal{X}} q(x,y)[f(y) - f(x)] \tag{3.5.1}$$

Proof. Here, we use the fact (via Taylor expansion) that for $i \neq j$, $\mathbb{P}_i(X_h = j) = q(i,j)h + o(h)$, so we have

$$\begin{aligned} \mathcal{L}f(x) &= \lim_{\delta \to 0} \frac{\mathbb{E}[f(X_\delta) - f(X_0) | X_0 = x]}{\delta} \\ &= \lim_{\delta \to 0} \sum_{y \neq x} \frac{(q(x,y)\delta + o(\delta))[f(y) - f(x)]}{\delta} \\ &= \sum_{y \neq x} q(x,y)[f(y) - f(x)] \end{aligned}$$

Using the properties of rate matrices, we also have the form

$$\mathcal{L}f(x) = \sum_{y \in S} q(x,y) f(y)$$

so $\mathcal{L}f = Qf$, identifying f with a the column vector. This will allow, in simple cases, for us to compute the QSD using linear algebra (see Prop. 3.8). ■

Corollary 3.6. *The adjoint of the generator $\mathcal{L}$, given transition rates $q(x,y)$ is:*

$$\mathcal{L}^* f(x) = \sum_{y \neq x} q(y,x)[f(y) - f(x)] + f(x) \sum_{y \neq x} [q(y,x) - q(x,y)]$$

Proof. For a proof, see the general construction in $L^2(\rho)$ in §9.1. ■

Returning to the eigenvalue problem, we have the following result, which identifies the importance of the generator $\mathcal{L}$ and its adjoint.

Theorem 3.7. *Suppose the Markov process X is irreducible on D, then*

(i) A vector $\bar{\nu}(x)$ is an eigenvector for $\mathcal{L}_\rho^*$ if and only if $\nu(x) = \rho(x)\bar{\nu}(x)$ is also an eigenvector of $\mathcal{L}^*$ with the same eigenvalue.

(ii) There exist a real eigenvalue of $\mathcal{L}$ and $\mathcal{L}^*$, $\lambda > 0$, for which $r > \lambda$ for any eigenvalue $r \neq \lambda$.

(iii) There exists a unique quasi-stationary distribution ν associated with this eigenvalue

(iv) There exist eigenvectors ψ and ϕ of $-\mathcal{L}$ and $-\mathcal{L}^*$ respectively, associated with the eigenvalue λ. ϕ is the unique quasi-stationary distribution and $(\phi, \lambda) = 1$. For any $\rho \in \mathcal{M}^+(D)$, $\bar{(\phi)} = \frac{\phi}{\rho}$ is an eigenvector of $-\mathcal{L}_\rho^*$ associated with the same eigenvector λ

(v) The function ψ characterizes the exit/decay rate from D, or, loosely, the rate at which mass leaves conditioned on its origin

$$\psi(X) = \lim_{t \to \infty} e^{\lambda t} \mathbb{P}_X(\tau_\delta > t)$$

Proof. (i) Notice that $\mathcal{L}_\rho^*(\bar{\nu}) = \frac{\mathcal{L}^*(\rho\bar{\nu})(x)}{\rho(x)} = \frac{\lambda\rho(x)\nu(x)}{\rho(x)} = \lambda\bar{\nu}(x)$

(ii) By the assumption of irreducibility, Q is a ML matrix, and thus by Theorem 2.12, Q has eigenvalue $r > 0$ associated with eigenvectors v^T and w (whose entries are positive). Then if $\lambda = -r$, we notice:

$$\mathcal{L}v(x) = \sum_{y \in S} q(x,y)\,[v(y) - v(x)] = \sum_{y \in S} q(x,y)v(y) = -rv(x) = \lambda v(x)$$

Then since $\mathcal{L}$ and $\mathcal{L}^*$ have the same eigenvalues:

$$\begin{aligned}\lambda \notin \sigma_{\mathcal{L}} &\iff \exists S : S(\mathcal{L} - \lambda I) = I \\ &\iff (\mathcal{L}^* - \lambda I)S^* = I \\ &\iff \lambda \notin \sigma_{\mathcal{L}^*}\end{aligned}$$

by (i) we are done.

(iii) This can be proven in a number of ways. The first, and simplest way, is to use the characterization of the QSD in Lemma 2.4 of [5].

(iv) If we set $\psi = v$ and $\phi = w$ from (ii) and Theorem 2.12, then we can see that ϕ is an eigenvector of $-\mathcal{L}$: Then if $\lambda = -r$, we notice:

$$-\mathcal{L}v(x) = -\sum_{y \in S} q(x,y)\left[v(y) - v(x)\right] = -\sum_{y \in S} q(x,y)v(y) = -rv(x) = \lambda v(x)$$

Then since $\mathcal{L}$ and $\mathcal{L}^*$ have the same eigenvalues:

$$\begin{aligned}\lambda \notin \sigma_{\mathcal{L}} &\iff \exists S : S(\mathcal{L} - \lambda I) = I \\ &\iff (\mathcal{L}^* - \lambda I)S^* = I \\ &\iff \lambda \notin \sigma_{\mathcal{L}^*}\end{aligned}$$

by (i) we are done. Then also ψ is an eigenvector of $-\mathcal{L}^*$:

$$\psi^T Q = \lambda \psi^T \implies \lambda\psi = Q^T\psi = Q^*\psi = \mathcal{L}^*\psi$$

(v) See [5]

■

The following proposition is not immediately obvious from the typical presentation of the generator, and allows for sanity-check computations of simple QSD.

Proposition 3.8. *If the rates $q(x,y)$ correspond with a rate matrix Q,*

$$\sum_{y \neq x} q(x,y)[f(y) - f(x)] + f(x)h(x) = (Q + H)(x)$$

Where H is a diagonal matrix with entries $H(x,x) = h(x)$

Proof.

$$\begin{aligned}\sum_{y \neq x} q(x,y)[f(y) - f(x)] + f(x)h(x) &= \sum_{y \neq x} q(x,y)f(y) - f(x)\sum_{y \neq x} q(x,y) + f(x)h(x) \\ &= \sum_{y \neq x} q(x,y)f(y) + f(x)q(x,x) + f(x)h(x) \\ &= \sum_{y \in S} q(x,y)f(y) + h(x)f(x) = (Q + H)(x)\end{aligned}$$

■

Applied to $\mathcal{L}f(x)$, we see $\mathcal{L} = \widetilde{Q}$, so a QSD, which is an eigenvector of $\mathcal{L}^*$, will be an eigenvector of $Q^* = Q^T$ when restricted to allowed states, and 0 otherwise.

4 Stochastic Control

The primary objects in Stochastic Control relevant to the study of QSDs are twofold: a minimal cost function, whose optimal policy is related to the quasi-stationary distribution for the chain, and the optimal ergodic cost, which relates to the decay rate. In this section, we begin by introducing both objects in the discrete-time setting, which we generalize to CTMCs in the following section

Consider the following setup, found in [5]:

- Take $\{X_n\}_{n\in\mathbb{N}}$, a discrete time Markov chain over finite S with $X_0 = x$
- A time horizon $N \in \mathbb{N}$, which is trivially sufficient for well-posedness of the optimization problem.
- A control space U, which can be interpreted as the set of controls that are available to minimize cost. U is a Polish space (a metrizable, separable topological space).
- A *running* cost, which contributes to total cost throughout the living time of the Markov Chain, $c : S \times U \to [0,\infty)$. This must be a **tightness function** over U [5], ie. for each $x \in S$, $M \in [0,\infty)$, the sublevel sets $\{u : c(x,u) \leqslant M\} \subset U$ are compact. We include a **terminal cost** F as well.
- A map of controlled transition probabilities which depend on the chosen control $p : S^2 \times U \to [0,1]$ which represent a discrete-time Markov transition matrix for each control, under the assumption that transition probabilities are continuous in control (i.e. $u \mapsto p(x,y|u)$ is continuous for any $x, y \in S$).
- The policy π representing the time and state-dependent control. $\pi = \{g_i\}_{i<N}$ where $g_i : S \to U$.

The finite-time problem asks the control to perform the following optimization, where $U_i = \pi(X_i) = g_i(X_i)$, and X_i follows transition probabilities governed by p:

$$V(x,i) = \inf_{\pi} \mathbb{E}^{\pi}_{x,i}\left[\sum_{j=i}^{N-1} c(X_j,U_j) + F(X_N)\right]$$

Such prototypical stochastic control problems are typically solved backwards using dynamic programming, noticing the following equality, formally described in [5]:

$$V(x,i) = \inf_{u\in U}\left[c(x,u) + \sum_{y\in S} p(x,y|u)V(y,i+1)\right]$$

The second object that is important in the study of QSDs is the optimal ergodic cost, which will be interpretable as the decay rate:

$$\gamma = \inf_{\pi} \lim_{n\to\infty} \frac{1}{n}\mathbb{E}^{\pi}_{x}\left[\sum_{i=0}^{n-1} c(X_i,U_i)\right]$$

4.1 Forward Ergodic Control

There are two main ergodic control representations that are discussed in the study of QSDs, one related to the generator $\mathcal{L}$ and one associated with its adjoint $\mathcal{L}^*$. Consider a modified process $\bar{X}_t$ with new, controlled, jump rates $\bar{b} = \{\bar{b}(x,y)\}_{x,y\in S}$. These modified jump rates are constrained by an *admissibility* condition:

