Convergence of Stratified MCMC Sampling of Non-Reversible Dynamics

Gabriel Earle⋆ and Jonathan Mattingly†

⋆Department of Mathematics, University of Massachusetts-Amherst
†Departments of Mathematics and of Statistical Science, Duke University

February 21, 2022

Abstract

We present a form of stratified MCMC algorithm built with non-reversible stochastic dynamics in mind. It can also be viewed as a generalization of the exact milestoning method, or form of NEUS. We prove convergence of the method under certain assumptions, with expressions for the convergence rate in terms of the process’s behavior within each stratum and large scale behavior between strata. We show that the algorithm has a unique fixed point which corresponds to the invariant measure of the process without stratification. We will show how the speeds of two versions of the new algorithm, one with an extra eigenvalue problem step and one without, relate to the mixing rate of a discrete process on the strata, and the mixing probability of the process being sampled within each stratum. The eigenvalue problem version also relates to local and global perturbation results of discrete Markov chains, such as those given by Van Koten, Weare et. al.

1 Introduction

Markov Chain Monte Carlo (MCMC) is an often used method to produce samples from a distribution, when a Markov kernel converging to that distribution is known. Stratification of MCMC methods is a well-studied form of rare event sampling. Cases of interest include systems where regimes of low probability have outsize importance, or systems with multiple regimes of high probability but rare transitions between them. In such cases, the sample space can be broken up into smaller parts, called strata, and a Markov Chain can be run within each. The results are then combined in some way, allowing for an estimate of a distribution on the whole space which can be obtained more quickly than via a simple MCMC method. There are many advantages of dividing the pace into strata. Including the fact that many of the computations can be run in parallel and computational resolution and effort can be concentrated in regions of interest.

In this article, we present and prove convergence results for a specific stratified MCMC scheme close to those in [4]. The algorithm, which we call the “injection measure method”, constructs an estimate of the invariant measure of a Markov chain, and is built specifically with non-reversible Markov chains in mind. It can also be viewed as a version of Non-Equilibrium Umbrella Sampling (NEUS), as in [4], or as an extension of the exact milestoning method, as detailed in [1]. Non-reversible Markov chains sampling problems typically arise in two settings. First, when one samples an invariant measure which does not satisfy detailed balance; and hence, has a nontrivial flux through the system in its stationary
state. Such stationary states are often referred to as non-equilibrium steady-states. Second, when one samples the non-reversible Markov chain can be obtained by adding time to the dynamics as one of the state-variables. Because of direction of time ensures that the system has a non trivial flux. As illustrated in [4], the resulting space-time dynamics can be used to study out-of-equilibrium transitions rates and other transient phenomenon. See [4] for more details.

The key object of the method, and our analysis of it, is a collection of distributions on the strata known as injection measures. These estimate how particles following the Markov chain are likely to be distributed on the step when they enter one stratum from another. Also associated to each stratum is a corresponding weight, estimating how likely particles are to enter the stratum, relative to the other strata. If the injection measures and weights were known exactly, they could be used to compute estimates of the invariant measure within each stratum, and patch those estimates together with the correct weights. In practice, the injection measures and weights will most likely not be known, and so they must themselves be estimated. We propose estimating them iteratively, finding new injection measures via trajectories started from the current ones. This iteration is the main step of our formulation and those discussed in [4].

The corresponding weights can be calculated in two ways, leading to two versions of the injection measure method. In the first, the weights are found by applying a transition matrix determined by the measures to the previous weights. We call this form of the method the basic version. In the other, the principle eigenvector of the matrix is found, and its entries are taken to be the new weights. This is called the eigenvector version. The eigenvector version performs strictly better in our numerical experiments, but we as yet only have a proof of local convergence for it, whereas we prove global convergence of the basic version.

In order to prove our main results, we need some assumptions about the behavior of the Markov chain being sampled and how it interacts with the chosen strata. In addition to some standard regularity assumptions, we will need to assumptions on two types of behavior. The first is a “microscopic” assumption, governing how particles following the chain move within a stratum. In effect, our assumption will be that any two particles starting somewhere in the same stratum have some chance of being coupled at or before the time of exit. We will also need a “macroscopic” assumption, about how the chain moves overall mass between strata. We will have several forms of this, depending on the version of the algorithm and precision of the theorem we wish to prove. However, each form of the assumption roughly states that the chain moves mass between the various strata at a suitable rate, or with enough regularity. The algorithm’s overall convergence speed can then be expressed in terms of the “microscopic” coupling probability and the “macroscopic” rate. That is the substance of our main theorems.

Specifically, our first result shows that, under the assumptions, the algorithm has a unique fixed point, and the injection measures given by the fixed point are correspond to the original invariant measure. Next, we prove that, if $c$ is the microscopic coupling probability and $\lambda$ is the macroscopic rate of convergence, then the algorithm converges to the fixed point in total variation at a rate that roughly looks like

$$O(\max(\sqrt{1-c}, \lambda)).$$

We also show that, for a constant $r$ reflecting the sensitivity of the weights to the entries of a macroscopic transition matrix, the eigenvector version converges, for sufficiently good starting estimates, at a rate approaching

$$1 - \frac{c}{1 + r}.$$ 

More precisely, we will show that, if $G$ is the matrix of transition rates between strata for the dynamics in equilibrium, then $r$ bounds the relative sensitivities of the invariant distribution of $G$ to small changes in the entries of $G$.

The structure of this article will be as follows. In Section 2, we give the basic notation needed for our results, and state the two versions of the injection measure method explicitly. Section 3, we outline the assumptions needed and state the main theorems we can now prove under them. In Section 4, we
give some results of numerical simulations of the method on a simple example system, and relate them to some of our theoretical results. In Section 5 we give the proofs of our theorems in detail.

2 The Injection Measure Method

2.1 Intuition Behind the Algorithm

To motivate the steps of the method, consider a space broken up into $J$ subsets, or strata, and a process $X^j_n$, following a kernel $P$ and confined to the $j$-th stratum. If we wish to sample the invariant distribution restricted to that stratum, then the question becomes what to do if a step drawn according to $P$ would have $X^j_n$ leave the stratum. One option would be to simply “bounce it back” from the boundary, which would mean setting $X^{n+1}_j = X^n_j$. We could then keep track of how many attempted exits occur between the strata and use that to form a matrix $G$ on the set of strata, or $\{1, \ldots, J\}$. The empirical distributions of the $X^n_j$ can then be combined, with weights given by the invariant distribution of $G$.

The above scheme works under a key assumption: that the kernel $P$ is reversible. The reason this is needed is that “bouncing a particle back” when it tries to leave a stratum is like saying that another particle comes to replace it at the exact location it left from. The assumption that this occurs for many particles in equilibrium is exactly the assumption that detailed balance holds, at least along the strata boundaries. But we are precisely interested in the non-reversible case. Therefore, we must have a better idea of how, and if, a particle leaving a stratum is replaced.

Thus we propose building an injection measure for each stratum, which captures how particles entering that stratum are distributed. In addition, we need weights that capture how many particles enter each stratum, relative to the other strata. Of course, we cannot expect these to reflect how particles entering a stratum are in equilibrium, at first. Therefore, we will use the starting injection measures to calculate the distributions of particles leaving each stratum, called exit measures. Then, for each $j$, we look at the distribution of all particles leaving other strata, via their exit measures, and entering stratum $j$. We use this to form a new injection measure and weight for the $j$-th stratum, for all $j$. Thus we have an iterative procedure, building new injection measures and weights from old ones. We wish to show that the injection measures thus defined converge over time to some fixed point, called the equilibrium injection measures, and that they give us an estimate of the invariant measure of $P$. We call the method thus outlined the basic version of our algorithm.

There is one further step we can add to the method. Just as in the reversible case, we can make a transition matrix $G$ based on how many particles leave each stratum into each other one, or how much mass the $k$-th exit measure gives to the $j$-th stratum. Then we can replace the old weights on each iteration by the invariant vector of $G$. We call the method, with this added step, the eigenvector version of the injection measure method. We will be able to show, under certain assumptions, an improved bound for the rate of convergence for this version, with the drawback that we have not yet shown that it converges for any starting injection measures. That is, our theorems only show that it converges to the correct distribution locally.

2.2 Setting and Notation

Suppose we are interested in sampling from the invariant distribution $\pi$ of a Markov kernel $P$, defined on a state space $A \subset \mathbb{R}^d$. The simplest version of the injection measure method of stratified sampling is to break $A$ into subsets, or strata, $A_1, \ldots, A_J$ which partition the space. One then runs the Markov chain $P$, starting from a measure $\nu^0_j$ concentrated with in each strata $A_j$, until it exist. The algorithm calculates new starting measures $\{\nu^\ell_j\}$ biased on the exists from the collection of $\{A_j\}$. This process, which is iterated until the measure $\{\nu^\ell_j\}$ converge also provides a collection of weights $\{a_j\}$ and occupation measure $\pi_j$ in each $A_j$ so that the desired target measure $\pi = \sum a_j \pi_j$. 

It the above setting is clear which of the strata \( A_j \) contain the process at any given times since they are disjoint. However, we are also interested in the case where there is overlap between the strata, and the point at which a particle is declared to have left one stratum and entered another is possibly random. Though we will consider a more general formulation, for the moment consider the following illustrative setting.

Consider the following setup: we have strata \( A_1, \ldots, A_J \) covering \( A \) as before, but the strata are no longer disjoint. We also assume we have a partition of unity \( \psi_1, \ldots, \psi_J \) whose support coincides with the \( A_1, \ldots, A_J \). Namely \( \text{supp}(\psi_j) = A_j \) and for all \( x \in A_j \), \( \psi_j(x) \geq 0 \) and

\[
\sum_j \psi_j(x) = 1.
\]

When a particle enters \( A_j \) at \( X_0 \sim \nu_j \), a value \( \kappa \) is chosen from some distribution \( \eta \) on \([0, 1]\). The particle then moves according to \( P \), with its position at time \( n \) being \( X_n \sim P(X_{n-1}, \cdot) \). When \( \psi_j(X_n) < \kappa \), the particle is declared to have exited \( A_j \). The index of the stratum it exits into is chosen from \( 1, \ldots, j-1, j+1, \ldots, J \), with probabilities proportional to \( \psi_1(X_n), \ldots, \psi_{j-1}(X_n), \psi_{j+1}(X_n), \ldots, \psi_J(X_n) \).

We want our results to cover this more general setting, because it is often computationally useful. In particular, one does not need to insure that the \( A_1, \ldots, A_J \) are disjoint which can be computationally intensive. Additionally, the “softening” of the exit boundary by introducing random exit times seems to soften any artifacts from the strata edges.

We can fit this more general case, into our original setting by considering the following augmented space

\[
A' = \{(x,k) : x \in A_k, 1 \leq k \leq J\}
\]

with strata given by

\[
A_j' = \{(x,j) : x \in A_j\}.
\]

The \( A_j' \) are then a partition of \( A' \) even when the \( A_j \) are only a cover of \( A \) (namely, not disjoint).

We then extend the Markov dynamics of \( P \) to \( A' \) in the stratified setting by defining a collection of kernels \( \{P'_{\kappa} : \kappa \in [0, 1]\} \) on \( A' \) as follows. If the initial state is \((X_0, j_0)\) then the new state \((X_1, j_1)\) is constructed as follows: \( X_1 \sim P(X_0, \cdot) \) and \( j_1 = j_0 \) if \( \psi_{j_0}(X_1) \geq \kappa \). If \( \psi_{j_0}(X_1) < \kappa \) then \( j_1 \) is chosen randomly according the probabilities \( P(j_1 = k) \propto \psi_k(X_1) \) for \( k \neq j_0 \) as described above. Then, when a particle is started in \( A_j' \) via the injection measure, a value of \( \kappa \) is drawn according to some probability measure \( \eta \) on \([0, 1]\), and the particle moves by \( P_{\kappa}' \) until it leaves \( A_j' \).

In light of this construction, we can recast this more setting into the initial framework of disjoint intervals.

