A Quotient Construction on Markov Chains with Applications to the Theory of Generalized Simulated Annealing

John R. Wicks
Department of Computer Science
Brown University, Box 1910
Providence, RI 02912

Amy Greenwald
Department of Computer Science
Brown University, Box 1910
Providence, RI 02912

Abstract

In an earlier paper [14], we developed the first algorithm (to our knowledge) for computing the stochastically stable distribution of a perturbed Markov process. The primary tool was a novel quotient construction on Markov matrices. In this paper, we show that the ideas and techniques in that paper arise from a more fundamental construction on Markov chains, and have much wider applicability than simply to game theory (the application discussed in [14]). Besides leading to new results, our quotient construction leads to simpler proofs of known results and to simpler algorithms for known computations. In this paper, we present one example of the former—we give necessary and sufficient conditions for a Markov matrix to have a unique stable distribution—and one of the latter—we show that a variant of the algorithm in [14] can be used to compute the virtual energy levels of a generalized simulated annealing in a straightforward, recursive manner using basic matrix arithmetic.

1 Introduction

Markov processes are fundamental to a wide variety of techniques in artificial intelligence. Much of the time, they arise as discrete-time, finite-state, stationary Markov processes (i.e., Markov chains), which are fully determined by a transition matrix and an initial distribution. For example, in an earlier paper [14], we analyzed the dynamics of Young's adaptive learning model in repeated games [16], which models the behavior of players over time as a Markov chain. In fact, we considered the more general case of a parameterized family of Markov processes, a so-called *perturbed* Markov process (PMP), and developed the first algorithm (to our knowledge) for computing the *stochastically stable distribution* of a PMP. The primary tool was a novel *quotient* construction on Markov matrices.

In this paper, we show that the ideas and techniques in that paper arise from a more fundamental construction on Markov chains and have much wider applicability than simply to game theory. Besides leading to new results, our quotient construction leads to simpler proofs of known results and to simpler algorithms for known computations. We present one example of the former—we give necessary and sufficient conditions for a Markov matrix to have a unique stable distribution—and one of the latter—we present a novel algorithm to compute the virtual energy levels of a generalized simulated annealing. Along the way, we highlight how several properties of the quotient construction, such as its *naturality* and its relationship to random walks, play a fundamental role in proving the correctness of our algorithm.

In the area of optimization, simulated annealing algorithms [6, 10] depend heavily on the theory of convergence of Markov processes to justify their usefulness. The analysis of parallelized schemes

for simulated annealing have led to the study of *generalized simulated annealing* (GSA) [9, 13]. One fundamental problem in the study of GSA is the computation of *virtual energies*, given the transition matrix of the process. While [3] and [13] provide algorithms to compute them, their algorithms are rather arcane. It turns out that the theory of GSA is very closed related to that of PMPs. We derive a variant of the algorithm from [14] which computes these energy levels in a straightforward, recursive manner using basic matrix arithmetic.

A fundamental observation in both the PMP and GSA applications is that both calculations may be performed on *equivalence classes* of perturbed matrices. The entries of the transition matrix of a perturbed Markov process are technically functions of a parameter, which in the former case is interpreted as an error probability, while in the latter corresponds to a "temperature". However, the stochastically stable distribution of a PMP and the virtual energies of a GSA depend only on certain coefficients of the entries. In particular, we show that the calculations may be carried out on real-valued matrices.

After establishing some notation and basic definitions in Section 2, in Section 3, we review the quotient construction for Markov matrices that we developed in [14] and highlight some of its key properties. In Section 5, we show how this construction arises from a more general construction on Markov processes, which allows us to prove additional, useful properties of the original construction on matrices. In Section 6, we apply the construction to PMPs as the basis for our algorithm to compute the virtual energies of a GSA, highlighting the key role played by the results from Section 5.

2 Markov Matrices

In this section, we establish some notation and review basic definitions associated with Markov matrices. For this paper, $J = [1, \ldots, 1]$ will ambiguously denote a row vector of 1's of arbitrary length. If $S \subset \mathbb{R}$ is a set of scalars, $\mathrm{Mat}_n(S)$ will denote the set of $n \times n$ matrices with entries in S. $S_n = \{1, \ldots, n\}$ will denote the index set for $n \times n$ matrices and $B_n = \{e_i \mid i \in S_n\}$ the standard basis vectors for $\mathbb{R}^n$.