Definition 4.1. *A control process is **admissible** if, for any $x_0 \in D$, $\mathbb{P}_{x_0}(\bar{X}_t \in D \ \forall t > 0) = 1$. We denote the class of admissible control processes $\mathcal{A}$.*

Remark 4.1.1. *It is not difficult to see that a control process is admissable iff $\bar{b}(x,y) = 0$ when $x \in D, y \notin d$.*

The formulation of the controls requires consideration for flows into the boundary under the control scheme. As such, we denote:

$$q(x,\partial D) = \sum_{x\notin D} q(x,z)$$

Then consider the following cost problem:

$$J(x_0,\bar{b}) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{x_0}\left[\int_0^T \left[q(\bar{X}_t,\partial D) + \sum_{y\neq\bar{X}_t, y\in D} q(\bar{X}_t,y)\mathscr{L}\left(\frac{\bar{b}(\bar{X}_t,y)}{q(\bar{X}_t,y)}\right)\right] dt\right]$$

where $\mathscr{L}(x) = x\log x - x + 1$ for $x = 0$ and $\mathscr{L}(0) = 1$. We notice two terms: the first term penalizes the control for being near the boundary, and the second term penalizes the control for varying from the original dynamics, similarly to relative entropy, or KL Divergence. The optimal cost is $\inf_{\bar{b}\in\mathcal{A}} J(x_0,\bar{b})$. Theorem 4.2 will give a characterization of J in terms of the below **Bellman equation**:

$$\gamma = \inf_{\bar{b}\in\mathcal{A}}\left\{\sum_{y\neq x, y\in D}\bar{b}(x,y)[\Gamma(y)-\Gamma(x)] + \sum_{y\neq x, y\in D} q(x,y)\mathscr{L}\left(\frac{\bar{b}(x,y)}{q(x,y)}\right) + q(x,\partial D)\right\} \tag{4.1.1}$$

Theorem 4.2. *Suppose the process X_t is irreducible on D. Then*

(i) If λ and ψ are as given in Theorem 3.7 and define $\Psi = -\log\psi$. Then (λ,Ψ) is a solution to the Bellman equation (4.1.1).

(ii) Solutions to the Bellman equation are unique in the first variable, and unique in the second variable up to constants.

(iii) If (λ,Ψ) is a solution to the Bellman equation, then a process Y_t with the following control rates:

$$b^*(x,y) = \begin{cases} 0 & x\in D, y\notin D \\ q(x,y)e^{\Psi(x)-\Psi(y)} & x,y\in D, x\neq y \\ \lambda + q(x,x) & x,y\in D, x = y \end{cases} \tag{4.2.1}$$

is a Markov process with unique invariant measure $\mu = \phi\psi$, where ϕ and ψ are from Theorem 3.7.

(iv) $\forall x_0 \in D$, $J(x_0) = \lambda$

(v) The control defined in (iii) is optimal, in fact, for any $x_0 \in D$,

$$\lambda = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{x_0}\left[\int_0^T \left[q(Y_t, \partial D) + \sum_{y\neq Y_t, y\in D} q(Y_t, y)\mathscr{L}(e^{\Psi(Y_t)-\Psi(y)})\right] dt\right]$$

Furthermore, the eigenvalue λ *can be determined from the invariant measure* μ *by the formula*

$$\lambda = \sum_{z\in D}\left[q(z, \partial D) + \sum_{y\neq z, y\in D} q(z, y)\mathscr{L}(e^{\Psi(Y_t)-\Psi(y)})\right]\mu(z)$$

Proof. Recall that ψ is a solution of

$$\begin{cases} \mathcal{L}\psi(x) + \lambda\psi(x) = 0 & x \in D \\ \psi(x) = 0 & x \in S\backslash D \end{cases}$$

If we define $\Psi = -\log\psi$, then by using our characterization of the infinitesimal generator for discrete-state Markov chains, and multiplying by $e^{-\Psi(x)}$, we have

$$\begin{aligned} 0 &= \sum_{y\neq x, y\in D} q(x, y)\left[e^{-\Psi(y)} - e^{-\Psi(x)}\right] + \lambda e^{-\Psi(x)} + \sum_{y\neq x, y\in\partial D} q(x, y)e^{-\Psi(x)} \\ &= \sum_{y\neq x, y\in D} q(x, y)\left[e^{-\Psi(y)+\Psi(x)} - 1\right] + \lambda - q(x, \partial D) \end{aligned}$$

Then we use the trick that, for any $a \in \mathbb{R}$,

$$e^{-a} - 1 = -\inf_{b\geqslant 0}\{ab + \mathscr{L}(b)\}$$

So with $a = \Psi(y) - \Psi(x)$, we have

$$\begin{aligned} \lambda &= \sum_{y\neq x, y\in D} q(x, y)\left[\inf_{b_x\geqslant 0}\{b_x[\Psi(y) - \Psi(x)] + \mathscr{L}(b_x)\} + q(x, \partial D)\right] \\ &= \inf_{b(x,y)\geqslant 0}\left\{\sum_{x\neq y, y\in D} q(x, y)b(x, y)[\Psi(y) - \Psi(x)] + \sum_{y\neq x, y\in D} q(x, y)\mathscr{L}(b(x, y)) + q(x, \partial D)\right\} \end{aligned}$$

Finally, we find the optimal control $\bar{b}$ by substitution $\bar{b}(x, y) = q(x, y)b(x, y)$ supported on nonabsorption (i.e. $\bar{b} : (x \in D, y \notin D) \mapsto 0$). So (i) follows. ∎

4.2 Backward Ergodic Control

Now consider another modified process $\widetilde{X}_t$ with new, controlled, jump rates $\tilde{b} = \{\tilde{b}(x, y)\}_{x,y\in S}$. These modified jump rates are again constrained by an *admissibility* condition:

Definition 4.3. *A control process on the backward dynamics is* ***admissible*** *if, for any* $x_0 \in D$, $\mathbb{P}_{x_0}(\widetilde{X}_t \in D \ \forall t > 0) = 1$. *We denote this class of admissible control processes* $\mathcal{A}^*$, *which is independent of* x_0 *by irreducibility.*

Remark 4.3.1. *We again notice that a control process is admissable iff* $\tilde{b}(x,y) = 0$ *when* $x \in D, y \notin d$.

The formulation of the controls requires consideration for flows into the boundary under the control scheme. Now, since $\mathcal{L}^*$ has a 0th order term, we need an additional term on our running cost, we define

$$c(x) = q(x, \partial D) + \sum_{y \in d, y \neq x} [q(x,y) - q(y,x)]$$

Then we consider the following cost problem:

$$J(x_0, \tilde{b}) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_{x_0} \left[\int_0^T \left[c(\widetilde{X}_t) + \sum_{y \neq \widetilde{X}_t, y \in D} q(y, \widetilde{X}_t) \mathscr{L} \left(\frac{\tilde{b}(\widetilde{X}_t, y)}{q(y, \widetilde{X}_t)} \right) \right] dt \right]$$

where $\mathscr{L}(x) = x \log x - x + 1$ for $x = 0$ and $\mathscr{L}(0) = 1$. The optimal cost is $\inf_{\tilde{b} \in \mathcal{A}^*} J(x_0, \tilde{b})$. Theorem 4.4 will give a characterization of J in terms of the below **Bellman equation**:

$$\gamma = \inf_{\tilde{b} \in \mathcal{A}^*} \left\{ \sum_{y \neq x, y \in D} \tilde{b}(x,y)[\Gamma(y) - \Gamma(x)] + \sum_{y \neq x, y \in D} q(y,x) \mathscr{L} \left(\frac{\tilde{b}(x,y)}{q(y,x)} \right) + c(x) \right\} \tag{4.3.1}$$

Theorem 4.4. *Suppose the process* X_t *is irreducible on* D. *Then*

(i) If λ *and* ϕ *are as given in Theorem 3.7 and define* $\Phi = -\log \phi$. *Then* (λ, Φ) *is a solution to the Bellman equation (4.3.1).*

(ii) Solutions to the Bellman equation are unique in the first variable, and unique in the second variable up to constants.

(iii) If (λ, Φ) *is a solution to the Bellman equation, then a process* Y_t *with the following control rates:*

$$b'(x,y) = \begin{cases} 0 & x \in D, y \notin D \\ q(y,x) e^{\Phi(x) - \Phi(y)} & x, y \in D, x \neq y \\ \lambda + q(x,x) & x, y \in D, x = y \end{cases} \tag{4.4.1}$$

is a Markov process with unique invariant measure $\mu = \phi\psi$, *where* ϕ *and* ψ *are from Theorem 3.7.*

(iv) $\forall x_0 \in D$, $J(x_0) = \lambda$

(v) The control defined in (iii) is optimal, in fact, for any $x_0 \in D$,

$$\lambda = \lim_{T \to \infty} \frac{1}{T} \mathbf{E}_{x_0} \left[\int_0^T \left[c(Y_t) + \sum_{y \neq Y_t, y \in D} q(y, Y_t) \mathscr{L}(e^{\Phi(Y_t) - \Phi(y)}) \right] dt \right]$$

Furthermore, the eigenvalue λ can be determined from the invariant measure μ by the formula

$$\lambda = \sum_{z \in D} \left[c(z) + \sum_{y \neq z, y \in D} q(y,z) \mathscr{L}\left(e^{\Phi(Y_t) - \Phi(y)}\right) \right] \mu(z)$$

5 Fleming-Viot Particle Systems

One computing scheme for finding quasi-stationary distributions (described in [8]), is based on a **Fleming-Viot process**. It is a class of **Interacting Particle Systems** (IPS) that is defined for the QSD problem as follows:

Algorithm 5.1 (Fleming-Viot)**.**

(1) Sample according to some distribution in D to determine the starting positions for N particles whose law is the same as the unconditioned semigroup.