Since all of the kernels \( P_{\kappa}' \) have the same action on \( x \), i.e. the \( x \)-marginal of \( P_{\kappa}'((x,j), \cdot) \) is independent of \( \kappa \), and the \( x \)-marginal of the invariant distribution of each kernel is \( \pi \). In particular, the kernel \( P'(x, \cdot) = \int_{[0,1]} P_{\kappa}'(x, \cdot) \eta(dx) \) acts the same on \( x \) as \( P \).

For the remainder of this article, we will use this last setting. We will drop the superscript of \( A', P_{\kappa}' \), etc. and simply assume that the strata \( A_j \) are a partition of \( A \), and that we have a family of Markov kernels \( P_\kappa \) on \( A \). We will also require that \( P(x, \cdot) = \int_{[0,1]} P_\kappa(x, \cdot) \eta(dx) \), so integrating the kernels over \( \kappa \) gives us back the kernel of the original process we wanted to sample. We will show that, in this setting, the injection measure method gives a way of approximating the invariant distribution of \( P \). In fact, our results apply to any family of kernels \( P_\kappa \) to an collection of kernels \( P_\kappa \) which we will make more explicit later. However, our primary interest is the specific choice of kernels outlined above.

**Remark 2.1.** Note that, if \( \kappa \) is always chosen to be 1, so that there is only one kernel \( P_1 \), then we are back in the original setting, so our results here cover all the cases in which we are interested.

Since the \( A_j \) are now assumed to be disjoint, we can define the index of a point in \( A \) as follows:

\[
\text{Idx}(x) = j \text{ if } x \in A_j.
\]
Now, we are interested in the point at which the sample process exits a stratum, so, for any \( X_0 = x \in A \), define the exit time starting from to be
\[
\tau = \inf \left\{ n \geq 1 : \text{Idx}(X_n) \neq \text{Idx}(X_0) \right\}
\]
where \( \kappa \) is first chosen according to \( \eta \) and then \( \{X_n\}_{n \geq 0} \) moves according to \( P_\kappa \). Note that, since \( \text{Idx}(X_0) \) is possibly random, we define the exit time starting from \( n = 1 \), so that an exit does not occur at time 0 even if \( \text{Idx}(X_0) \neq \text{Idx}(x) \).

We will write \( P_x \) and \( E_x \) to be the respectively the probability and expected value when the process starts from the initial condition \( x \). Similarly we will write \( P_\nu \) and \( E_\nu \) to be probability and expected value when the initial condition is distributed as a the probability measure \( \nu \).

We can now define the main object of our algorithm, the exit kernel \( Q \) defined on \( A \) by
\[
Q(x, \cdot) = P_x(X_\tau \in \cdot).
\]
Note that \( Q \) does not depend on \( \kappa \), because the step where \( \kappa \) is chosen is included in its definition.

Given any injection measure \( \nu \) with \( \nu(A_j) > 0 \) for all \( j \), we define the associated weights \( a_j(\nu) \) and stratified injection measure \( \nu_j \) by
\[
a_j = \nu(A_j) \quad \text{and} \quad \nu_j(\cdot) = \frac{\nu(\cdot \cap A_j)}{a_j}.
\]
Note that \( \nu = \sum_j a_j \nu_j \). Next we define transition matrix \( G \) by
\[
G_{ij} = P_{\nu_j}(X_\tau \in A_i) = \nu_j Q(A_i).
\]
Finally, for each \( j \), define the exit measure from \( A_j \) by
\[
\xi_j(\cdot) = P_{\nu_j}(X_\tau \in \cdot) = \nu_j Q(\cdot).
\]
Note that \( G_{ij} = \xi_j(A_i) \).

We will denote the above quantities associated with the associated with the equilibrium injection measure \( \nu^* \) by \( a^*_j, \nu^*_j, G^* \) and \( \xi^*_j \). Notice that without loss of generality, we can assume that \( a^*_j > 0 \) for all \( j \). Also, observe that \( G^* \) is the matrix which the index process in equilibrium follows. In equilibrium, this is truly a Markov process on the index space since \( \nu^* = \nu^* Q \).

In this framework, the basic version of our algorithm proceeds by\[ \nu^{n+1} = \nu^n Q, \]
and we will want to show that \( \nu^n \to \nu^* \) as \( n \to \infty \).

Analyzing the stratified processes will require understanding their transition kernels. To this end, define, for each \( j \in \{1, \ldots, J\} \) and \( x \in A_j \), the restricted kernel
\[
\tilde{P}_{\kappa,j}(x, \cdot) = \frac{P_{\kappa}(x, \cdot \cap A_j)}{P_{\kappa}(x, A_j)}
\]
We will assume throughout that \( P_{\kappa}(x, A_j) > 0 \) if \( x \in A_j \). Define the corresponding un-normalized kernel
\[
\tilde{P}_{\kappa,j}(x, \cdot) = P_{\kappa}(x, \cdot \cap A_j)
\]
and finally the un-normalized kernel restricted to leaving \( A_j \):
\[
\tilde{P}_{\kappa,j}(x, \cdot) = P_{\kappa}(x, \cdot \cap A^c_j).
\]

\[\text{In practice, the algorithm will build a finite approximation of } \nu^{n+1} \text{ given } \nu^n, \text{ by sampling a finite number of starting points from } \nu^n \text{ and calculating an exit point from each.}\]
Denote the quasi-stationary distribution (QSD) of $P_\kappa$ on $A_j$ by $\tilde{\nu}_{\kappa,j}$. That is, $\tilde{\nu}_{\kappa,j}$ is the unique invariant distribution of $\tilde{P}_{\kappa,j}$. The kernels $\hat{P}_j$, $\tilde{P}_j$, and $\bar{P}_j$ are defined by integrating the associated kernel over $\kappa \sim \eta$, just as $P$ was defined from $P_\kappa$.

The last notation we need to introduce has to do with the relationship between injection measures and the invariant measure on the whole space (and approximations of it), which is what we originally wanted to sample from.

Given an injection measure $\nu_j$ on a stratum $A_j$, define the corresponding occupation measure, and its normalization, by

$$
\mu_j(B) = \mathbb{E}_{\nu_j} \left[ \sum_{k=0}^{\tau_{\nu_j} - 1} 1_{(X_k \in B)} \right] \quad \text{and} \quad \pi_j(B) = \frac{1}{\mathbb{E}_{\nu_j} \mathbb{E}_{\nu_j} \left[ \sum_{k=0}^{\tau_{\nu_j} - 1} 1_{(X_k \in B)} \right]} \mu_j(B),
$$

for $B \subseteq A$. Here $\mathbb{E}_{\nu_j}$ means the expected value of the exit time from $A_j$ for a particle started at $\nu_j$. If we also have weights $a_j$ for the strata, then we can define the total occupation measure on $A$:

$$
\mu = \sum_j a_j \mu_j \quad \text{and} \quad \pi = \frac{1}{\mu(A)} \sum_j a_j \mu_j = \frac{1}{\mu(A)} \mu.
$$

(As before, $\pi_j, \pi^*$, etc. are defined analogously).

Throughout this article, $\| \cdot \|$ will mean $\| \cdot \|_{TV}$ unless otherwise stated. $a, a^*$ will mean the vectors $(a_1, ..., a_J)$, $(a_1^*, ..., a_J^*)$, and similarly for other vectors and matrices. We will also denote the weights given by a total injection measure as $a(\nu) = (a_1(\nu), ..., a_J(\nu)) = (\nu(A_1), ..., \nu(A_J))$.

### 2.3 Statement of the Algorithm

We are now ready to state how the injection measure method proceeds formally. The idea behind the algorithm is to start with some injection measures and weights, and on each iteration, calculate the exit measures given by them, then combine those into new exit measures and weights. The eigenvector version adds a step in which the starting weights are replaced by the eigenvector weights of the transition matrix given by the injection measures.

For reference, we state the precise form of the injection measure method in Algorithm 1.

### 3 Main Results

#### 3.1 Assumptions

Roughly, in order to prove our convergence theorems, we need to first assume that the process is well-behaved both within each stratum, and in terms of how it moves mass between the strata. Our strategy will then be to show that the exit process moves the weights closer to the true weights, at which point the behavior within each stratum lets us show that coupling can occur.

The first two assumptions below can be thought of as non-degeneracy and regularity conditions on $P$ and the strata. The first says, in effect, that $P$ is well behaved and has a unique equilibrium, and that it can move mass into each strata from at least one of the others. The second is a Lyapunov-type condition we can use to control the growth of the exit measure for a well-behaved injection measure, even when some strata are unbounded.

**Assumption (A0).** The Markov transition kernel $P$ has a unique invariant distribution $\pi$. For all $\kappa$, $P_\kappa$ is irreducible and has a unique invariant distribution $\pi_\kappa$, such that for all $j$,

$$
\int_{A_j^c} P_\kappa(x, A_j) \pi(dx) > 0.
$$
Algorithm 1 The Injection Measure Method

1: \( N \leftarrow \# \text{ iterations} \)
2: \( M \leftarrow \# \text{ Points per Strata} \)
3: Initial weights \( a_j^0 \) and strata measures \( \nu_j^0 \) \( \triangleright \) Initial measure \( \nu^0 = \sum_j a_j^0 \nu_j^0 \)
4: for \( n \leftarrow 0 \) to \( N - 1 \) do
5: \hspace{1em} for \( j \leftarrow 1 \) to \( J \) do
6: \hspace{2em} for \( i \leftarrow 1 \) to \( M \) do
7: \hspace{3em} \( X_{j,i}^n(0) \leftarrow \text{Random as } \nu_j^n \)
8: \hspace{3em} \( \kappa_{j,i}^n \leftarrow \text{Random as } \eta \)
9: \hspace{2em} while \( k < \tau \) do
10: \hspace{3em} \hspace{1em} \( X_{j,i}^n(k+1) \leftarrow \text{Random as } P_{\kappa_{j,i}^n}(X_{j,i}^n(k), \cdot) \)
11: \hspace{3em} \hspace{1em} \( \mu_{j,i}^n \leftarrow \sum_{k=0}^{\tau-1} \delta X_{j,i}^n(k) \)
12: \hspace{3em} \hspace{1em} \( \mu_j^n \leftarrow \frac{1}{M} \sum_{i=1}^M \mu_{j,i}^n \)
13: \hspace{2em} \hspace{1em} for \( j,k \leftarrow 1 \) to \( J \) do
14: \hspace{3em} \hspace{2em} \( G_{kj}^n \leftarrow \{|i : \text{Idx}(X_{k,i}^n(\tau)) = j|\} \)
15: \hspace{2em} \hspace{2em} Normalize \( G \) to be a probability transition matrix
16: \hspace{2em} if Basic Version then
17: \hspace{3em} \hspace{1em} for \( j \leftarrow 1 \) to \( J \) do
18: \hspace{4em} \( a_j^{n+1} \leftarrow \sum_k a_k^n G_{kj}^n \) \( \triangleright \) \( a^{n+1} \leftarrow a^n G^n \)
19: \hspace{4em} \( \nu_j^{n+1} \leftarrow \frac{1}{a_j^{n+1}} \sum_k a_k^n \sum_{i : X_{k,i}^n(\tau) \in A_j} \delta X_{k,i}^n(\tau) \)
20: \hspace{4em} \( \nu^{n+1} \leftarrow \sum_k a_k^{n+1} \nu_k^{n+1} \) and \( \mu^{n+1} \leftarrow \sum_k a_k^{n+1} \mu_k^{n+1} \)
21: \hspace{2em} if Eigenvector Version then
22: \hspace{3em} \hspace{1em} \( z^n \leftarrow \text{Normalize solution of } z^n G^n = z^n \) \( \triangleright \) \( z^n = (z_1^n, \ldots, z_J^n) \)
23: \hspace{3em} \hspace{1em} \( \nu_j^{n+1} \leftarrow \frac{1}{z_j^n} \sum_k z_k^n \sum_{i : X_{k,i}^n(\tau) \in A_j} \delta X_{k,i}^n(\tau) \)
24: \hspace{3em} \hspace{1em} \( \nu^{n+1} \leftarrow \sum_j z_j^n \nu_j^{n+1} \) and \( \mu^{n+1} \leftarrow \sum_k z_k^{n+1} \mu_k^{n+1} \)
25: Return: \( \mu^N \) \( \triangleright \) Approximate of \( \pi \)
and $P(x, \cdot) = \int_{[0,1]} P_\kappa(x, \cdot) \eta(dx)$.

**Assumption (A1).** There exists a continuous function $V : A \to [0, \infty)$, a compact set $K \subset A$, and $b, K > 0$, $\gamma \in (0, 1)$ such that for all $x \in A$, $\kappa \in [0, 1]$,

$$P_\kappa V(x) \leq \gamma V(x) + b 1_K$$

and if $x \in K$ and $\text{Idx}(x) = j$, then

$$E_x \tau_j \leq K.$$

Next is our primary assumption on the system’s “microscopic” behavior, i.e. it’s behavior inside a stratum. Intuitively, it states that any particle starting from any point in a stratum has a chance of coupling with the quasi-stationary distribution (QSD) of $P_\kappa$ in that stratum before or at the time of exit. Therefore, it’s exit distribution will look like, with some probability, what it would have been had the particle been injected according to the QSD.

**Assumption (A2).** For all $j$ and $\kappa \in [0, 1]$, there exists a unique QSD $\tilde{\nu}_{\kappa, j}$ of $P_\kappa$ in $A_j$. Let $\tilde{\nu}_j = \int_{[0,1]} \tilde{\nu}_{\kappa, j} \eta(dx)$, where $\eta$ is the distribution $\kappa$ is drawn from at the start of a trajectory. Then there exists a constant $0 < c < 1$ such that, for any $j$ and $\nu_j^0$,

$$\xi_j^0 \geq c \cdot \tilde{\xi}_j$$

where $\tilde{\xi}_j$ is the exit measure from $A_j$ started from $\tilde{\nu}_j$, via the kernel $Q$. That is, $\tilde{\xi}_j = \tilde{\nu}_j Q$.

Next, we need an assumption on the “macroscopic” behavior, i.e. how mass moves between strata. (Assumption A0 ensures some movement between strata, but not does not ensure that it is global or give quantitative information.) We have three versions of this assumption, and a convergence theorem that holds under each. The first form says that after enough exits, a particle has a probability, bounded from below, of being in any of the strata. i.e. the exit process explores the space after enough time. The second choice of assumption says that the transition matrix in equilibrium, $G^*$, is geometrically ergodic, with mixing rate $\lambda$. This will allow us the prove a convergence result for the basic version with a more precise estimate of the convergence rate, combining the microscopic rate $c$ and macroscopic rate $\lambda$.

Note that the above assumptions could fail due to the index of a particle on the $n$-th exit being periodic. For example, if there are only two strata, then the index will keep flipping between one stratum and the other between exits. If this is the only problem keeping the assumptions from holding, it can be remedied by introducing laziness to the exit kernel. That is, $Q$ can be replaced by $pQ + (1 - p)I$, where $I$ is the identity kernel, and the periodicity will be removed.

The last form of the macroscopic assumption pertains to the eigenvector version, and is borrowed from the perturbation results in [14]. Instead of giving the mixing rate of $G^*$, we assume that the invariant distribution vector $G^*$ has bounded sensitivity to perturbations in its entries. We will then prove a convergence theorem for the eigenvector version, this time combining $c$ and the sensitivity constants. We state all three macroscopic assumptions below.

**Assumption (B1).** There exists $u > 0$ and $m \geq 1$ such that, for any $j$, if $X_0 \sim \tilde{\nu}_j$ and $X_n$ is the position of the $n$-th exit from one stratum to another, then

$$\text{Law}(\text{Idx}(X_n)) \geq u a^*$$

for all $n \geq m$.

**Assumption (B2).** There exists $m \geq 1$ and $\lambda^* \in (0, 1)$ such that, for $n \geq m$ and any two probability vectors $a, b$ on $\{1, \ldots, J\}$,

$$\|a(G^*)^{n+1} - b(G^*)^{n+1}\| \leq \lambda \|a(G^*)^n - b(G^*)^n\|.$$

\footnote{If $V(x) \to \infty$ as $|x| \to \infty$, then the condition is equivalent to requiring that $PV(x) \leq \gamma' V(x) + b'$ for some different $\gamma' \in (0, 1)$ and $b' > 0$. We will use the slightly more general condition in Assumption A1.}
Assumption (B3). There exists \( \theta_{ik} > 0 \) for \( i, k \in \{1, ..., J\} \), depending only on \( G^* \) such that, if \( G \) is an irreducible transition matrix on \( \{1, ..., J\} \) such that, if \( G, G^* \geq cG^* \) and \( z, z^* \) are the invariant measures of \( G, G^* \), resp., then for all \( j \)

\[
\sup_j (\log(z_j) - \log(z_j^*)) \leq \sum_{i \neq k} \theta_{ik} |G_{ik} - G_{ik}^*|
\]

3.2 Main Theorems

We are now ready to state our main theorems, as well as give a brief strategy for proving them. The detailed proofs will be reserved for section 5.

The first theorem states in effect that the equilibrium injection measure exists uniquely, and that the corresponding occupation measure is the original measure we wanted to sample. Our strategy for proving this involves manipulating the sub-stochastic kernels \( \hat{P}_{n,j} \) and \( \hat{P}_{n,j} \), and how they relate the injection and occupation measures. It is essentially a Poisson equation argument at heart: \( \mu_j \) solves a Poisson-like equation in terms of \( \nu_j \), and combining these gives an equation for \( \mu \) that reduces to \( \mu(t - P) = 0 \) in the equilibrium case.

**Theorem 3.1.** Suppose that (A0) and (A1) hold. Then there exists a unique probability measure on \( A \), \( \nu^* \) such that \( \nu^* V < \infty \) and \( \nu^* Q = \nu^* \). Furthermore, for the corresponding occupation measure, \( \mu^* = \mu^* P \), where \( P(x, \cdot) = \int_{[0,1]} \hat{P}_n(x, \cdot) \eta(dx) \).

Our second main theorem establishes convergence of the basic version after enough steps are performed. We use a fairly straightforward coupling argument to prove it. The idea is that assumption (B1) states that two particles following the exit process \( Q \) can eventually be in the same stratum, at which point (A1) says that they have a chance of being coupled after the next iteration of \( Q \). This establishes geometric ergodicity of \( Q \).

**Theorem 3.2.** Under (A1) and (B1), for any \( n \geq m \), \( Q^{n+1} \) is a global contraction on probability measures on \( A \), with contraction constant \( 1 - c^2 u \). In particular, \( \| \nu^{k(m+1)} - \nu^* \| \leq (1 - c^2 u)^k \) for \( k \geq 1 \).

Next, we have a theorem giving a more precise rate of geometric convergence in the long term. To prove it, we will need to construct a new metric which balances the total variation distance between \( \nu^\alpha \) and the true \( \nu^* \), and the difference between their strata weights. The idea is that, even if no coupling is possible on a given step, the weights will still get closer to the truth, which will allow coupling with \( \nu^\alpha \) to occur at some future time. Since the metric we show contraction in is equivalent to TV-distance, we get the desired result.

**Theorem 3.3.** Suppose that (A0)-(A2) and (B1)-(B2) hold. For any \( \nu^0 \) such that \( \nu^0 V < \infty \), there exists \( q_n \), \( \forall n \geq 0 \), such that \( \| \nu^{n+1} - \nu^* \| \leq q_n \| \nu^n - \nu^* \| \) and as \( n \to \infty \),

\[
q_n \to q := \inf_{\beta \in (0, 1)} \left( \inf_{0 < \alpha < \frac{\log_q 1}{1 - \beta c}} \max \left( 1 - \beta c + \alpha S(1 - c), \frac{1 + (1 - \beta)\alpha \lambda}{1 + (1 - \beta)\alpha} \right) \right)
\]

where \( S = \frac{1}{1 - \lambda} \).

The expression we get for the limiting rate is complex, but has a meaningful interpretation. Note that it takes the form of a maximum of a rate in terms of \( c \) and one in terms of \( \lambda \). Therefore, it suggests that, of the macro- and microscopic rates, whichever is slower acts as a sort of bottleneck. It can also be shown straightforwardly that if \( 1 - c = 0 \), then the limiting rate is \( \lambda \), and if \( \lambda = 0 \), then the limiting rate is order \( \frac{1}{2} \) in \( 1 - c \). So we can conjecture that, to first order, it behaves like \( \max(\sqrt{1 - c}, \lambda) \).

Our final theorem establishes geometric convergence of the eigenvector version, for starting guess \( \nu^0 \) sufficiently close to \( \nu^* \). The final rate involves the microscopic coupling parameter \( c \) and a constant \( r^\infty \)

\[3\] Where \( \tilde{G}_{ij} = \xi_j(A_i) \) is the transition matrix between strata when the injection measures are the QSD’s.
relating to how sensitive $G^*$’s eigenvector is to perturbations. In this way, it is like an analogue to the previous theorem, with $\lambda$ replaced by $r^\infty$. It also does not have the bottleneck form, suggesting that if $\lambda$ is a slow rate, then the eigenvector version may be faster than the basic version, which is to be expected.

**Theorem 3.4.** Let $\nu^0, \nu^1, \ldots$ be the total injection measures for a run of the eigenvector version. Suppose that (A0)-(A2) and (B3) hold. Let

$$r^\infty = 2(1-c) \sup_j (a^*_j) \sup_{i,k} \left( \frac{e^{\theta_{ik}}}{a^*_i} - 1 \right)$$

$$E = \exp \left( 2(1-c) \sum_i \sup_{k \neq i} (\theta_{ik}) \right)$$

Also suppose that

$$\| \nu^0 - \nu^* \| < \frac{1 + r^\infty E}{2(1-c) \sup_{i \neq k} \frac{\theta_{ik}}{a^*_i}} \left( \frac{1}{q} - 1 \right),$$

where $q = 1 - \frac{c}{r^\infty + 1}$. Then there exists $p_n, \forall n \geq 0$, such that $\| \nu^{n+1} - \nu^\ast \| \leq p_n \| \nu^n - \nu^\ast \|$ and

$$p^n \to 1 - \frac{c}{r^\infty + 1}.$$

**4 Numerical Simulations**

We now turn to analyzing the results of numerical simulations of our algorithms. We focus mainly on the eigenvector version here, as in all our simulations its convergence is strictly faster than the basic version, but does not immediately suggest how to prove this. We include the code used to generate our results in the following git repository: https://gitlab.com/gabeearle/julia-code-stratified-2021

We test our method on the two-dimensional Maier-Stein system, as outlined in [11], with parameters used in [8]. This is a relatively simple low-dimensional system, but one which displays the non-reversibility which our methods are suited for. Furthermore, with the chosen parameters, the system displays both increased non-gradient effects and a double-well type invariant distribution.

For reference, we state the version of the Maier-Stein system we use here. The dynamics evolve according to the following SDE:

$$du = (u - u^3 - \beta uv^2)dt + \sqrt{\epsilon}dW^1_t$$

$$dv = -(1 + u^2)vdv + \sqrt{\epsilon}dW^2_t$$

where $\beta, \epsilon > 0$ and $W^1, W^2$ are independent Brownian motions. For $\beta \neq 1$, this system is non-reversible. For $\beta > 4$, it displays additional non-gradient behavior and unusual minimum action paths, as detailed in [8]. We use the parameters $\beta = 10$ and $\epsilon = 0.01$. Fig. 1 shows an approximate 2D histogram plot of the invariant distribution of this system, computed via a discretization of the dynamics.

In Fig. 2, we show several ways to choose the strata. One is a subdivision of the space into 3 vertical ellipses, another into 5 smaller ones. The next is a version in which the 5 strata are rotated, so that they do not perfectly line up with the axes of the system. Finally, we also study the case where the space is divided into 6 circular strata, with varying sizes and more than two strata overlapping at once.

In Fig. 3, we compare the 1-dimensional projections of both estimates to get a sense of how well the algorithm approximates the double-well structure of the system. Fig. 4 shows the un-stratified and stratified (with 3000 exits) estimates of the $u$-marginal after 30 iterations. As we can see, they agree quite well, demonstrating the accuracy of our method.