(2) Evolve the system while particles remain in D (and no particle is killed)

(3) When a particle is killed, replace it with a particle sampled on the uniform distribution over the other $N-1$ particles.

(4) Repeat steps (2-3)

The goal of this section, then, will be to formalize the language relating to Fleming-Viot systems, and to prove the following theorem:

Theorem 5.2. *Suppose that an IPS system (see RAS), $\{X_t^{(i)}\}$, evolves according to the rates q with particle killing rates $c(X_t^{(i)})$. Further, suppose that particles are reborn according to the distribution $\frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(dx)$. Then*

$$\left(\frac{1}{N} \sum_{i=1}^{N} \delta_{X^{(i)}}(dx), X^{(1)} \right) \xrightarrow{w} \left(\mathbb{P}_x(Z \in dx), Z \right)$$

on $[0,T]$ as $N \to \infty$.

This scheme extends to linear functionals but may not extend to nonlinear functionals because of the correlation of paths; in this case, N is critical, and we notice that approximations can do poorly near the boundary.

5.1 Formalism

We will primarily use the notation from [2]. Suppose we continue in a hard-killing regime, where particles evolve according to rates $Q_{x,y} = q(x,y)$ over a state space S, and that particles are killed as soon as they enter states $\partial D := S \backslash D$; D are the allowed states,

and ∂D are the killed/forbidden states, which are absorbing. We denote $c : D \to \mathbb{R}^+$ as $c : x \mapsto q(x, \partial D)$. As before, we define $\mathcal{L}$ to be the generator

$$\mathcal{L}f(x) = \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{E}[f(X_\delta) - f(X_0) | X_0 = x]$$

Then if X has law μP_t (with $P_t = e^{tQ}$), then its law at time t conditioned on non-absorption is defined, over bounded test functions supported on D, as

$$\mu T_t f = \frac{\mu P_t f}{\mu P_t \mathbb{1}_D} = \frac{\sum_{x \in D} P_t f(x) \mu(x)}{\sum_{x \in D} P_t \mathbb{1}_D(x) \mu(x)}$$

We set $T_t f(x) = \delta_x T_t f$, and recall that a quasi-stationary distribution is a distribution satisfying $\nu T_t f = \nu f$ for any test function f.

If, suppressing the dependence on N, $\eta^{(N)}(k) = \eta(k)$ is the number of particles in state k, the chain η_t is a Markov chain with state space

$$E = E^{(N)} = \left\{ \eta : D \to \mathbb{N}_0 : \sum \eta(k) = N \right\}$$

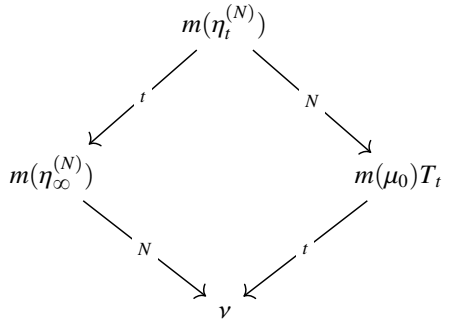
and the generator of this chain is, again suppressing the dependence on N, $\mathcal{N} = \mathcal{N}^{(N)}$. Since η_t has two dynamics that govern it—resampling according to $c(\eta(k))$ and moving according to $q(\eta(k), y)$—$\mathcal{N}$ will be

$$\mathcal{N} f(\eta) = \sum_{i \in D} \eta(i) \left[\sum_{j \in D} (f(T_{i \to j} \eta) - f(\eta))(Q_{i,j} + c(i) \frac{\eta(j)}{N-1}) \right]$$

Where $T_{i \to j} \eta(i) = \eta(i) - 1$ and $T_{i \to j} \eta(j) = \eta(j) + 1$. The empirical distribution of the N particle system is

$$m(\eta)(dx) = \frac{1}{N} \sum_{k \in D} \eta(k) \delta_{\{k\}}(dx)$$

Since both increasing N and increasing t can improve the accuracy of the Fleming-Viot estimate $m(\eta_t^{(N)})$, there are two ways in which our particle system may be shown to converge to the desired QSD:



Where convergences are in distribution. In order to show that the Fleming-Viot system is asymptotically consistent, we will first show that with sufficiently large N, the correlations between particles in the system goes to zero. We will then show that for fixed N, the empirical measure of the system converges to the system's stationary distribution in an almost-sure sense. We are using the following definition of asymptotic consistency:

Definition 5.3. *A sequence of estimators (θ_N^t) is said to be an **asymptotically consistent** estimator of a parameter θ if for any $\varepsilon > 0$,*

$$\lim_{N\to\infty}\lim_{t\to\infty}\mathbb{P}(|\theta_N^t - \theta| > \varepsilon) = 0$$

5.2 Convergence of Time-Averaged Empirical Measure

We will start by showing that the empirical measure of a generic continuous-time chain converges to the system's stationary distribution. We do this first by showing the property for a generic discrete-time chain with an ergodic transition kernel, using a convenient pair measure, then use the semigroup of the continuous-time chain to generalize. The convenient trick in the first theorem does not work for nonlinear operators, and is presented as a testament to the difficulty of solving nonlinear problems, like the QSD problem. To provide bounds for the Fleming-Viot system, a more involved approach is necessary, which will result not only in the convergence for both the normal and swapped chain, but also rates of convergence.

Theorem 5.4. *Let $\{X_i\}_{i\in\mathbb{N}}$ be a Markov chain with an ergodic transition kernel P and stationary measure π. Then*

$$\mu^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i} \Rightarrow \pi$$

almost surely.

Proof. Consider the pair measure

$$m^N(A\times B) = \frac{1}{N}\sum_{i=1}^{N}\delta_{(X_i,X_{i+1})}(A\times B)$$

Then clearly the difference between the first marginal $[m^N]_1 = \mu^N$ and the second marginal $[m^N]_2$ converges to zero:

$$([m^N]_1 - [m^N]_2)(A) = \frac{1}{N}[\delta_{X_i}(A) - \delta_{X_{N+1}}(A)] \to 0$$

Additionally, by the Markov property, for any test function f

$$\mathbb{E}\left[f(X_{i+1}) - \sum_{y\in S}P(X_i,y)f(y)\;\middle|\;X_i\right] = 0$$

So we can compare $[m^N]_2 f$ and $[m^N]_1 Pf$:

$$
\begin{aligned}
&\mathbb{P}\left(|[m^N]_2 f - [m^N]_1 Pf| \geqslant \varepsilon\right) \\
&= \mathbb{P}\left(\left|\sum_{y\in S}[m^N]_2(y)f(y) - \sum_{y\in S}\sum_{x\in S}P(x,y)[m^N]_1(x)f(y)\right| \geqslant \varepsilon\right) \\
&= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(X_{i+1}) - \frac{1}{N}\sum_{i=1}^{N}\sum_{y\in S}P(X_i,y)f(y)\right| \geqslant \varepsilon\right) \\
&\leqslant \frac{1}{\varepsilon^2}\mathbb{E}\left[\frac{1}{N^2}\left(\sum_{i=1}^{N}\left(f(X_{i+1}) - \sum_{y\in S}P(X_i,y),f(y)\right)\right)^2\right] \\
&\overset{*}{=} \frac{1}{\varepsilon^2}\mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}\left(f(X_{i+1}) - \sum_{y\in S}P(X_i,y),f(y)\right)^2\right] \\
&\leqslant \frac{4\|f\|_\infty}{\varepsilon^2 N}
\end{aligned}
$$

Where $\overset{*}{=}$ relies on a conditioning argument; for $i < j$,

$$
\begin{aligned}
&\mathbb{E}\left[\left(f(X_{i+1}) - \sum_{y\in S}P(X_i,y)f(y)\right)\left(f(X_{j+1}) - \sum_{y\in S}P(X_j,y)f(y)\right)\right] \\
&= \mathbb{E}\left[\left(f(X_{i+1}) - \sum_{y\in S}P(X_i,y)f(y)\right)\mathbb{E}\left[f(X_{j+1}) - \sum_{y\in S}P(X_j,y)f(y)\;\middle|\; X_j\right]\right] \\
&= 0
\end{aligned}
$$

Then we have convergence along any subsequence, so by tightness, with probability 1,

$$[m^\infty]_2 f - [m^\infty]_1 Pf = 0$$

So $\mu^\infty(y) = \sum_x P(x,y)\mu^\infty(x)$ and $\mu^\infty = \pi$ with probability 1. ∎

Extending to the continuous-time case, we have

Theorem 5.5. *Let $\{X_t\}_{t\in\mathbb{R}^+}$ be an irreducible Markov chain with transition rates Q over a finite state-space. Then*

$$\frac{1}{T}\int f(X_t)dt \Rightarrow \pi f \quad a.s.$$

Proof. Let $Q_1 = \sup_x \sum_{y\neq x} Q_{x,y} < \infty$. First, we know that if N_t is the number of jumps of X_t in the interval $(0,t]$, and N_t^* is a Poisson process with jump rate Q_1, then

$$\mathbb{P}(N_t \geqslant n) \leqslant \mathbb{P}(N_t^* \geqslant n) \leqslant \frac{Q_1 t}{n}$$