In Fig. 4, we plot, for each of the strata setups, the final total injection measure, i.e. our approximation of $\nu^n$, where $n = 30$ is our final iteration number. Note that in each case, in injection measure has
Figure 1: A 2D histogram of the invariant measure for the Maier-Stein system. This approximation of the true density was generated by an (un-stratified) Runge-Kutta Method, with noise, run for $10^8$ steps, and averaged over two runs. The $u$-marginal density of this distribution will be used as our benchmark for the true density when calculating the error of the stratified method in $u$.

Figure 2: Four choices of elliptical strata for the Maier-Stein system. We use setups with 3 and 5 vertically oriented strata, tilted strata, and 6 circular strata which cover the space more tightly.
Figure 3: 1-dimensional histograms for the true $u$-marginal density (again from a high-resolution un-stratified run) and the approximations of it generated by the algorithm with the strata setups above. Note that, with some deviation, each setup produces quite an accurate approximation of the marginal density.

Figure 4: 2D Histogram of the density of the total injection measure, for the strata setups in Fig. 2. The densities are calculated from the points generated on the last iteration of a 30-iteration run of the algorithm, with 3000 exits per iteration. Because we are using strata with hard boundaries, the injection measures lie on a portion of the boundaries of each ellipse.
Figure 5: Log-Log total variation error in $u$ vs time, for each of the strata setups. The error values are averaged over 20 full runs of the algorithm for each setup. Computational time is measured as the total number of steps of time $h$ that each particle has taken, up to the current iteration. Note that, for the two 5-strata setups, a dramatic increase in accuracy occurs somewhere between 1000 and 3000 exits run per iteration.

We can also examine how much the number of exits per iteration affects performance. Fig. 5 shows a log-log plot of the error in the $u$-marginal of the invariant distribution vs. time, for a range of exits the algorithm was run with and for each setup. The error is approximately computed by taking the TV-distance between the histogram from the un-stratified run shown in Fig. 3, and the histogram computed from the stratified run, averaged up to up to iteration $n$. The time spent is computed as the total number of steps of the SDE taken in all strata up to iteration $n$. The results, are shown in Fig. 5.

We can see that, in the 3-strata setting, increasing the number of exits (hence the number of sample from each exit measure) per iteration does not seem to affect the accuracy per computational time greatly. The runs of the algorithm with more exits do in fact decrease the error more quickly on each iteration, but the iterations also take longer, and in the end the effects balance out, and all three runs have similar trend lines of error vs. time. For both 5-strata setups, this behavior changes dramatically. The version of the algorithm run with 3000 exits still takes longer to perform, but also sees error decrease much more rapidly. The reason for such an increase in accuracy is not immediately clear, and could be a subject of future work. The interesting phenomenon is that there appears to be some threshold, in terms of the work put in on each iteration, at which the compounding error becomes much less of an issue.

A possible hint as to why the accuracy threshold occurs can be found by examining how much the injection measure changes for each strata setup and each choice of the number of exits. We measure fluctuations in the total injection measure similarly to the error in the invariant measure: by taking the converged to a density roughly along the boundaries of the ellipses, illustrating that the algorithm can be used to estimate this density as well as the invariant distribution of the system.
Figure 6: Fluctuation of the injection measure vs. time for the strata setups. The fluctuation is computed as the TV distance between the $\nu$-marginal of the total injection measure on one iteration and on the next. Computational time is computed as in Fig. 6. Corresponding to the jump in accuracy in the 5-strata setups is a lower and more regular fluctuation in the injection measure.

TV-distance between a histogram of the approximate $\nu$ marginal in the $\nu$-direction on one iteration, and the same histogram on the next iteration. The results are shown, for all the setups, in Fig. 6. First, we can see that the total variation change in the injection measure from one iteration to the next remains quite small for each case. However, corresponding to the large drops in error for the 5-strata setups are cases where, when going from 1000 to 3000 exits, the fluctuation not only decreases but seems to behave periodically. What this means for the algorithm’s deeper behavior is not immediately clear, but it could be related to the accuracy threshold observed in Fig. 5.

In Fig. 7, we plot the number of points in the injection measure of each stratum, for each setup and some choices of the number of exits per iteration. Unsurprisingly, we see in the 5-strata case that having more exits results in the more points in each injection measure. More specifically, we can see a large difference in how many points are in the “smallest” of the strata. Perhaps the accuracy threshold has some connection to there being enough exits that the least common strata still gets enough points to approximate $\nu^j$ reasonably well. In that case, an alternate version of the algorithm could be devised, in which more exits are sampled from the strata neighboring the “smallest” ones, as that is where it is most important to get samples. This would have the advantage of avoiding unnecessary computational work in the strata which naturally get many points injected already. We can already see from Fig. 7 that one of the main advantages of stratified sampling holds here: that the different strata have different sizes, in terms of how likely particles are to enter them, but we put in similar amounts of work in them, by always drawing the same number of points from each injection measure.

We can derive some insight into how the error behaves by decomposing into the error in the weights or in each of the strata. In doing so, we can see how the error in the weights relates to the overall error,
Figure 7: Injection measures sizes vs. iteration number for the strata setups. The injection measure size is the total number of points that exited into a given stratum on the current iteration. The figure illustrated that the size of each measure converges to some value and then fluctuates around it. Note that some strata have significantly more points injected than others, but the same number of injections are drawn and run to exit on an iteration for each, illustrating that similar amounts of work are done in each strata despite their different sizes.
and in each stratum. In Fig. 8, we first plot the “occupation weights” of the strata for the 5 vertical strata setup, that is, their weight as a fraction of the total mass in the invariant measure. We then plot the error in the occupation measure for each stratum. For this, we use as a benchmark the occupation measure computed from an un-stratified run, where the measure in each stratum is formed from all points visited in that stratum over the run. The error is then taken to be the total variation distance between a histogram of the stratified run’s occupation measure and this benchmark. Finally, in Fig. 8 we also show in error in the injection weights over time and the overall total variation error over time.

Comparing the results for stratified runs with different numbers of exits per iteration, we see that the overall error behaves almost identically to the error in strata 2 and 3. This is not surprising, as those are the strata with by far the highest occupation weight. Since they dominate the mass of the invariant measure, most of the total error will be composed of the error in them. We can see that the error in the dominant strata sees a dramatic drop somewhere between 1000 and 3000 exits, suggesting that the same behavior we saw before in the total error is caused by the drop occurring in strata 2 and 3.

Another observation to note is that the error in the weights also behaves very similarly to the error in the dominant strata. A sharp increase in accuracy in those occupation measures when going from 1000 to 3000 exits corresponds to a similar rise in accuracy in the injection weights. Recall that the injection weights at iteration \( n \) are given by the eigenvector of the transition matrix \( G^n \), which itself is a function of the injection measures and the exit kernel. The occupation and injection measures are also directly related to each other. The behavior in the weights being similar to that of strata 2 and 3 suggests that, for this choice of strata, the weights are especially sensitive to perturbations in those strata. Therefore, it seems that the weights are especially sensitive to perturbations in the corresponding entries in the \( G \) matrix.

Hence, in this case, the strata which are dominate the invariant measure’s mass and those to which the eigenvector weights are most sensitive are the same.

On the other hand, for a different choice of strata, we see slightly different behavior. In Fig. 9, we plot the occupation weights, error in occupation measures and weights, and total error in the same way as before, but for the 6 circular strata setup. As before, we can see that the total error behaves similarly to that of the highest weight strata, this time strata 5 and 6. However, the injection weight error behaves more like the occupation error in the previous 4 strata, suggesting that the transition matrix \( G \) for this setup is more sensitive to the entries corresponding to transitions from those strata. So as before, some strata are more important to the overall error and some are more important than the weights, but this time the two do not line up. Strata with small occupation weight have a large influence on the accuracy of the eigenvector weights. This is important to note, as it illustrates one of the main motivations for our method: that strata which are visited rarely may still have outsize influence on something of interest.

5 Proofs

5.1 Existence and Uniqueness of the Equilibrium Injection Measure \( \nu^* \)

The first result we must prove is that the fixed point of \( Q \) exists, and has the appropriate distribution. We will show this by manipulating the sub-stochastic kernels derived from \( P_\kappa \), and using the ways they relate the injection, exit and occupation measures. First, however, we must establish some regularity. To this end, our first lemma states that, if a given injection measure has controlled growth with respect to the Lyapunov function in assumption (A1), then the corresponding exit and occupation measures exist, and also have controlled growth.

**Lemma 5.1.** Suppose that \( (A1) \) holds. Let \( \nu_j \) be a probability measure on \( \Lambda_j \), with exit time, exit measure and occupation measure \( \tau_j, \xi_j \) and \( \mu_j \), such that \( \nu_j V < \infty \). Then \( \nu_j \tau_j < \infty \), \( \xi_j \) and \( \mu_j \) are well defined, and \( \xi_j V < \infty \), \( \mu_j V < \infty \). Furthermore, if \( \pi \) is the invariant distribution of \( P_\kappa \) for any \( \kappa \in [0,1] \), then \( \pi V < \infty \) and \( \pi |_{\Lambda_j} (I - P_j) < \infty \).
Figure 8: Occupation weights, individual strata errors, weights error and total error, for 5 vertical strata setup. The error in individual occupation measures is calculated as the TV distance between the $u$-histogram generated by an un-stratified run and the $u$-histogram for the occupation measure from the stratified algorithm, with the same bins as for the overall error. The total error behaves similarly to the error in the largest-weight strata, as does the weight error.
Figure 9: Occupation weights, individual strata errors, weights error and total error, for the 6 circular strata setup. Errors are calculated the same way as in Fig. 10. In this case, the total error still behaves similarly to the error in the largest strata, but the weight error is more similar to that of the smaller strata, showing that low-weight strata can have significant importance.
Proof. Fix \( \kappa \in [0, 1] \). Let \( B \subset \mathcal{K} \) where \( \mathcal{K} \) is defined in A1, and let \( \tau_B = \inf\{ n : X_n \notin B \} \) Where \( X_0 = x \) and \( X_n \) follows \( P_x \). Define

\[
V^*_B = \inf\{ V(x) : x \in B \}
\]

Following ideas from \([13, 12]\), we define the following process:

\[
M_n = \frac{V(X_n)}{V^*_B \gamma^n} - b \sum_{k=0}^{n-1} \frac{1(\{x \in K\})}{\gamma^{k+1}}
\]

with the filtration \( \mathcal{F}_n = \sigma(X_n : k \leq n) \). Then, using the condition of (A1) to bound \( E[V(X_{n+1}) | \mathcal{F}_n] = PV(X_n) \), we can simplify

\[
E[M_{n+1} | \mathcal{F}_n] = \frac{E[V(X_{n+1}) | \mathcal{F}_n]}{V^*_B \gamma^{n+1}} - b E\left[ \sum_{k=0}^{n-1} \frac{1(\{x \in K\})}{\gamma^{k+1}} | \mathcal{F}_n \right] - b E\left[ \frac{1(\{x \in K\})}{\gamma^{n+1}} | \mathcal{F}_n \right]
\]

to show that

\[
E[M_{n+1} | \mathcal{F}_n] \leq \frac{V(X_n)}{V^*_B \gamma^n} - b \sum_{k=0}^{n-1} \frac{1(\{x \in K\})}{\gamma^{k+1}} = M_n.
\]

Furthermore, if \( E | M_n | < \infty \), then \( EV(X_n) < \infty \) because the term \( b \sum_{k=0}^{n-1} \frac{1(\{x \in K\})}{\gamma^{k+1}} \) is bounded by \( bn \), and so

\[
E[M_{n+1}] = E\left[ \frac{V(X_{n+1})}{V^*_B \gamma^{n+1}} - b \sum_{k=0}^{n-1} \frac{1(\{x \in K\})}{\gamma^{k+1}} \right] \\
\leq E\left[ \frac{V(X_{n+1})}{V^*_B \gamma^{n+1}} \right] - b E\left[ \frac{1(\{x \in K\})}{\gamma^{n+1}} \right] \\
\leq \frac{\gamma V(X_n) + b}{V^*_B \gamma^{n+1}} \leq \frac{\gamma E[V(X_n)] + b}{V^*_B \gamma^{n+1}} < \infty.
\]

Therefore, if \( EV(X_0) < \infty \), then \( E | M_n | < \infty \) for all \( n \geq 0 \), and so \( M_n \) is a super-martingale with the filtration \( \mathcal{F}_n \).