Then if f is finitely supported and càdlàg, then any intervals for which the integral over f is different from its left-Riemann sum must have at least one discontinuity, so,

$$\left|\int_0^T f(x)dx - \frac{T}{n}\sum_{k=1}^{n} f(\frac{TK}{n})\right| \leqslant \frac{T\|f\|_\infty|\{t\in[0,T] : f(t) \neq lim_{s\nearrow t} f(s)\}|}{n}$$

Then setting $T = N$ and $n = N^3$ and dividing by T,

$$\left|\frac{1}{N}\int_0^N f(x)dx - \frac{1}{N^3}\sum_{i=1}^{N^3} f(i/N^2)\right| \leqslant \frac{\|f\|_\infty|\{t\in[0,N] : f(t) \neq lim_{s\nearrow t} f(s)\}|}{N^3}$$

So for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\int_0^N f(x)dx - \frac{1}{N^3}\sum_{i=1}^{N^3} f(i/N)\right| \geqslant \varepsilon\right) \leqslant \mathbb{P}\left(N_t \geqslant \frac{\varepsilon N^3}{\|f\|_\infty}\right) \leqslant \frac{Q_1\|f\|_\infty}{\varepsilon N^2}$$

which is summable. So by Borel-Cantelli, and the previous theorem, we are done (to be exact, we would need to use Borel Cantelli twice, once to show convergence of the integral to the sum, and once to show that increasing the subdivision size does not meaningfully change the rate of convergence of the empirical measures). ■

5.3 Convergence to the Conditioned Process

The pair-measure trick that was used in Theorem 5.4 does not work for chains that evolve through nonlinear dynamics. Although over a small time interval ε, the evolution of the empirical measure $m(\nu_t)$ evolves similarly to T_ε, the pair measure does not evolve according to T_ε, as T_ε is nonlinear (see §2.2.1). Instead, we need to demonstrate that as N is large, the covariance between particles goes to 0, so that they are, in a sense, asymptotically independent. Then, by a bias-variance argument, we are able to show that the evolution of the empirical measure of the Fleming-Viot system follows T_t closely in N (and uniformly in t), which yields the desired convergence. This is trivially extended to the time average.

The first result, relying on a coupling argument, is what [2] dubs "Wasserstein exponential ergodicity". Let

$$\lambda = \inf_{i,i'\in D}\left(Q_{i,i'} + Q_{i',i} + \sum_{j\neq i,i'} Q_{i,j} \wedge Q_{i',j}\right)$$
$$\rho = \lambda - (\sup(c) - \inf(c))$$

Theorem 5.6 (Wasserstein Exponential Ergodicity)**.** *For any two η_t, η'_t generated by Algorithm 5.1,*

$$\mathcal{W}_d(\mathrm{Law}(\eta_t), \mathrm{Law}(\eta'_t)) \leqslant e^{-\rho t} W_d(\mathrm{Law}(\eta_0), \mathrm{Law}(\eta'_0))$$

Where d_1 is a scaled version of total-variation distance:

$$d_1(\eta,\eta') = \frac{1}{2}\sum_{j\in F}|\eta(j) - \eta'(j)| = N d_{TV}(m(\eta), m(\eta'))$$

and $\mathcal{W}$ is the Wasserstein distance

$$\mathcal{W}_d(\mu,\mu') = \inf_{\substack{\eta\sim\mu \\ \eta'\sim\mu'}} \mathbb{E}[d(\eta,\eta')]$$

Proof. We will construct a coupling of (η_t) and (η'_t), starting at η_0 and η'_0 respectively, which maximizes the probability of coalescence. If (ξ^n_t) and (ξ'^n_t) are the positions of the particles in each configuration, then we have the identity

$$d_1(\eta_t,\eta'_t) = |\{1 \leqslant k \leqslant N | \eta^k_t \neq \eta'^k_t\}|$$

We will then decompose the generator of the joint process

$$\mathbb{L}f(\eta,\eta') = \sum_{i,i',j,j'\in D} A(i,i',j,j')(f(T_{i\to j}\eta, T_{i'\to j'}\eta') - f(\eta,\eta'))$$

as a generator associated with the dynamics and a generator associated to the killing (we use $\mathbb{L}$ to emphasize that this is the generator of our designed coupled dynamics, and distinct from the $\mathcal{L}$ dynamics that govern the original process). So $A = A_Q + A_h$. Lets start by describing how A_Q couples the two systems, using the decomposition

$$\eta(i) = (\eta(i) \wedge \eta'(i)) + (\eta(i) - \eta'(i))_+$$

which decomposes the particles by particles that can be coupled and excess particles.

There are $\eta(i) \wedge \eta'(i)$ particles that are at the same site i as their counterpart. We couple these particles so that they move together, and

$$A_Q(i,i,j,j) = (\eta(i) \wedge \eta'(i))Q_{i,j}$$

If a particle is not paired, the good option would be to choose a particle that sees similar states as itself and maximize the chance they go to the same state; you could think of even better couplings to improve the convergence estimate. We will use the more brazen approach of pairing particles at random. For a η particle at i, the probability that its corresponding pair particle is at site i' is

$$\frac{(\eta'(i') - \eta(i'))_+}{d_1(\eta,\eta')}$$

since the numerator are the number of η' particles that have not been pair off and the denominator is the total number of not paired particles. So paired particles can coalesce in two ways: by moving them simultaneously to the same state, or by moving one to the others state and not changing the original dynamics. For a state $j \neq i, i'$, this gives

$$\begin{aligned}
A_Q(i,i',j,j) &= \frac{(\eta(i') - \eta'(i'))_+ \cdot (\eta'(i') - \eta(i'))_+}{d_1(\eta,\eta')}(Q_{i,j} \wedge Q_{i',j}) \\
A_Q(i,i',i',i') &= \frac{(\eta(i') - \eta'(i'))_+ \cdot (\eta'(i') - \eta(i'))_+}{d_1(\eta,\eta')}(Q_{i,i'}) \\
A_Q(i,i',i,i) &= \frac{(\eta(i') - \eta'(i'))_+ \cdot (\eta'(i') - \eta(i'))_+}{d_1(\eta,\eta')}(Q_{i',i})
\end{aligned}$$

Of course the particles can move with dynamics that do not increase the number of paired particles, but we are concerned about the convergence, so we will omit these values of A_Q. We now proceed with the coupling of the killing dynamics.

Take a coupled pair from $\eta(i) \wedge \eta'(i)$. Since the killing rates depend only on the position of the particles, they have the same killing rates, so when one is killed, we can kill the other and try to maximize the probability they move to the same state, say j. We may do this with probability

$$\frac{\eta(j) \wedge \eta'(j) - \delta_{i,j}}{N-1}$$

So we set

$$A_p(i,i,j,j) = p_0(i)\,(\eta(i) \wedge \eta'(i))\,\frac{\eta(j) \wedge \eta'(j) - \mathbb{1}_{i=j}}{N-1}$$

For unpaired particles, we again pair them uniformly, and choose to kill them at the same time when possible. Once killed, since the empirical distributions might differ, we send them to the same particle with the same probability as before, this time without the $\delta_{i,j}$ term:

$$A_p(i,i',j,j) = (\eta(i) - \eta'(i))_+ \cdot \frac{(\eta'(i') - \eta(i'))_+}{d_1(\eta,\eta')} \cdot (c(i) \wedge c(i'))\frac{\eta(j) \wedge \eta'(j)}{N-1}$$

To demonstrate that this is a valid coupling, we would need to show that when a function $f : E \times E \to \mathbb{R}$ satisfies $f(\eta,\eta') = g(\eta)$ for some g, then $\mathbb{L}f(\eta,\eta') = \mathcal{L}g(\eta)$, and similarly for η'. Continuing with the exponential convergence, we reach, decomposing $\mathbb{L} = \mathbb{L}_Q + \mathbb{L}_p$ and using the fact that d_1 is only changed in the cases we identified,

$$\mathbb{L}_Q d_1(\eta,\eta') \leqslant -\lambda d_1(\eta,\eta')$$

and

$$\mathbb{L}_p d_1(\eta,\eta') \leqslant (\sup(c) - \inf(c))d_1(\eta,\eta')$$

So $\mathbb{L}d_1(\eta,\eta') \leqslant \rho d_1(\eta,\eta')$, and by the Kolmogorov forward equation and Gronwall's inequality,

$$\mathbb{E}[d_1(\eta_t,\eta'_t)] \leqslant e^{-\rho t}\mathbb{E}[d_1(\eta_0,\eta'_0)]$$

And by the definition of Wasserstein distance, we are done. ∎

The following two theorems are also found in [2]. We will omit the proofs of the following theorems and corollary.

Theorem 5.7 (Covariance Estimates)**.** *If $\mathbb{E}_\eta$ represents expectation of a functional over a system generated by 5.1 under initial distribution η, then for any $k,l \in D, \eta \in E$, and $t > 0$,*

$$\left|\mathbb{E}_\eta\left[\frac{\eta_t(k)}{N}\frac{\eta_t(l)}{N}\right] - E_\eta\left[\frac{\eta_t(k)}{N}\right]E_\eta\left[\frac{\eta_t(l)}{N}\right]\right| \leqslant \frac{2(Q_1 - c_1)}{N-1}\frac{1-e^{-2\rho t}}{\rho}$$

Where, since we are working over a finite number of states,

$$Q_1 = \sup_i(-Q_{i,i}) < \infty, \quad c_1 = \sup_i(c(i))$$

Theorem 5.8 (Convergence to the Conditioned Process)**.** *There exist $B, C > 0$ such that for any initial distribution $\eta \in E^{(N)}$ and $\mu \in \mathcal{P}(S)$,*

$$\sup_{\|\varphi\|_\infty \leqslant 1} \mathbb{E}_\eta[|m(\eta_t)(\varphi) - \mu T_t \varphi|] \leqslant Ce^{Bt}\left(\frac{1}{\sqrt{N}} + d_{TV}(m(\eta), \mu)\right)$$

Corollary 5.9. *Under the same conditions as the previous theorem, we can find $K_0, \gamma > 0$ for which*

$$\sup_{t \geqslant 0} \sup_{\|\varphi\|_\infty \leqslant 1} \mathbb{E}_\eta\left[|m(\eta_t)(\varphi) - m(\eta) T_t \varphi|\right] \leqslant \frac{K_0}{N^\gamma}$$

Furthermore, if η is distributed according to the stationary distribution of the system η_N, then there exist $K_0 > 0$ and $\gamma > 0$ such that

$$\mathbb{E}\left[|m(\eta)(\varphi) - \nu(\varphi)\right] \leqslant \frac{K_0}{N^\gamma}$$

6 Particle Swapping

6.1 Glauber Dynamics

In statistical mechanics, the likelihood of observing a system in equilibrium in a particular state is related to its energy; states with low energy are more likely and states with high energy are less likely. In particular, the probability of a state i is proportional to $e^{-H(i)}$, where H is the Hamiltonian: the energy of a particular state. Distributions that assign probabilities in this manner are calleds **Gibbs Distributions**. A system for which such a state is the equilibrium, and for which the dynamuics are reversible, is called a **Glauber Dynamic** for that state. Recall the definition of reversibility:

Definition 6.1. *A Markov chain X_t is said to be **reversible** with respect to a distribution π if the backwards distribution of the chain, starting at any $T > 0$, is the same as the forwards distribution when both are started in distribution π. When X_t is has transition rates $P_t(x, y)$, we have*

$$\begin{aligned}\mathbb{P}(X_T = y | X_{T-t} = x)\mathbb{P}(X_{T-t} = x) &= \mathbb{P}(X_{T-t} = x | X_T = y)\mathbb{P}(X_T = y) \\ \implies P_t(x, y)\pi_x &= P_t(y, x)\pi_y \qquad \text{(Detailed Balance Equation)}\end{aligned}$$

We notice here that detailed balance implies stationarity. From this we get Kolmogorov's criteria; for $\{j_i\} \in S$,

$$\begin{aligned}&q(j_1, j_2)q(j_2, j_3)\cdots q(j_{n-1}, j_n)q(j_n, j_1) \\ &= q(j_1, j_n)q(j_n, j_{n-1})\cdots q(j_3, j_2)q(j_2, j_1)\end{aligned}$$

Furthermore,

$$\pi(j_n) = \pi(j_1)\frac{q(j_1, j_2)q(j_2, j_3)\cdots q(j_{n-1}, j_n)}{q(j_n, j_{n-1})\cdots q(j_3, j_2)q(j_2, j_1)}$$

Now take a finite state space $\mathbb{S}$, and a reference assignment $\nu \in (0,\infty)^{\mathbb{S}}$ of weights, not necessarily summable (for our purposes, ν will be uniform). We are given a Hamiltonian $H : \mathbb{S} \to [0,\infty)$ that satisfies

$$Z(\beta) \coloneqq \sum_{x\in\mathbb{S}} e^{-\beta H(x)}\nu_x < \infty \quad \forall \beta \in (0,\infty)$$

Z is called the partition function, and encodes all the relevant physical information of the underlying distribution; intuitively, the higher the temperature, the more likely high energy events are. The gibbs state γ_β is the probability vector where

$$\gamma(\beta)_x = \frac{1}{Z(\beta)} e^{-\beta H(x)}\nu_x$$

The Glauber dynamics corresponding to this distribution will not be unique, in particular, they depend on some communication matrix A that has non-negative entries, with diagonal entries of 0. Furthermore A will be irreducible:

$$\sup(A^n)_{x,y} > 0 \quad \forall (x,y) \in \mathbb{S}^2$$

and A will be reversible with respect to ν, so that

$$\nu_x A_{x,y} = \nu_y A_{y,x}$$

As a final condition, for the rates to be summable we need

$$\sum_{y\in\mathbb{S}} e^{-\beta H(y)} A_{x,y} < \infty$$

When ν is uniform, A denotes the states that are allowed to communicate.

Lemma 6.2 (Glauber Dynamics)**.** *Given a reference distribution ν, Gibbs measure $\gamma(\beta)$, and communication matrix A, the rates given by*

$$q(x,y) = e^{-\beta(H(y)-H(x))^+} A_{x,y}$$

are reversible with respect to $\gamma(\beta)$

Proof.

$$\begin{aligned}
\gamma(\beta)_x Q_{x,y} &= \frac{1}{Z(\beta)} e^{-\beta(H(x)+(H(y)-H(x))^+)}\nu_x A_{x,y} \\
&= \frac{1}{Z(\beta)} e^{-\beta(H(y)\vee H(x))}\nu_x A_{x,y} \\
&= \frac{1}{Z(\beta)} e^{-\beta(H(x)\vee H(y))}\nu_y A_{y,x} \\
&= \gamma(\beta)_y Q_{y,x}
\end{aligned}$$

■

6.2 Setup

Suppose that a nearest-neighbor chain X_t evolves according to the rates $q(x,y)$, or equivalently, has generator:

$$\mathcal{L}f(x) = \sum_{y \neq x} q(x,y)[f(y) - f(x)]$$

and is killed using soft killing rates $c(x)$. We recall from Cor 3.6 that $\mathcal{L}$ has adjoint

$$\mathcal{L}^* = \sum_{y \neq x} [f(y) - f(x)] + f(x) \sum_{y \neq x} [q(y,x) - q(x,y)]$$

Allowing ϕ and ψ to denote the QSD of this forward chain, as in Theorem 3.7—with Φ and Ψ denoting the corresponding energy potentials taking values in $\bar{\mathbb{R}}$—ϕ and ψ will be the unique solutions to:

$$-\mathcal{L}\psi(x) + c(x)\psi(x) = \lambda\psi(x) \tag{6.2.1}$$

$$-\mathcal{L}^*\phi(x) + d(x)\phi(x) = \lambda\phi(x) \tag{6.2.2}$$

The time reversal of X_t, say Y_t, is the chain with transitions $q(y,x)$. If Y_t is killed according to rate $d(x) = c(x) + h(x)$, where h is the 0th order term of $\mathcal{L}^*$, we see that the QSD of Y_t also solves (6.2.1). Then if $\{X_t^{(i)}\}_{i=1}^n$ and $\{Y_t^{(i)}\}_{i=1}^n$ denote IPS converging in measure to the QSD, as in Theorem (5.2), then we can pair particles with

$$(X,Y)_t^{(i)} \coloneqq (X_t^{(i)}, Y_t^{(i)})$$

so the law of (X,Y)—an abuse of notation for a chosen $(X,Y)^{(i)}$—converges weakly to the distribution $\psi(x)\phi(y) = e^{-(\Psi(x)+\Phi(y))}$.

6.3 Swapping Method

If we consider a pair of particles as an element in $S \times S$, then Lemma (6.2) tells us that, with uniform reference measure, the swaps

$$\begin{aligned} r_{x,y} \coloneqq q((x,y),(y,x)) &= e^{-(H((y,x))-H((x,y)))^+} A_{(x,y),(y,x)} \\ &= e^{-(\Psi(y)+\Phi(x)-\Psi(x)-\Phi(y))^+} A_{(x,y),(y,x)} \end{aligned}$$

are reversible under stationarity for any A that is reversible with respect to ν, which in this case is uniform, so $A_{i,j} = A_{j,i}$. These rates do not change the marginal distributions of the individual particles under stationarity.

A priori, these rates are not known, and involve the QSDs themselves. However, the forward and backward control problems allow for computation of these rates without needing the distributions.

Proposition 6.3. *Nearest-neighbor rate matrices satisfying a particular stationary distribution are unique up to multiples of the rates.*

Theorem 6.4 (Particle Swapping Rates)**.**

$$e^{-(\Psi(y)+\Phi(x)-\Psi(x)-\Phi(y))^+} = \left[\prod_1^n \frac{q(z_{i+1},z_i)}{q(z_i,z_{i+1})}\right] \vee 1 = \frac{\pi(y)}{\pi(x)} \vee 1$$

Where $\{z_i\}_1^n$ *is a path with positive probability from x to y. Or equivalently, a sequence with* $q(z_i,z_{i+1}) > 0\ \forall i \in [n-1]$ *where* $x = z_1$ *and* $z_n = y$*.*

Proof. By Theorems 4.2 and 4.4, the rates $b^*(x,y)$ and $b'(x,y)$ are nearest-neighbor rates with the same stationary distribution, $\phi(x)\psi(x)$. Since such rates are unique up to constant multiples (by the above proposition), and the total rates are the same:

$$b^*(x,x) = b'(x,x) = \lambda + q(x,x)$$

The rates are the same, so for any x that communicates with y in the original chain, neither being hard killed states (nearest neighbors),

$$\begin{aligned} b^*(x,y) &= q(x,y)e^{\Psi(x)-\Psi(y)} \\ = b'(x,y) &= q(y,x)e^{\Phi(x)-\Phi(y)} \end{aligned}$$

so

$$e^{-(\Psi(y)+\Phi(x)-\Psi(x)-\Phi(y))} = \frac{q(y,x)}{q(x,y)}$$

By telescoping along these states, we have the first equality, and the second equality just follows from the Kolmogorov criteria.