Now suppose that \( x \) is a fixed point in \( B \). Then \( 1 \leq \frac{V(x)}{V^*_B} \) and \( 1 \{x \in K\} = 0 \). Therefore, using the bounded stopping time \( n \wedge \tau_B \) for fixed \( n \), the optional stopping theorem implies that

\[
\frac{1}{\gamma^n} P(\tau_B > n) \leq E\left[ \frac{V(X_n)}{V^*_B \gamma^n} 1(\{\tau_B > n\}) \right] \leq EM_n \wedge \tau_B \leq EM_0 = \frac{V(x)}{V^*_B}
\]

Therefore, \( P(\tau_B > n) \leq \gamma^n \frac{V(x)}{V^*_B} \), and so summing over \( n \), we have

\[
E \tau_B \leq \frac{1}{1 - \gamma} \frac{V(x)}{V^*_B}
\]

It follows that if \( x \sim \nu \) and \( \nu V < \infty \), then \( E \tau_B < \infty \).

Now suppose that \( \nu V < \infty \), and define the measure

\[
\mu_B(\cdot) = \int_{\mathcal{K}} \nu \cdot \sum_{k=0}^{\tau_B - 1} 1(\{x_k \in \cdot\}).
\]

We now have that

\[
\mu_B(B) = E_\nu \tau_B \leq \frac{1}{1 - \gamma} \frac{\nu V}{V^*_B}
\]
so the total mass of $\mu_B$ is bounded. Next, let $\xi_B(\cdot) = \mathbb{E}1_{(X_n \in \cdot)}$ be the measure of the point where $X_n$ hits $B$. Using $M_0 \geq EM_0 \geq \mathbb{E}\left[\frac{V(X_n \cap \tau_B)}{V_B^\ast}B\right]1_{\tau_B > n}$ and letting $n \to \infty$, we have

$$M_0 \geq \frac{\mathbb{E}V(X_{\tau_B})}{V_B^\ast}$$

and so

$$\mathbb{E}V(X_{\tau_B}) \leq V(X_0).$$

Therefore, $\xi_B V \leq \nu V \leq \infty$. Setting $B = A_j \cap \mathcal{K}^c$, and with compactness of $A_j \cap \mathcal{K}$, the result follows. \qed

Our next lemma, which is the key to proving the first main theorem, relates an injection measure to its occupation and exit measure through the sub-stochastic kernels. Recall that, when a trajectory to exit is started, the threshold that determines when an exit occurs, $\kappa$, is chosen from a distribution $\eta$ on $[0, 1]$.

**Lemma 5.2.** Suppose that (A0) and (A1) hold. Let $\nu_j$ be a distribution on $A_j$ such that $\nu_j V < \infty$. Fix $\kappa \in [0, 1]$, and let $\tau_{\kappa,j}, \mu_{\kappa,j}, \xi_{\kappa,j}$ be the exit time, occupation, and exit measures for a trajectory in $A_j$ started at $\nu_j$, if the threshold is chosen to be $\kappa$. Then

$$\mu_{\kappa,j} = \sum_{k=0}^{\infty} \nu_j \hat{P}_{\kappa,j}^k, \quad \mu_{\kappa,j}(I - \hat{P}_{\kappa,j}) = \nu_j, \quad \text{and} \quad \xi_{\kappa,j} = \mu_{\kappa,j} \hat{P}_{\kappa,j}$$

where $I$ is the identity kernel.

Note that the true occupation and injection measures are obtained from those above by integrating over $\kappa$: $\mu_j(\cdot) = \int_{[0, 1]} \mu_{\kappa,j}(\cdot) \eta(d\kappa)$, $\xi_j(\cdot) = \int_{[0, 1]} \xi_{\kappa,j}(\cdot) \eta(d\kappa)$.

**Proof.** For the first equality, the definition of the occupation measure gives

$$\mu_{\kappa,j}(\cdot) = E_{\nu_j} \left[ \sum_{k=0}^{\tau_{\kappa,j}-1} 1(X_k \in \cdot) \right]$$

$$E_{\nu_j} \left[ \sum_{k=0}^{\infty} 1(X_k \in \cdot) \cdot 1(\tau_{\kappa,\nu_j} > k) \right] = \sum_{k=0}^{\infty} \mathbb{P}_{\nu_j} (X_k \in \cdot, \tau_{\kappa,\nu_j} > k) = \sum_{k=0}^{\infty} \nu_j \hat{P}_{\kappa,j}^k(\cdot)$$

Because $\nu_j \hat{P}_{\kappa,j}^k(\cdot)$ is the probability of a particle being in $\cdot$ at time $k$ and not having left $A_j$ at any time before $k$. The series on the third line above converges because $E\tau_0 < K$ by assumption (A0). The second equality is given by applying $I - \hat{P}_{\kappa,j}$ to the first.

For the final equality,

$$\xi_{\kappa,j}(\cdot) = P_{\nu_j} (X_{\tau_{\kappa,j}} \in \cdot) = \sum_{k=1}^{\infty} \mathbb{P}_{\nu_j} (X_k \in \cdot, \tau_{\kappa,j} = k) = \sum_{k=1}^{\infty} \nu_j \hat{P}_{\kappa,j}^{k-1} \hat{P}_{\kappa,j}(\cdot)$$

$$= \left( \sum_{k=0}^{\infty} \nu_j \hat{P}_{\kappa,j}^k \right) \hat{P}_{\kappa,j} = \mu_{\kappa,j} \hat{P}_{\kappa,j}. \quad \square$$

We can now put the equalities in the last lemma together to derive, after integrating over $\kappa$, that a fixed point of $Q$ must have a corresponding occupation measure which is a fixed point of $P$, as defined in [3.1]. We then can establish existence and uniqueness of the fixed point of $Q, \nu^*$, since $P$ is assumed to have a unique invariant measure.
Proof of Theorem 3.1. Let \( \nu = \sum_j a_j \nu_j \) be a total injection measure, with corresponding occupation measure \( \mu = \sum_j a_j \mu_j \). By integrating the equations of Lemma 5.2 we get, for each \( j \),

\[
\nu_j = \int_{[0,1]} \mu_{\kappa,j}(I - \hat{P}_{\kappa,j}) \eta(\text{d}k) \quad \text{and} \quad \xi_j = \int_{[0,1]} \mu_{\kappa,j} \hat{P}_{\kappa,j} \eta(\text{d}k).
\]

Furthermore, for any \( \kappa \) and \( j \),

\[
\hat{P}_{\kappa,j} + \hat{P}_{\kappa,j} = P_\kappa \quad \text{and} \quad \sum_j a_j \xi_j = \nu Q, \quad \text{by construction. Therefore, we can subtract the second equation above from the first to get}
\]

\[
\nu - \nu Q = \int_{[0,1]} \sum_j a_j \mu_{\kappa,j}(I - P_\kappa) \eta(\text{d}k) = \mu(I - P)
\]

Therefore, \( \nu = \nu Q \), if and only if \( \mu(I - P) = 0 \), in which case \( \mu \sim \pi \), where \( \pi \) is the unique invariant distribution of \( P \).

Now, for a given occupation measure \( \mu_j \) on \( A_j \), define an injection measure \( \nu_j \) on \( A_j \) by

\[
\nu_j \sim \mu_j(I - \hat{P}_j)
\]

recalling that \( \hat{P}_j = \int_{[0,1]} \hat{P}_{\kappa,j} \eta(\text{d}k) \). Then the occupation measure given by \( \nu_j \) on \( A_j \) is, by Lemma 5.2

\[
\int_{[0,1]} \nu_j(I + \hat{P}_{\kappa,j} + \hat{P}^2_{\kappa,j} + \cdots) \eta(\text{d}k) = \int_{[0,1]} \mu_j(I + \hat{P}_j)(I + \hat{P}_{\kappa,j} + \hat{P}^2_{\kappa,j} + \cdots) \eta(\text{d}k) = \mu_j(I - \hat{P}_j)(I + \hat{P}_j + \hat{P}^2 + \cdots) = \mu_j
\]

Note that the infinite sum of operators converges, because for any \( B \subset A_j \) and any probability measure \( \eta \),

\[
\sum_{k=0}^{\infty} \eta \hat{P}^k_j(B) \leq \sum_{k=0}^{\infty} \eta \hat{P}^k_j(A_j) = \sum_{k=0}^{\infty} \mathbb{P}_j(\tau_j \geq k) \leq K
\]

Where \( K \) is as in (A1). Therefore, given an occupation measure \( \mu_j \), the associated injection measure is \( \mu_j(I - \hat{P}_j) \), up to normalization.

We can now show existence of the fixed point. Let \( \pi |_{A_j} \) be the un-normalized restriction of \( \pi \) to \( A_j \). Define the injection measure on \( A_j \) by

\[
\nu^*_j \sim (\pi |_{A_j})(I - \hat{P}_j)
\]

and the weights by

\[
a^*_j = \frac{1}{W} \pi(A_j) / \mathbb{E}_{\nu^*_j}
\]

Where \( W = \sum_j \pi(A_j) / \mathbb{E}_{\nu^*_j} \). Then the above calculation show that if \( \nu^* = \sum_j a^*_j \nu^*_j \), then \( \nu^* - \nu^* Q \sim \pi(I - P) = 0 \).

Uniqueness follows similarly. If \( \nu Q = \nu \), then we must have \( \mu P = \mu \), so \( \mu \sim \pi \). But then \( \nu_j = \mu_j(I - \hat{P}_j) \sim (\pi |_{A_j})(I - P_j) \), and since \( \nu_j \) must be a probability measure by construction, this determines all the \( \nu_j \). Similarly, the weights must be given by the formula for \( a^*_j \) above, in order to have \( \mu \sim \pi \), so they are determined as well. Therefore, the fixed point of \( Q \) is unique. Furthermore, it satisfies \( \nu^* V < \infty \), because \( (\pi |_{A_j})(I - \hat{P}_j)V < \infty \) for all \( j \).

This concludes our proof that the algorithm has the appropriate fixed distribution. Notice that, in a sense, the argument we used is really a Poisson equation argument. We use that \( \mu_j \) satisfies the Poisson-like equation \( \mu_j(I - \hat{P}_j) = \nu_j \) for each \( j \) to show that \( \mu(I - P) = \nu - \nu Q \), which is like a Poisson equation on the whole space \( A \). This equation is what lets us relate the fixed point of \( P \) to that of \( Q \).

In the rest of the proofs, we approach the problem of showing that the injection measures given by the algorithm actually approach the fixed point of \( Q \).
5.2 Convergence Proof for the Basic Algorithm

We will first prove that the measures $\nu Q^n$, which are the distributions found by the basic algorithm without finite approximations, do in fact converge to $\nu^*$ as $n \to \infty$, for any initial $\nu$. Later we will improve our estimate of how fast the convergence is.

The proof we give below has a simple intuition. Consider a particle following the exit kernel $Q$, whose position at the $n$-th exit from a stratum is $X_n$. Assumption (B1) implies that, after $m$ exits, the particle has at least a certain probability of being in any stratum. This means that, for two such particles, a coupling can be constructed so they have a chance, bounded from below, of being in the same stratum. Assumption (A2) then implies that they have a probability of being at the same location on the next exit, for an appropriate coupling. Therefore, the operator that applies the algorithm $m + 1$ times, i.e. $Q^{m+1}$, is a contraction in total variation. With this intuition, we give the formal proof now.

Proof of Theorem 3.2: First note that, because $Q$ is a time-homogeneous Markov kernel, it suffices to consider is a contraction in the case where the initial measures are single-point delta distributions.

Therefore, let $\nu^1 = x_1, \nu^2 = x_2$ where $x_1, x_2 \in A$ and $x_1 \neq x_2$, and let $n \geq m$.