■

In principle, this would work for any irreducible chain for which we can make the same uniqueness argument, and there are adjustments to the reference measure and communication matrix that can be made, perhaps to improve convergence.

6.4 Consistency of the Swapped System

The argument for the asymptotic consistency of the typical Fleming-Viot particle system relied on a propagation-of-chaos argument. After showing that the system mixed sufficiently well, it was shown that the particles were asymptotically uncorrelated, so that each particle, when resampled, effectively sees the averaged dynamics. Then from a bias-variance argument, the consistency was reached. We will extend the consistency to the Fleming-Viot estimator through a much simpler topological argument, which will almost directly follow from Corollary 5.9. Then, we will present a bound on the exponential ergodicity of the swapped system.

Lemma 6.5. *If* $(\eta,\mathfrak{N})_{A,N}$ *is distributed according to the invariant distribution of a N-particle swapped Fleming-Viot system with swapping rate A, then there exists a* $A_N > 0$*, such that* $A_N \to \infty$ *and for any sequence* $B_N \leqslant A_N$

$$\lim_{N\to\infty} \mathbb{E}\left[|m(\eta_{B_N,N})(\varphi) - \nu(\varphi)|\right] = 0$$

for any $\|\varphi\|_\infty \leqslant 1$*.*

6.5 Numerics

To demonstrate the efficacy of the swapping method, we will show that for a simple particle system with known quasi-stationary distribution and clear metastability issues, the swapped system explores the state space more, holding computations fixed, and seems consistent in finding the QSD. Let X_t be an interpolated diffusion process over $[1, 1]$ satisfying the SDE (see 8.2)

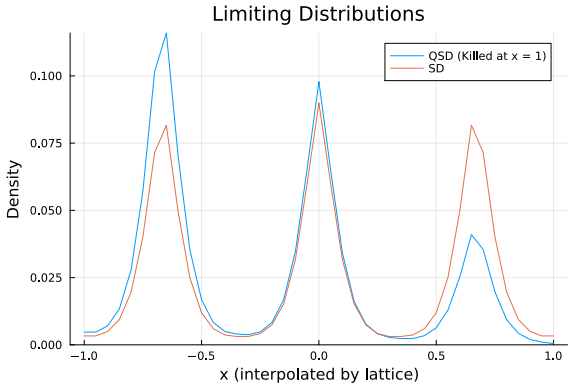
$$dX_t = U(X_t)dt + dW_t$$

where

$$U(x) = -11 \cdot \sin(3\pi x)$$

with a mesh size of $h = 0.05$ and reflective boundary conditions. Then using linear algebra (as discussed in 3.8) we can compute the quasi distribution for X_t if it is killed at state $\partial = \{1\}$ (Figure 1).

Figure 1: Gibbs distributions for the interpolated diffusion process.

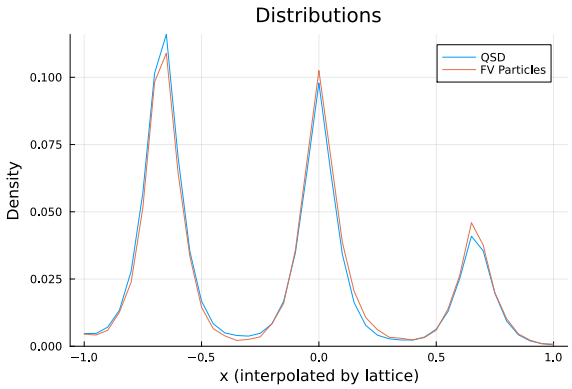


If the swapping multiplier is set to $A = 1$, then using the typical estimator for the QSD from the Fleming-Viot particles, averaged over time, our numerics indicate that for large T, the estimator appears to converge to the QSD (Figure 2).

6.5.1 Well Method

To determine whether this is, in a suitable sense, a better estimator for the QSD, we will measure how much individual particles explore the state space. The way we do this is by defining "wells" of the state space to be places where the dynamics make it hard for particles to escape, which corresponds to peaks of ϕ or troughs of Φ. To avoid doing this by hand, we set a cutoff for the energy potential by some average value over the states, and find the connected components of the states with potential lower than this. These components are seen in Figure 3: a contiguous red line is a connected component, and a red dot is a state in that component.

Figure 2: Consistency of the swapped Fleming-Viot particles.



In order to determine whether a particle has traversed between wells, we label particles by their last explored well, and note a "well-to-well jump" when a particle moves into a well that is different than their last explored well. Normalizing the number of well-to-well jumps by the computations performed, we notice that the swapped system moves between wells at a significantly higher rate, sometimes 3 to 4 times as often (Figure 4).

It should be noted that the dynamics of a swapped system comprise a smaller proportion of computations than a non-swapped system, so normalizing by computations, in a sense, favors the non-swapping system.

Figure 3: Construction of wells.

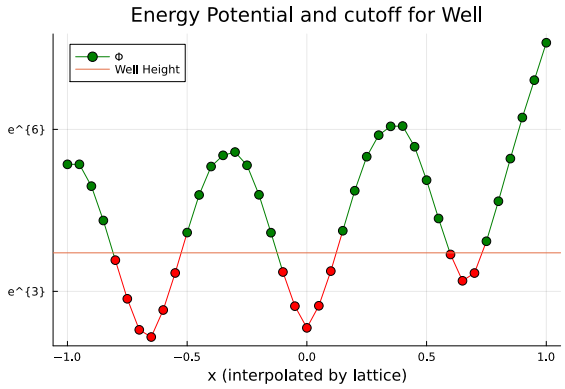
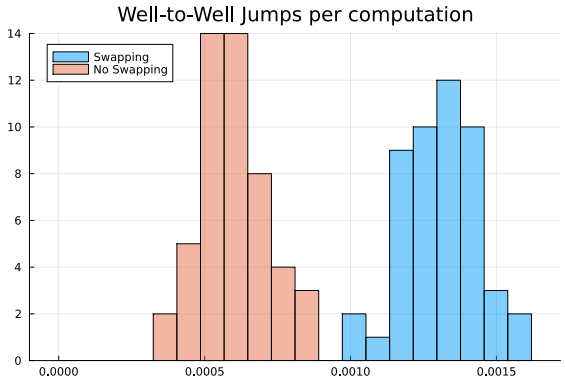


Figure 4: Comparison of the frequency of well-to-well jumps between a typical Fleming-Viot system and a swapped system.



7 Infinite Swapping

7.1 Two-Time-Scale Markov Chains

The theory of two-time-scale Markov chains is based on a chain whose generator is

$$Q_\varepsilon = \frac{1}{\varepsilon}\widetilde{Q} + \widehat{Q} \tag{7.0.1}$$

As $\varepsilon \to 0$, the dynamics of $\widetilde{Q}$ dominate the chain over time scales of order ε, while for order 1 time intervals, the dynamics of $\widetilde{Q}$ are averaged, and $\widehat{Q}$ drives the dynamics. Lets formalize what we would expect this to look like. First, we impose a condition on $\widetilde{Q}$

Definition 7.1. *A rate matrix $\widetilde{Q}$ is said to be **locally irreducible** if $\widetilde{Q}$ is block diagonal by irreducible sub-matrices, possibly after permuting indices. In other words, $\widetilde{Q}$ has a partition $\{S_k\}$ by non-communicating irreducible classes.*

Let $|\{S_k\}| = l$ and $|S_k| = m_k$. Then when ε is small, the average amount of time that the chain spends in the ith element of $|S_k|$ before evolving by $\widehat{Q}$ is ν_i^k, where ν^k is the stationary distribution of $\widetilde{Q}^k$, the S_k minor of $\widetilde{Q}$. We would expect the dynamics of $\widehat{Q}$ between irreducible classes to be, reordering the indices so that $\widetilde{Q}$ is block-diagonal,

$$\bar{Q} = \begin{pmatrix} \nu^1 & & & \\ & \nu^2 & & \\ & & \ddots & \\ & & & \nu^l \end{pmatrix} \widehat{Q}\tilde{\mathbb{1}}$$

where

$$\tilde{\mathbb{1}} = \begin{pmatrix} \mathbb{1}_{m_1} & & & \\ & \mathbb{1}_{m_2} & & \\ & & \ddots & \\ & & & \mathbb{1}_{m_l} \end{pmatrix}$$

So if ϑ^k is the marginal representing the time spent in a particular class, we expect ϑ to evolve by $\bar{Q}$:

$$\frac{d}{dt}\vartheta(t) = \vartheta(t)\bar{Q}, \quad \vartheta^k(0) = \mu(S_k)$$

and so the probability that a chain starting with distribution μ is at site (k, j) at time t, where k is the irreducible class and j is an index of the class, is approximately $\nu_j^k \vartheta^k(t)$.

The following theorem is Corollary 4.31 of [4]. It is a generalization of a much more general theorem on the distribution of possibly inhomogenous two-time-scale Markov chains, which approximates the distribution to higher order, depending on the regularity of the semigroup.

Theorem 7.2. *Suppose that, for a Markov chain (X_t) over finite state space S is locally irreducible through a partition $\{S_k\}$, then there exist $K, \kappa_0 > 0$ such that*

$$|\mu P_t^\varepsilon \mathbb{1}_{(k,j)} - \nu_j^k \vartheta^k(t)| \leqslant K\left(\varepsilon(t+1) + \exp\left(-\frac{\kappa_0 t}{\varepsilon}\right)\right) \tag{7.2.1}$$

where $\mathbb{1}_{(k,j)}$ is one in class k and index j and zero otherwise, $\mathrm{Law}(X_0) = \mu$, *and P_t^ε is the semigroup of Q_ε.*

7.2 Infinite Swapping Limit

The last theorem directly applies to our swapping case. It is important to notice that as defined, our swapping dynamics do not have well-defined distributions. This is because as A, the swapping parameter from Lemma 6.2, gets large, the pairs of particles oscillate quickly between two states; there is no well-defined dynamic, as the sample paths become erratic as $A \to \infty$.

To fix this issue, we swap the dynamics of a particle pair, rather than their locations. Formally, we define new random variables $S_t^{(n)}$ that take values over $\{1, -1\}$. $S_t^{(n)}$ will represent whether the particle $X_t^{(n)}$ is a forwards or a backwards particle, so that, for instance, when $S_t^{(n)} = 1$, $X_t^{(n)}$ is a forward particle and is in a forward Flemming-Viot system with all other $X^{(k)}$ particles for which $S_t^{(k)} = 1$ and all other $Y^{(k)}$ particles for which $S_t^{(k)} = -1$. These are clearly the same processes, just with different "accounting".

This system is exactly a two-time-scale Markov chain where the high-frequency process is a locally irreducible process. To be precise, the elements of Q^ε represent transitions in the entire configuration of the swapping system, and thus can be indexed by the positions of all particles and the configuration of $S^{(n)}$; in other words we index over

$$F = \{\varsigma : 2N \to D \times \{1, -1\} \mid \varsigma_2(n) = -\varsigma_2(n + N \mod 2N)\}$$

The irreducible minors of $\widetilde{Q}$ are just 2×2 matrices that represent swaps between forward and backward configurations for one pair at one location. Since the swapping dynamics do not change the positions of the particles, but rather the polarity S, the quotient process $\bar{Q}$ is well-defined as a dynamic over the quotient (X, Y). It is easiest to parameterize the partition by two indices, the first representing the pair of particles (with N possibilities) that is to be moved, and the second representing the location of each particle (with $|D|^2$ possibilities):

$$\widetilde{Q}^{(n,(x,y))} = \begin{pmatrix} -r_{x,y} & r_{x,y} \\ r_{y,x} & -r_{y,x} \end{pmatrix}$$

This way of describing the dynamics is inefficient from a combinatorial perspective, since many of the swapping and evolution values are duplicated (swapping rates for pairs do not depend on the other particle's positions, for instance).

Invoking Theorem 7.2, we see that sending $\varepsilon \to 0$, the (X, Y) dynamics of the system (X, Y, S) converge to the following system:

Algorithm 7.3 (Infinite-Swapping System)**.** *Let*

$$\nu_{x,y} = \frac{r_{x,y}}{r_{x,y} + r_{y,x}}$$

Then the infinite swapping limit of the swapped Fleming-Viot system has the following rates

- *If* $(X,Y)^{(n)} = (x,y)$ *then X moves through its dynamics to z with rate*

$$\nu_{x,y} Q_{x,z} + \nu_{y,x} Q_{z,x}$$

.

- *If* $(X,Y)^{(n)} = (x,y)$ *then Y moves through its dynamics to z with rate*

$$\nu_{x,y} Q_{z,y} + \nu_{y,x} Q_{y,z}$$

.

- *If* $(X,Y)^{(n)} = (x,y)$*, then X is killed with rate*

$$\nu_{x,y} c(x) + \nu_{y,x} d(y)$$

Once killed, it chooses another pair uniformly, and then, if the pair is at positions (q,p)*, it moves to q with probability* $\nu_{q,p}$ *and p with probability* $\nu_{p,q}$

- *If* $(X,Y)^{(n)} = (x,y)$*, then Y is killed with rate*

$$\nu_{x,y} d(y) + \nu_{y,x} c(x)$$

Once killed, it chooses another pair uniformly, and then, if the pair is at positions (q,p)*, it moves to q with probability* $\nu_{p,q}$ *and p with probability* $\nu_{q,p}$

And the following is an asymptotically consistent estimator of the QSD ϕ*:*

$$\phi(dx) = \frac{1}{T} \int_{t=0}^{T} \sum_{n=1}^{N} \nu_{X_t^{(n)}, Y_t^{(n)}} \, \delta_{X_t^{(n)}}(dx) + \nu_{Y_t^{(n)}, X_t^{(n)}} \delta_{Y_t^{(n)}}(dx) \, dt$$

8 Applications

8.1 Standard Markov-Chain Monte Carlo

The simplest application of this swapping system is in performing MCMC on a normal Gibbs measure. The rationale for using such a system is simple: we can use the reversed dynamics of a system to reduce the metastability issues of a Metropolis-Hastings estimator without knowing any additional structure about the chain. The construction through QSD was necessary to formalize how the backward dynamics interacted with the Fleming-Viot system. A Fleming-Viot system that approximates a stationary distribution is just a set of N independent particles; by introducing the backwards system, we allow these particles to interact in a way that is constructive in finding the distribution.

8.2 Diffusions and their Discrete-state Approximations

Quasi-stationary distributions may exist in absorbed processes whose state space may be finite, countable, or uncountable and whose time indices may be countable or uncountable. In most cases, the diffusion processes—a subset of continuous state continuous-time Markov processes whose paths are continuous—are the easiest to manipulate, despite the relatively steep theory. For this reason, it will be prudent to consider the effects of our improvements on estimators of QSDs on these spaces. When we computationally model these effects, discrete-state approximations will be critical.

The **SDE**,

$$dX_t = -X_t dt + dW_t, \quad X_0 = 0$$

where W_t is a **Weiner process**, describes a standard **Ornstein–Uhlenbeck process**. This differential notation represents integrals on time so that $X_t = \int_0^t -X_s ds + W_t + X_0$. One can see that such a process has two terms: a symmetric diffusion term, and a term that pushes the process back to the origin.

Our motivation for such a process is simple: although a Weiner process is always recurrent—it can be shown that for any $\epsilon > 0$, W_t takes strictly positive and negative values on $(0, \epsilon)$—its discrete-time analog is not. Adding an inward drift term makes it positive recurrent, which is important for existence of a stationary distribution in the diffusion process.

Proposition 8.1. *It can be shown, using variation of parameters, that the explicit equation for this process is*

$$X_t = X_0\, e^{-t} + \int_0^t e^{s-t} dW_s$$

Proof. First, we notice that

$$\dot{X}_t^h + X_t^h = 0 \Rightarrow X_t^h = Ce^{-t}$$

So, by variation of parameters, letting $X_t = v(t)e^{-t}$ yields:

$$v'(t)e^{-t} - v(t)e^{-t} + v(t)e^{-t} = \frac{dW_t}{dt}$$

$$\int_0^t v'(t)dt = \int_0^t e^s \frac{dW_s}{ds} ds$$

$$v(t) - v(0) = \int_0^t e^s dW_s$$

$$X_t = v(0)e^{-t} + \int_0^t e^{s-t} dW_s, \quad v(0) = X_0$$

■

Returning to the concept of generators, it can be shown that the generator for our standard Ornstein-Uhlenbeck process is $\mathcal{L}f(x) = -xf'(x) + \frac{1}{2}f''(x)$. To find a discrete-state analog of the diffusion process, we consider jumps on the lattice $h\mathbb{Z}$ given by the process $\{X_t^h\}$ and design jump rates so that $\mathcal{L}^h f \approx \mathcal{L}f = -xf'(x) + \frac{1}{2}f''(x)$. To this end, we use difference estimates:

$$f'(x) \approx \begin{cases} \frac{f(x+h)-f(x)}{h} \\ \frac{f(x)-f(x-h)}{h} \end{cases}$$
$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

The expression for $f''(x)$ is easily verified using l'Hopital and $f \in C^2$. Using Prop 3.5 and the generator for our diffusion process, we obtain new jump rates. We split our $f'(x)$ term into two cases to ensure that we are left with a generator that represents a legitimate process in the form $r^h(x,y)[f(y) - f(x)]$.

$$\begin{aligned} \mathcal{L}^h f(x) = &-\mathbb{1}_{x>0}\frac{hx}{h^2}\left[f(x) - f(x-h)\right] \\ &- \mathbb{1}_{x<0}\frac{hx}{h^2}\left[f(x+h) - f(x)\right] \\ &+ \frac{1}{2}\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \\ \Downarrow& \\ r^h(x, x \pm h) = &\frac{1}{h^2}\left[\mp xh\mathbb{1}_{\{\mp x>0\}} + \frac{1}{2}\right] \end{aligned}$$

Here, when x is negative, we ensure the generator takes the required form by using the $[f(x+h) - f(x)]$ expression to ensure our rates will be positive ($-x \geqslant 0$). This process is called **upwinding**.

Staying in our discrete-state process, we have seen that we can manipulate the expression $\langle g, \mathcal{L}f\rangle = \langle \mathcal{L}^* g, f\rangle$ to find the ℓ^2 adjoint for $\mathcal{L}$, which for discrete-states is characterized:

$$\mathcal{L}^* f(x) = \sum_{y\in\mathcal{X}} q(y,x)[f(y) - f(x)] + f(x)\sum_{y\in\mathcal{X}}[q(y,x) - q(x,y)]$$

where q is the rate matrix on the underlying process.