Let $j_1 = \text{Idx}(x_1)$ and $j_2 = \text{Idx}(x_2)$. Let $X^i_n = x_i$, and $X^i_0$ be the location of the $n$-th exit from a stratum starting from $X^i_0$, for $i = 1, 2$. $X^i_0$ is in $A_{j_i}$, so by (A1), with probability at least $c$, a particle started at $X^i_0$ exits as if it started at $\tilde{\nu} j_i$. That is, there exists measures $\eta^i_{j_i}$ on $A^i_{j_i}$ for $i = 1, 2$ such that

$$\nu^iQ = \delta_{\tilde{\nu} j_i}Q = c \xi_{j_i} + (1 - c)\eta^i_{j_i} = c\tilde{\nu} j_iQ + (1 - c)\eta^i_{j_i}$$

So applying $Q^{n-1}$ to both sides, where $n \geq m$,

$$\nu^iQ^n = c\tilde{\nu} j_iQ^n + (1 - c)\eta^i_{j_i}Q^{n-1}.$$ 

Now, by assumption (B1), the probability that a particle starting from the QSD in $A_{j_i}$ is in $A_k$ after $n$ exits is minorized by $ua^*_k$, for any $j$, which means that

$$\tilde{\nu} j_iQ^n(A_k) \geq u a^*_k \forall i, k.$$ 

Therefore, for $i = 1, 2$,

$$\nu^iQ^n(A_k) \geq u \cdot a^*_k.$$ 

Therefore the distributions of $\text{Idx}(X^1_n), \text{Idx}(X^2_n)$ are both minorized by $cu$ times the same probability vector, $a^*_k$. Combining these two observations, there exists a coupling $X^1_n, X^2_n$ of $\nu^1Q^n, \nu^2Q^n$ such that $P(\text{Idx}(X^1_n) = \text{Idx}(X^2_n)) \geq cu$. Now, by (A2), conditioned on two particles starting in the same stratum $A_k$, there exists a coupling such that the probability of their exit points being equal is $\geq c$, because their exit distributions are both minorized by $c \xi_k$. Therefore, there exists a coupling $X^1_{n+1}, X^2_{n+1}$ of $\nu^1Q^{n+1}, \nu^2Q^{n+1}$ such that

$$P(X^1_{n+1} = X^2_{n+1}) \geq P(\text{Idx}(X^1_{n+1}) = \text{Idx}(X^2_{n+1})) \cdot P(X^1_n = X^2_n | \text{Idx}(X^1_n) = \text{Idx}(X^2_n)) \geq cu \cdot c = c^2 u.$$ 

Therefore, $\| \nu^1Q^{n+1} - \nu^2Q^{n+1} \| \leq 1 - c^2 u = (1 - c^2 u) \| \nu^1 - \nu^2 \|$. 

\[ \square \]

5.3 Basic Version - Long Term Convergence Rate

Proving our more precise estimate of the rate of convergence in the long term will require more machinery. As we saw above, assumption (A2) allows us to show that coupling occurs if two particles start in the same stratum. If the difference between two injection measures, $\nu^1$ and $\nu^2$, is mostly in their individual injection measures (i.e. the $\nu_j$), but their weights are mostly the same, then we can conclude that they
get closer when evolved by $Q$. This is because two particles distributed by $\nu^1$, $\nu^2$ which are not already coupled have a good chance of being in the same stratum, and then (A2) says they have a chance of coupling on the next exit. But what if the difference is mostly in the weights? Then the particles have little chance to couple on the next exit, since they are most likely not in the same stratum. So we can’t conclude that $Q$ is a contraction. This can be thought of as the “problem case”, as far as a coupling argument is concerned.

We solve this issue with two tools. First, given the equilibrium $\nu^*$ and a given measure $\nu$, we want to decompose them into the the two above cases. We can start with the standard way of decomposing the measures into equal and mutually singular parts:

$$
\nu = (1 - \epsilon)\bar{\nu} + \epsilon \nu^1
$$

$$
\nu^* = (1 - \epsilon)\bar{\nu} + \epsilon \nu^2
$$

where $\bar{\nu}$, $\nu^1$, and $\nu^2$ are probability measures with $\nu^1 \perp \nu^2$ and $\epsilon = \| \nu - \nu^* \|$. We can further decompose the mutually singular parts into a part that looks like the “easy case” above, and a part that looks like the “problem case”. We get the following expression:

$$
\nu^1 = (1 - \gamma) \sum_j b_j \nu_j + \gamma \sum_j \eta_j j
$$

$$
\nu^2 = (1 - \gamma) \sum_j b^*_j \nu^*_j + \gamma \sum_j \eta^*_j j
$$

for some probability vectors $p$, $b^1$, $b^2$ and measures $\eta^1$, $\eta^2$ where $b^1 \perp b^2$, $\eta^1 \perp \eta^2$, and some $\gamma \in [0, 1]$. The first term represents the parts of $\nu^1$, $\nu^2$ that have different weights. The second term represents the parts that have the same weights but orthogonal distributions on the strata. $\gamma$ is the share of the second term relative to all of the TV distance between $\nu$ and $\nu^*$. We will handle the cases where $\gamma$ is large or small separately, and get a bound for each.

However, as explained above, $Q$ may not actually bring the two measures closer together in the “problem case”, where $\gamma << 1$. To fix this, we introduce the following new metric, which is equivalent to TV distance in the sense that they induce the same topology:

$$
d_w(\nu, \nu^*) = \| \nu - \nu^* \|_{TV(A)} + \alpha d_w(a, a^*)
$$

where $a, a^*$ are the weight vectors of $\nu, \nu^*$, and $G^*$ is the transition matrix between strata in equilibrium. The metric $d_w$ on weight vectors is given by

$$
d_w(a, b) = \sum\limits_{k=0}^{m} \lambda^k \| a(G^*)^k - b(G^*)^k \| + \sum\limits_{k=m+1}^{\infty} \lambda^m \| a(G^*)^k - b(G^*)^k \|
$$

where $\lambda, m$ are as in (B2). The intuition behind the new metric is as follows. Even if we are in the problem case, the weights of our injection measure should get closer to the true weights, which should allow for coupling to occur at some later step. So we add a term to the standard TV-distance that should contract if the weights move approximately by $G^*$ under $Q$. We choose the metric $d_w$ on probability vectors because under it, $G^*$ is a contraction with constant $\lambda$.

With the decomposition and metric defined, our strategy is now as follows: First, we consider the “easy case”, where $\gamma$ is close to 1. Then (A1) lets us conclude that coupling occurs on the next exit, so $\nu$ gets closer to $\nu^*$ when acted on by $Q$. In the “problem case”, where $\gamma$ is close to 0, the weights of $\nu$ get closer to those of $\nu^*$ because they move approximately by $G^*$, so we still get contraction in $d_w$. The complication we encounter is that the weights do not move exactly by $G^*$, but by the transition matrix $G$ given by $\nu$. The result is that we only get local contraction of $Q$ in $d_w$. But we already have global convergence by [3,2]. So we still get a rate of convergence that eventually applies.

With the strategy now laid out, we can now proceed through the proof of [3,3]. We begin by showing the decomposition [2].

23
Lemma 5.3. Let $\nu = \sum a_j \nu_j$ be an injection measure on $A$, and let $\nu^* = \sum_j a_j^* \nu_j^*$ be the fixed point of $Q$ then there exist probability vectors $p$, $b^1 \perp b^2$, measures $\eta^1 \perp \eta^2$ and $\gamma \in [0,1]$ such that (3) holds.

Proof. Start by decomposing the weight vectors of $\nu$, $\nu^*$ into equal and orthogonal parts:

$$a = (1 - \hat{\epsilon})a + \hat{\epsilon}b^1 \quad \quad a^* = (1 - \hat{\epsilon})\bar{a} + \hat{\epsilon}b^2.$$

with $b^1 \perp b^2$. Similarly, we can decompose the injection measures on each stratum:

$$\nu_j = (1 - \epsilon_j)\bar{\nu}_j + \epsilon_j \eta_j^1 \quad \quad \nu_j^* = (1 - \epsilon_j)\bar{\nu}_j + \epsilon_j \eta_j^2.$$

We can now substitute these expressions for the injection measures and weights into $\nu = \sum a_j \nu_j$, $\nu^* = \sum_j a_j^* \nu_j^*$. We get

$$\nu^0 = (1 - \hat{\epsilon})(1 - Z) \left[ \sum_j \bar{a}_j \left( 1 - \epsilon_j \right) \nu_j \right] + \hat{\epsilon} \left[ \sum_j b^1_j \nu_j \right] + (1 - \hat{\epsilon})Z \left[ \sum_j p_j \eta_j^1 \right]$$

$$\nu^* = (1 - \hat{\epsilon})(1 - Z) \left[ \sum_j \bar{a}_j \left( 1 - \epsilon_j \right) \nu_j \right] + \hat{\epsilon} \left[ \sum_j b^1_j \nu_j^* \right] + (1 - \hat{\epsilon})Z \left[ \sum_j p_j \eta_j^2 \right]$$

where $Z = \sum_j \bar{a}_j \epsilon_j$ and $p_j = \frac{\lambda_j + \epsilon}{2\lambda}$. Setting $\gamma = \frac{(1 - \hat{\epsilon})Z}{(1 - \hat{\epsilon})Z}$, the result follows.

□

As mentioned above, our proof strategy is complicated by the fact that the weight vectors of two injection measures move by different matrices under $Q$. However, if all $a_j^*$ are non-zero, i.e. every stratum has some weight in equilibrium, then $G, G^*$ will be close if $\nu, \nu^*$ are close in total variation. So our next step is to bound the difference in weight vectors after applying $Q$, in the case where the transition matrices are close.

Lemma 5.4. Suppose that (A0)-(A2) and (B2) hold. If $G, G^*$ are the transition matrices given by the exit measures for $\nu, \nu^*$, and if $\| G - G^* \|_\infty < \delta$, then

$$d_w(aG, a^* G^*) \leq S ((1 - \epsilon) \gamma + \delta (1 - \gamma)) \| \nu - \nu^* \|_{TV} + \lambda d_w(a, a^*).$$

(3)

Where $S = 1 + \lambda + \lambda^2 + \lambda^3 + \cdots = \frac{1}{1 - \lambda}$.

Proof. By decomposing the weight vectors $a$, $a^*$ into equal and orthogonal parts, as in the proof of Lemma 5.3, we have

$$aG - a^* G^* = a(G - G^*) + (a - a^*)G^* = (1 - \hat{\epsilon})\bar{a}(G - G^*) + \hat{\epsilon}\bar{b}^1(G - G^*) + (a - a^*)G^*.$$

Therefore, by the triangle inequality,

$$d_w(aG, a^* G^*) \leq (1 - \hat{\epsilon})d_w(\bar{a}G, \bar{a}G^*) + \hat{\epsilon}d_w(\bar{a}G^*, a^* G^*) + \hat{\epsilon}d_w(\bar{b}^1G, b^1G^*).$$

Now, $G^*$ is a contraction in $d_w$, so $d_w(a^* G^*, a^* G^*) \leq \lambda d_w(a^0, a^*)$. The construction of $d_w$ and assumption (B2) imply that

$$d_w(b^1G, b^1G^*) \leq S \| b^1 G^* - b^1 G^* \|$$

which is $\leq S \| G - G^* \|_\infty \leq s \delta$ because $b^1$ is a probability vector.
Using the definition of $G, G^*$ and assumption (A2),
\[
\left| (\bar{a}(G - G^*))_j \right| = \left| \sum_k \bar{a}_k (G_{jk} - G^*_{jk}) \right| = \left| \sum_k \bar{a}_k (\xi_k(A_j) - \xi_k(A_j)) \right| \\
\leq \sum_k \bar{a}_k \left\| \xi_j - \xi_j^* \right\| \leq (1 - c) \sum_k \bar{a}_k \left\| \nu_j - \nu_j^* \right\|_{TV}.
\]
Therefore,
\[
(1 - \hat{c}) \left\| \bar{a}(G - G^*) (G^*)^k \right\|_{TV} \leq (1 - c)(1 - \hat{c}) \sum_k \bar{a}_k \epsilon_k \\
= (1 - c) \gamma \left\| \nu - \nu^* \right\|_{TV}
\]
Which implies that $(1 - \hat{c}) d_{\alpha}(aG, aG^*) \leq (1 - c) \gamma S \left\| \nu - \nu^* \right\|_{TV}$. Putting the inequalities just derived together gives the desired result.