9 Other Possible Improvements

9.1 Change of Reference Measure

To improve the approximations, we could use a reference measure *other* than the regular Lebesgue/uniform measure, and will then find the Lebesgue/uniform adjoint of that new operator, which will be $(\mathcal{L}_\rho^*)^*$. After simulating Fleming-Viot particles on this new process (we saw that Generators characterize processes from Lemma 2.10) renormalization will yield an approximation of the QSD of the original process.

In a general setting, for any reference measure $\rho(dx)$ that is absolutely continuous with respect to the lebesgue measure, so that $\rho(dx) = \rho(x)dx$, we will have:

$$\mathcal{L}^*_\rho f(x) = \frac{1}{\rho(x)}\mathcal{L}^*(f\rho)(x)$$

and if $\widetilde{\varphi}$ and φ represent eigenfunctions in $L^2(\rho)$ and L^2 respectively, then $\widetilde{\varphi} = \varphi/\rho$ [1]. We have a proof of both statements, as well as a formulation of the adjoint in the finite-state case, below.

Proof. We recall the adjusted inner product

$$\langle f(x), g(x)\rangle_\rho = \int f(x)g(x)\rho(dx) = \sum_{x\in S} f(x)g(x)\rho(x)$$

The first statement follows; take any test function $g(x)$, then

$$\langle \mathcal{L}^*_\rho f, g\rangle = \langle \mathcal{L}^*_\rho f, \frac{1}{\rho}g\rangle_\rho = \langle f\rho, \mathcal{L}\frac{1}{\rho}g\rangle = \langle \frac{1}{\rho}\mathcal{L}^*(f\rho), g\rangle$$

Furthermore, if ψ is an eigenfunction of $\mathcal{L}^*$

$$\mathcal{L}^*_\rho\left(\widetilde{\psi}\right) = \mathcal{L}^*_\rho\left(\frac{\psi}{\rho}\right) = \frac{\mathcal{L}^*(\psi)}{\rho} = \frac{\psi}{\rho} = \widetilde{\psi}$$

In the finite state case, with rates $q(x,y)$, we use this expression to find the general adjoint, and manipulate the expression to find $\mathcal{L}^*_\rho$. Continuing with the aforementioned inner product,

$$\begin{aligned}
&= \sum_{x\in S}\left[\sum_{y\in S} q(x,y)g(y)\right] f(x) \\
&= \sum_{y\in S}\sum_{x\in S} q(x,y)f(x)g(y) \\
&= \sum_{x\in S}\left[\sum_{y\neq x} q(y,x)f(y) - \sum_{y\neq x} q(x,y)\rho(x)\right] g(y) \\
&= \sum_{x\in S}\left[\sum_{y\neq x} q(y,x)[f(y)-f(x)] + f(x)\sum_{y\neq x}[q(y,x)-q(x,y)]\right] g(y)
\end{aligned}$$

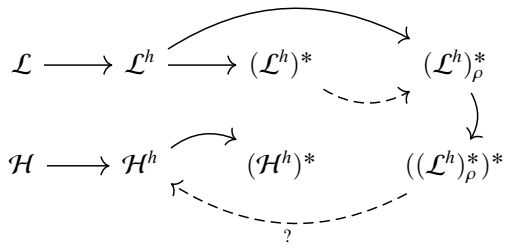
and we see the expression of $\mathcal{L}^*$. Then for a reference measure ρ, our earlier conclusion leads to:

$$\mathcal{L}^*_\rho f(x) = \frac{1}{\rho(x)}\left[\sum_{y\neq x} q(y,x)[f(y)\rho(y) - f(x)\rho(x)] + f(x)\sum_{y\neq x}[q(y,x)-q(x,y)]\right]$$

■

Here, it is important to note the distinction between operators and generators. Not all operators can be represented as generators of a Markov chain, and although all generators are operators, heuristically, we interpret generators as objects that encode information about some underlying process (specifically, the expected infinitesimal change in functions with time).

In the context of a diffusion, to see how the dual generator in effect "reverses" the dynamics of the process, and the effects of our change of reference measures, we could study the following relationship:



Here, our new operator $\mathcal{H}$, represents the generator for the reversed diffusion process: $\mathcal{H}f(x) = xf(x) + \frac{1}{2}f''(x)$, a process that, trivially, has no limiting distribution. For this uni-modal distribution, we could look at three reference measures:

(1) Lebesgue measure: $\rho(dx) = dx$. Here, our process will not change, of course, and the dynamics will still push inwards.

(2) Stationary distribution: $\rho(dx) = e^{-\frac{x^2}{2\epsilon}}dx$. Here, we expect the dynamics of our new process to be reversed ($\approx \mathcal{H}$). The increased frequency of killings and consolidation of particles towards the boundary is expected to improve the Fleming-Viot approximation.

(3) Weakened version of stationary distribution: $\rho(dx) = e^{\frac{-x^2}{\epsilon}}dx$. Here, we expect our new dynamics to be a middle ground between $\mathcal{L}$ and $\mathcal{H}$, a Weiner process with no drift.

9.2 Associated Markov Chain

We now consider the uniform adjoint of the reference-measure adjoint (which is also the operator whose adjoint is $\mathcal{L}^*_\rho$ as the adjoint operator is an involution) for the finite-state CTMC. Using the identity $\langle \mathcal{L}^*_\rho g(x), f(x)\rangle = \langle g(x), (\mathcal{L}^*_\rho)^* f(x)\rangle$, and the proof from the last section, we compute the second adjoint as

$$\begin{aligned}\langle \mathcal{L}^*_\rho g(x), f(x)\rangle &= \sum_{x\in S}\sum_{y\in S} q(y,x)\frac{\rho(y)}{\rho(x)}g(y)f(x) \\ &= \sum_{y\in S}\left[\sum_{x\in S} q(y,x)\frac{\rho(y)}{\rho(x)}f(x)\right]g(y)\end{aligned}$$

$$
\begin{aligned}
\implies (\mathcal{L}_\rho^*)^* f(x) &= \sum_{y\in S} q(x,y)\frac{\rho(x)}{\rho(y)} f(y) \longleftarrow \textit{Matrix form} \\
&= \sum_{y\neq x} q(x,y)\frac{\rho(x)}{\rho(y)} f(y) - \sum_{y\neq x} q(x,y) f(x) \\
&+ \sum_{y\neq x} q(x,y)\frac{\rho(x)}{\rho(y)} f(x) - \sum_{y\neq x} q(x,y)\frac{\rho(x)}{\rho(y)} f(x) \\
&= \sum_{y\neq x} q(x,y)\frac{\rho(x)}{\rho(y)} [f(y) - f(x)] + f(x) \sum_{y\neq x} q(x,y) \left[\frac{\rho(x)}{\rho(y)} - 1\right] \\
&= \sum_{y\neq x} q(x,y)\frac{\rho(x)}{\rho(y)} [f(y) - f(x)] + \left(\sum_{y\in S} q(x,y)\frac{\rho(x)}{\rho(y)}\right) f(x) \\
&:= \mathcal{L}_\rho f(x) + h_\rho(x) f(x)
\end{aligned}
$$

Here, $\mathcal{L}_\rho$ is the generator of the process with rates $q(x,y)\rho(x)\rho(y)^{-1}$, and h_ρ is a 0th order term, so the eigenvalue problem, which we have shown to be equivalent to the original QSD after renormalization, can (**??**) be solved using a Flemming-Viot scheme.

9.3 Speedup

It can be shown that for non uniform $\rho' = \rho^{-1}$, $\exists x$ such that $h_{\rho'}(x) < 0$. Take $x = \arg\max_z \rho(z)$. Then since ρ is irreducible

$$
\begin{aligned}
(Q\rho)(x) &= q(x,x)\rho(x) + \sum_{y\neq x} q(x,y)\rho(y) \\
&= \sum_{y\neq x} q(x,y)(\rho(y) - \rho(x)) < 0
\end{aligned}
$$

So the conclusion holds, as

$$
\begin{aligned}
h_{\rho'}(x) < 0 &\iff \sum_{y\in S} q(x,y)\frac{\rho(y)}{\rho(x)} < 0 \\
&\iff Q\rho < 0 \text{ at } x
\end{aligned}
$$

In order to perform Flemming-Viot when the 0th order term is not non-negative, there are two main approaches. One involves introducing particles, rather than killing them, at rates determined by the negative positions/magnitudes of the 0th order term. Another is to add a constant to the 0th order term, which requires a correction to the computed killing rate.

References

[1] Amarjit Budhiraja, Paul Dupuis, Pierre Nyquist, and Guo-Jhen Wu. Quasistationary distributions and ergodic control problems. *Stochastic Processes and Their Applications*, 2021.

[2] Bertrand Cloez and Marie-Noémie Thai. Quantitative results for the fleming-viot particle system and quasi-stationary distributions in discrete space. *Stochastic Processes and their Applications*, 2014.

[3] Pierre Collet, Servet Martínez, and Jaime San Martín. *Quasi-Stationary Distributions: Markov Chains, Diffusions and Dynamical Systems*. Springer, 2013.

[4] Qing Zhang George Yin. *Continuous-Time Markov Chains and Applications: A Two-Time-Scale Approach*. Springer, 2013.

[5] Hao Liu. Quasi-stationary distribution, ergodic control problems, and a jacobi type iteration, 2022.

[6] Eugene Seneta. *Non-negative Matrices and Markov Chains*. Springer, 1981.

[7] Daniel W. Strook. *An Introduction to Markov Processes*. Springer, 2014.

[8] Denis Villemonais. General approximation method for the distsribution of markov processes conditioned not to be killed. *ESAIM Probability and Statistics*, 2014.