Now we are ready to show that $Q$ acts as a local contraction in $d_{\alpha}$, for both small and large $\gamma$. We start with the case where $\gamma$ is bounded away from 0. In this case, the idea is that, if two particles coupling $\nu$ and $\nu^*$ are not at equal positions, they have a chance, bounded below, of being in the same stratum. (A2) then lets us assure that they can couple when moving by $Q$. The rest of the proof of the following lemma is simply keeping track of how all the terms in $d_{\alpha}(\nu Q, \nu^* Q)$ behave.

**Lemma 5.5.** Suppose that (A0)-(A2) and (B2) hold and fix $\beta \in (0,1)$. Then for initial $\nu$ and equilibrium $\nu$ injection measures, if $\gamma > \beta$ and $\parallel G - G^* \parallel_{\infty} < \delta$,
\[
d_{\alpha}(\nu Q, \nu^* Q) \leq \max (1 - \beta c + \alpha S (1 - c) + \alpha S \delta, \lambda) \cdot d_{\alpha}(\nu, \nu^*).
\]

**Proof.** Suppose $\nu \neq \nu^*$, since otherwise the result is trivial. Let $X, X^*$ be a coupling of $\nu, \nu^*$. By Lemma 5.3, the coupling can be chosen so that with probability $\gamma \parallel \nu - \nu^* \parallel$, $\text{Idx}(X) = \text{Idx}(X^*)$ and $X \neq X^*$, so that
\[
P \left( \text{Idx}(X) = \text{Idx}(X^*) \mid X \neq X^* \right) = \frac{P \left( \text{Idx}(X) = \text{Idx}(X^*), X \neq X^* \right)}{P(X = X^*)} \\
= \frac{\gamma \parallel \nu - \nu^* \parallel}{\parallel \nu - \nu^* \parallel} = \gamma \\
\geq \beta.
\]
Assumption (A2) now implies that if $Y, Y^*$ are a coupling of $\nu Q, \nu^* Q = \nu^*$, then they can be chosen so that
\[
P(Y = Y^* \mid X \neq X^*) \geq c \gamma \geq c \beta.
\]
Therefore,
\[
\parallel \nu Q - \nu^* \parallel \leq (1 - c \beta) \parallel \nu - \nu^* \parallel.
\]
Combining this with Lemma 5.4 and the definition of $d_{\alpha}$ implies that
\[
d_{\alpha}(\nu Q, \nu^* Q) \leq (1 - \beta c + \alpha S (1 - c) + \alpha S \delta) \parallel \nu - \nu^* \parallel_{TV} + \alpha \lambda d_{\alpha}(a, a^*) \\
\leq \max (1 - \beta c + \alpha S (1 - c) + \alpha S \parallel G^0 - G^* \parallel_{\infty}, \lambda) d_{\alpha}(\nu^0, \nu^*).
\]

Now we can approach the case where $\gamma$ is bounded away from 1 in a similar way. The idea this time is that, if we are close to the problem case, but $\parallel G - G^* \parallel$ is small, then the weights will get more accurate under $Q$ and so the distance in $d_{\alpha}$ should still contract, even if no coupling occurs.
Lemma 5.6. Suppose that (B2) holds, and fix \( \beta \in (0, 1) \). Suppose that \( \alpha < \frac{1}{\sqrt{4 - \epsilon}} \). Then for any \( \nu, \nu^* \) such that \( \gamma \leq \beta, \| G - G^* \| < \delta \),

\[
d_\alpha(\nu Q, \nu^* Q) \leq \frac{(1 + \alpha S \delta) + (1 - \beta)\alpha \lambda}{1 + (1 - \beta)\alpha} d_\alpha(\nu, \nu^*).
\] (4)

Proof. As before, we can assume that \( \nu \neq \nu^* \). First, as in the previous proof, we have

\[\| \nu Q - \nu^* Q \| \leq (1 - \gamma c) \| \nu - \nu^* \|\]

and by the construction of \( \gamma \),

\[\| a - a^* \| = (1 - \gamma) \| \nu - \nu^* \| \geq (1 - \beta) \| \nu - \nu^* \| \]

\[\Rightarrow \| \nu - \nu^* \| \leq \frac{1}{1 - \beta} \| a^0 - a^* \| \leq \frac{1}{1 - \beta} d(a, a^*).
\]

Now, putting the above inequalities together with the result of Lemma 5.4 implies that

\[d_\alpha(\nu, \nu^*) \leq \frac{(1 + (\alpha S(1 - c) - c)) \gamma + \alpha S \delta}{1} \| \nu^0 - \nu^* \|_{TV} + \lambda d_w(a^0, a^*) \]

\[\leq \rho \beta d_\alpha(\nu^0, \nu^*)
\]

where

\[\rho = \frac{(1 + (\alpha S(1 - c) - c)\gamma_0 + \alpha S \delta)}{\| \nu - \nu^* \| + d_w(a, a^*)}.
\]

Simplifying \( \rho \) gives the desired result.

Notice that the contraction constants found in the previous two lemmas are \( < 1 \) if \( \delta \) is sufficiently small. This is why we can conclude that \( Q \) is a local contraction in \( d_\alpha \). With both cases now out of the way, we can put them together to get our next main convergence theorem.

Proof of Theorem 3.3. Theorem 3.2 implies that, under the given assumptions, \( \| \nu^n - \nu^* \| \to 0 \), where \( \nu^n = \nu Q^n \). Since, by assumption, \( a_j^* > 0 \) for all \( j \), this implies that \( \nu^0_j \to \nu^*_{j} \) for all \( j \), in total variation, and therefore \( G^n \to G^* \), where \( G^n \) is the transition matrix given by \( \nu^n \). Therefore, there exists \( l \geq 0, \delta > 0 \) such that, for all \( n \geq l \) and \( \alpha < \frac{1}{\sqrt{4 - \epsilon}}, \| G - G^* \| < \delta \) and

\[q_{\alpha, \beta}^\delta := \max \left( 1 - \beta c + \alpha S(1 - c) + \alpha S \delta, \lambda, \frac{(1 + \alpha S \delta) + (1 - \beta)\alpha \lambda}{1 + (1 - \beta)\alpha} \right) < 1.
\]

It follows that \( d_\alpha(\nu^{n+1}, \nu^*) \leq q_{\alpha, \beta}^\delta \cdot d_\alpha(\nu^n, \nu^*) \) for \( n \geq l \). Furthermore, as \( l \to \infty \), the above holds for arbitrarily small \( \delta \). Also note that for \( \lambda < 1, \frac{(1 + \alpha S \delta) + (1 - \beta)\alpha \lambda}{1 + (1 - \beta)\alpha} \geq \frac{1 + (1 - \beta)\alpha \lambda}{1 + (1 - \beta)\alpha} > \lambda \). Therefore, there exist \( q_n \in (0, 1) \) such that \( d_\alpha(\nu^{n+1}, \nu^*) \leq q_n \cdot d_\alpha(\nu^n, \nu^*) \) for sufficiently large \( n \), and

\[q_n \to \lim_{\beta \to 0} q_{\alpha, \beta}^\delta = \max \left( 1 - \beta c + \alpha S(1 - c), \frac{1 + (1 - \beta)\alpha \lambda}{1 + (1 - \beta)\alpha} \right).
\]

Finally, optimizing the limit of \( q_n^\delta \) over all choices of \( \alpha, \beta \) proves the desired result.
5.4 Local Convergence of the Eigenvector Version

Now we turn to proving our final result, that the eigenvector version converges locally with a rate we can bound. The strategy again revolves around the decomposition (3). This time, however, we do not have to deal with the “problem case”. The reason is that, as a result of the perturbation bound in (B3), we can show that the parameter $\gamma$ is never too small when the weights are given by the eigenvector of the transition matrix $G$ associated to $\{\nu_j\}$. Therefore, we always get that coupling can occur, by assumption (A1).

However, we also run into a new problem: this version of the algorithm cannot be represented as evolving the measure $\nu$ forward by a single Markov kernel, the way the basic version evolves by $Q$. The reason is that each of the eigenvector weights depends on all the measures $\nu_j$. The eigenvector version does not act meaningfully on just one $\nu_j$. Therefore, we cannot assure that no decoupling occurs in the step where the new weights are chosen. However, similarly to the previous proof, we can control the size of this decoupling in the case where the $\nu_j$ are already close to $\nu_j^*$. Therefore, we will end up with a local convergence result.

Our first goal is to show that, when using the eigenvector weights, the parameter $\gamma$ is bounded below. This follows somewhat straightforwardly from assumption (B3), and the fact that the transition matrix entries are themselves determined by the injection measures. Note that the eigenvector weights for the equilibrium injection measure are simply given by entries are themselves determined by the injection measures. Note that the eigenvector weights for the equilibrium injection measure are simply given by $a^*$, because by construction of the fixed point of $Q$, $a^*G^* = a^*$.

Lemma 5.7. Suppose that (B3) holds. Let $G^{*}$ be the transition matrix given by the fixed point measures $\nu_j$, with principal eigenvector $a^*$, i.e. $a^*G^* = a^*$. Let $G$ be another transition matrix on $\{1, \ldots, J\}$ with eigenvector $z$, such that $z_j > 0$ for all $j$. Then for any $\delta > 0$, there exist constants $\theta_{ik}$ for $i \neq k$ such that

$$\frac{|z_j - a^*_j|}{z_j \wedge a^*_j} \leq \sum_{i \neq k} \theta_{ik} |G_{ik} - G^*_{ik}| .$$

if $\sup_{i,k} |G_{ik} - G^*_{ik}| < \delta$. Furthermore, as $\delta \to 0$, $\theta_{ik}' \to e^{\theta_{ik}} - 1$.

Proof. By (B3), there exist $\theta_{ik} > 0 \forall i, k \in \{1, \ldots, J\}$ such that for all $j \leq J$,

$$\log \left( \max \left( \frac{z_j}{a^*_j} \right) \right) \leq \sum_{i \neq k} \theta_{ik} |G_{ik} - G^*_{ik}|$$

$$\Rightarrow z_j \wedge a^*_j = 1 + \frac{|z_j - a^*_j|}{z_j \wedge a^*_j} \leq \exp \left( \sum_{i \neq k} \theta_{ik} |G_{ik} - G^*_{ik}| \right).$$

Now, for any $M > 0$, there exists $K > 0$ such that $e^x - 1 \leq Kx$ for $0 \leq x \leq M$. Specifically, since $|G_{ik} - G^*_{ik}| \leq 1$, we have

$$\exp \left( \sum_{i \neq k} \theta_{ik} |G_{ik} - G^*_{ik}| \right) = \prod_{i \neq k} \left( 1 + \sum_{m=1}^{\infty} \frac{\theta_{ik} |G_{ik} - G^*_{ik}|^m}{m!} \right) - 1$$

$$\leq \prod_{i \neq k} \left( 1 + |G_{ik} - G^*_{ik}| \sum_{m=1}^{\infty} \frac{\theta_{ik}^m}{m!} \right) - 1 = \prod_{i \neq k} \left( 1 + |G_{ik} - G^*_{ik}| (e^{\theta_{ik}} - 1) \right) - 1$$

$$\leq \sum_{i \neq k} |G_{ik} - G^*_{ik}| \left( (e^{\theta_{ik}} - 1) + O(|G - G^*|) \right),$$

where $O(|G - G^*|)$ represents terms which are order 1 or higher in $\{ |G_{ik'} - G^*_{ik'}| \}_{i \neq k}$. Since there are only finitely many of these terms, setting $\theta_{ik}'$ to be the coefficient of $|G_{ik} - G^*_{ik}|$ in

$$\prod_{i \neq k} \left( 1 + |G_{ik} - G^*_{ik}| (e^{\theta_{ik}} - 1) \right) - 1,$$
with all terms of the form $|G_{i'k'} - G_{ik'}^*|$ replaced by 1, suffices.

Now we can use the above lemma to show that $\gamma$ is bounded in the eigenvector version.

**Lemma 5.8.** Suppose that (A0)-(A2) and (B3) hold. Let $\nu_j$ be injection measures for $1, \ldots, J$, with transition matrix $G$ and corresponding (normalized) eigenvector $z$. Let $\nu = \sum_j z_j \nu_j$, and decompose $\nu, \nu^*$ as in [4]. Then there exists $r$, independent of $\nu_j$, such that

$$\gamma \geq \frac{1}{1 + r}.$$  \hfill (5)

Furthermore, as $\nu_j \to \nu_j^*$ in TV, $r$ can be chosen to be arbitrarily close to $r^\infty$, as in Theorem 5.4.

**Proof.** First, note that by (A2), $G^*, G \geq \tilde{G}$, where $\tilde{G}$ is the transition matrix if the injection measure in $A_j$ is the QSD $\hat{\nu}_j$. Therefore, (B3) applies to $G$ and $G^*$. Now write $\hat{e} = \|z - a^*\|$ and $e_j = \|\nu_j - \nu_j^*\|$ for all $j$.

From Lemma 5.7, we have that for all $j$,

$$\frac{|z_j - a_j^*|}{z_j \land a_j^*} \leq \sum_{i \neq k} \theta_{ik} |G_{ik} - G_{ik}^*| = \sum_i \left( \sum_{k \neq i} \theta_{ik} |G_{ik} - G_{ik}^*| \right) \leq \sum_i \sup_{k \neq i} \theta_{ik} |G_{ik} - G_{ik}^*| \leq \sum_i \sup_{k \neq i} \theta_{ik} \cdot 2 \|G_i - G_i^*\| \leq \sum_i 2 \cdot \sup_{k \neq i} \theta_{ik} \cdot (1 - c) e_i \leq 2(1 - c) \sup_{i,k} \left( \frac{\theta_{ik}}{a_i^*} \right) \sum_i a_i^* e_i$$

Where the third line follows from (A2), which implies that $\|G_i - G_i^*\| \leq \|\nu_j Q - \nu_j^* Q\| \leq (1 - c) e_j$. Now, (B3) implies that

$$\frac{a_i^*}{z_i \land a_i^*} \leq \frac{z_i \lor a_i^*}{z_i \land a_i^*} \leq \exp \left( \sum_{i \neq k} \theta_{ik} |G_{ik} - G_{ik}^*| \right) \leq \exp \left( 2(1 - c) \sum_{k \neq i} \sup_{i,k} \theta_{ik} \right) =: E$$

Therefore, for all $j \in \{1, \ldots, J\}$,

$$\hat{e} \leq |z_j - a_j^*| \leq (z_j \land a_j^*) \cdot 2(1 - c) \sup_{i,k} \left( \frac{\theta_{ik}}{a_i^*} \right) \sum_i (z_i \land a_i^*) e_i \leq r \sum_i (z_i \land a_i^*) e_i$$

where

$$r = 2(1 - c) \sup_{j} \left( \frac{\theta_{jk}}{a_j^*} \right) E.$$

We also now have that $r \to r^\infty$ as $\nu \to \nu^*$, because $\theta_{ik} \to e^{\theta_{ik}} - 1$. Now, observe that in the notation of Lemma 5.3

$$\frac{(1 - \hat{e}) Z}{(1 - \hat{e}) Z + \hat{e}} = \frac{\sum_j (z_j \land a_j^*) e_j}{\sum_j (z_j \land a_j^*) e_j + \hat{e}}$$

and so the result follows. \hfill \square
We can now show that the eigenvector version acts as a local contraction on injection measures. We choose to state the local contraction in another metric, for which the calculations are simpler. Given injection measures $\nu_j$, with eigenvector weights $z_j$ and $\nu = \sum_j z_j \nu_j$, define

$$d(\nu, \nu^*) = \sum_j a_j^* \| \nu_j - \nu_j^* \| = \sum_j a_j^* \epsilon_j.$$ 

Note that, since the measures $\nu_j$ determine the weights $z_j$, this is a well-defined metric on the set of injection measures with eigenvector weights. The proof of Lemma 5.8 also implies that $\hat{\epsilon} < rd(\nu, \nu^*)$ so $d$ is equivalent to total variation distance for such choices of $\nu$. Since we are in the case where the measures determine the weights, $d$ can be thought of as the metric that only looks at the difference in measures, appropriately weighted. By moving between $d(\nu, \nu^*)$ and $\| \nu - \nu^* \|$, we can use a coupling argument and the boundedness of $\gamma$ to show local contraction in $d$, as we do below.

**Lemma 5.9.** Suppose that (A0)-(A2). Let $\nu' = \sum_j z_j' \nu_j'$ be the measure obtained from $\nu$ via the eigenvector injection measure method, with $\epsilon_j' = \| \nu_j' - \nu_j^* \|$. Then

$$\sum_j (z_j \land a_j^*) \epsilon_j' \leq q \sum_j (z_j \land a_j^*) \epsilon_j$$

Where

$$q := 1 - \frac{1}{1 + c}$$

**Proof.** First, by Lemma 5.8 and the same coupling argument as in Lemma 5.5

$$\| \nu Q - \nu^* \| \leq (1 - \gamma c) \| \nu - \nu^* \| \leq q \| \nu - \nu^* \|.$$

(6) The individual injection measures of $\nu Q$ are $\nu_j'$, by construction of the algorithm. Furthermore, the weights of $\nu Q$ are $z$, because $zG = z$. Therefore,

$$\| \nu Q - \nu^* \| = \sum_j (z_j \land a_j^*) \epsilon_j + \hat{\epsilon}.$$ 

We also have

$$\| \nu - \nu^* \| = \sum_j (z_j \land a_j^*) \epsilon_j' + \hat{\epsilon}.$$ 

Substituting the above expressions into (6), we get

$$\sum_j (z_j \land a_j^*) \epsilon_j' \leq q \sum_j (z_j \land a_j^*) \epsilon_j - (1 - q) \hat{\epsilon} \leq q \sum_j (z_j \land a_j^*) \epsilon_j.$$ 

$$\square$$

**Lemma 5.10.** Fix $\delta > 0$. Under the assumptions of Lemma 5.9, if $2(1 - c)(\sup_{j \neq k} \frac{a_k}{a_j})d(\nu, \nu^*) < \delta$, then

$$d(\nu', \nu^*) \leq q(1 + \delta)d(\nu, \nu^*)$$

29
Proof. Continuing on from Lemma 5.9, we have

\[ d(\nu', \nu^*) = \sum_j a_j^* \epsilon_j \leq \sup_j \left( \frac{z_j}{z_j + a_j^*} \right) \sum_j (z_j \wedge a_j^*) \epsilon_j \leq \frac{q_j}{z_j} \left( \sum_j (z_j \wedge a_j^*) \epsilon_j \right) \]

\[ \leq q \cdot \left( 1 + \frac{|z_j - a_j^*|}{z_j} \right) \sum_j a_j^* \epsilon_j \]

\[ \leq q \cdot \exp \left( 1 + 2(1-c) \sup_{k \neq i} (\sup \theta_{ik}) \epsilon_i \right) \]

where the second line holds by (A2), and the last line holds because \( d(\nu, \nu^*) \geq \frac{1}{\sup_{i \neq k} \frac{\theta_{ik}}{a_i}} \sum_i \sup_{k \neq i} (\theta_{ik}) \epsilon_i. \)

\[ \square \]

Note, in particular, that if \( \delta < \frac{1}{q} - 1 \), then \( q(1 + \delta) < 1 \), and so have have local contraction. It is now straightforward to use the equivalence of TV distance and \( d \), when the weights are given by the eigenvector of \( G \), to prove our final convergence result.

Proof of Theorem 3.4 Suppose that the initial injection measure is \( \nu^0 = \sum_j z_j^0 \nu_j^0 \), and that successive \( \nu^n \) are given by steps of the eigenvector version, with. Lemma 5.10 implies that, if \( 2(1-c) (\sup_{i \neq k} \frac{\theta_{ik}}{a_i}) d(\nu^0, \nu^*) < \frac{1}{q} - 1 \), then \( d(\nu^n, \nu^*) \to 0 \), and for all \( n \), \( d(\nu^{n+1}, \nu^*) \leq p^n d(\nu^n, \nu^*) \) where \( p^n \) is the value \( q(1+\delta) \), computed from \( \nu = \nu^0 \). By construction of \( p^n \), we have \( p^n \to 1 - \frac{c}{1+r^\infty} \).

Next, by the bounds derived in Lemma 5.8

\[ \| \nu^0 - \nu^* \| \leq \sum_j (z_j^0 \wedge z_j^*) \| \nu_j^0 - \nu_j^* \| \leq (1+r) \sum_j (z_j^0 \wedge z_j^*) \| \nu_j^0 - \nu_j^* \| \]

\[ \leq (1+r) d(\nu^0, \nu^*). \]

Therefore, we have convergence if \( \frac{2(1-c) (\sup_{i \neq k} \frac{\theta_{ik}}{a_i})}{1+r} \| \nu^0 - \nu^* \| < \frac{1}{q} - 1 \). Solving for \( \| \nu^0 - \nu^* \| \) and using \( 1+r > 1 + r^\infty E \) gives the bound used in Theorem 3.4

**Acknowledgments:** We thank Jonathan Weare for introducing JCM to this general area. We also thank the NSF grant DMS-1613337 and both SAMSI (DMS-1638521) and Duke TRIPODS (CFF-1934964) for partial support of this work.

**References**

[1] David Aristoff, Juan M. Bello-Rivas, and Ron Elber. A mathematical framework for exact milestoning. Multiscale Modeling and Simulation, 14(1):301–322, 2016.

[2] Grace E. Cho and Carl D. Meyer. Comparison of perturbation bounds for the stationary distribution of a markov chain. Linear Algebra and its Applications, 335(1-3):137–150, 2001.

[3] Aaron Dinner, Erik Thiede, Brian Koten, and Jonathan Weare. Stratification of markov chain monte carlo. 05 2017.
[4] Aaron R. Dinner, Jonathan C. Mattingly, Jeremy O. B. Tempkin, Brian Van Koten, and Jonathan Weare. Trajectory stratification of stochastic dynamics. *SIAM Review*, 60(4):909–938, 2018.

[5] Aaron R. Dinner, Erik Thiede, Brian Van Koten, and Jonathan Weare. Stratification as a general variance reduction method for markov chain monte carlo, 2017.

[6] Gabriel Joseph Earle. *Stratified MCMC Sampling of non-Reversible Dynamics*. PhD thesis, Duke University, 2020.

[7] Martin Hairer and Jonathan C. Mattingly. Yet another look at Harris’ ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, volume 63 of *Progr. Probab.*, pages 109–117. Birkhäuser/Springer Basel AG, Basel, 2011.

[8] Matthias Heymann. *The geometric minimum action method: A least action principle on the space of curves*. PhD thesis, New York University, 2007.

[9] Matthias Heymann and Eric Vanden-Eijnden. The geometric minimum action method: A least action principle on the space of curves. *Communications on Pure and Applied Mathematics*, 61(8):1052–1117, 2008.

[10] Stephen J. Kirkland, Michael Neumann, and Nung-Sing Sze. On optimal condition numbers for markov chains. *Numerische Mathematik*, 110(4):521–537, 2008.

[11] Robert S. Maier and D. L. Stein. A scaling theory of bifurcations in the symmetric weak-noise escape problem. *Journal of Statistical Physics*, 83(3-4):291–357, 1996.

[12] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, 101(2):185–232, 2002.

[13] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993.

[14] Erik Thiede, Brian Van Koten, and Jonathan Weare. Sharp entrywise perturbation bounds for markov chains. *SIAM Journal on Matrix Analysis and Applications*, 36(3):917–941, 2015.

[15] Konstantin S. Turitsyn, Michael Chertkov, and Marija Vucelja. Irreversible monte carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414, 2011.