Any $M \in \mathrm{Mat}_n(\mathbb{R}^+)$ (i.e., $M \geq 0$) is called *Markov* iff $JM = J$, i.e., all columns sum to 1. Likewise, a *distribution* is a vector $v \geq 0$ *s.t.* $Jv = 1$. Given a Markov matrix M, a *stable* distribution of M is one which is also an eigenvector with eigenvalue 1: i.e., $Mv = v$. If $\dim M = n$ and Δ is the standard n-simplex, then the set of stable distributions of M, $\mathrm{stab}(M) = \ker(M - I) \cap \Delta$.

We will say that M_2 is *D-equivalent* to M_1 iff there exists an injective mapping D such that:

$$\ker(M_1 - I) = D \ker(M_2 - I) \tag{1}$$

This equivalence condition says that D maps $\ker(M_2 - I)$ onto $\ker(M_1 - I)$, implying that D is in fact a bijective mapping between the two kernels. Note that, in general, D-equivalence is *not* an equivalence relation. It is a partial order, since it is not symmetric, but it is reflexive (choose $D = I$) and transitive (if M_2 is D-equivalent to M_1 and M_3 is D'-equivalent to M_2, then M_3 is DD'-equivalent to M_1). However, this terminology is justified by the observation that, if $D \geq 0$ and has a positive left-inverse, $D^*(v) = \frac{Dv}{\|Dv\|_1}$ gives 1-1 correspondence between $\mathrm{stab}(M_2)$ and $\mathrm{stab}(M_1)$.

In other words, if we are only interested in calculating stable distributions, we may work equally well with either M_1 or M_2. If we restrict attention to matrices of a fixed dimension, this is, in fact, an equivalence relation. For example, we may take D to be a diagonal matrix with sufficiently small, positive entries to "scale" the columns of a Markov matrix as follows, $M_2 = (M_1 - I)D + I$.

More importantly, if we can find a non-square, D which gives such an equivalence, we may reduce the dimensionality of the problem in our search for stable distributions.

We can associate a weighted graph $\mathcal{G}(N) = (V, E)$ with a square matrix, N, in two ways, depending on whether $N \in \mathrm{Mat}_n([0,1])$ or $\mathrm{Mat}_n([0,\infty])$. In either case, we take $V = B_n$. If $N \in \mathrm{Mat}_n([0,1])$, we define the edge set so that $(e_i, e_j) \in E$ iff $N_{j,i} > 0$; if $N \in \mathrm{Mat}_n([0,\infty])$, $(e_i, e_j) \in E$ iff $N_{j,i} < \infty$. In either case, we take the weight on edge (e_i, e_j) to be $N_{j,i}$. In the former case, we interpret the edge weights as probabilities (i.e., we cannot utilize an edge with 0 probability), while in the later, we interpret them as costs (i.e., we cannot utilize an infinitely expensive edge).

For any graph, $\mathcal{G} = (V, E)$, we will call the vertex set of a strongly connected component (SCC) a *communicating class* of $\mathcal{G}$. In addition, we will say that a set of vertices, $\mathcal{S}$, is *invariant* iff $\mathcal{S}$ has no outgoing edges, i.e., $\forall\,(e_i, e_j) \in E, e_i \in \mathcal{S} \Rightarrow e_j \in \mathcal{S}$. We will call an invariant communicating class a *closed* class. Vertices that are not in any closed class are called *transient*. Any set of vertices that does not contain a closed class will be called *open*.

Using the natural correspondence, $S_n \leftrightarrow V$, we may carry over the terminology of communicating classes, closed classes, open, invariant, and transient sets of vertices in $\mathcal{G}(M)$ and apply it to sets of indices of M. Given a set of indices, $s \subset S_n$, we will denote the corresponding set of vertices, $V_s = \{e_i \mid i \in s\}$. A Markov matrix is said to be *reducible* if it possesses more than one communicating class; otherwise it is said to be *irreducible*. More generally, since every Markov matrix has at least one closed class, we will say that it is *regular* if it possesses exactly one closed class.

3 The Quotient Construction on Markov Matrices

In this section, we review the quotient construction of [14], which takes a Markov matrix, M, and an open set of indices, s, to produce an equivalent Markov matrix of strictly smaller dimension. Given a set of indices $s \subset S_n$ we may uniquely enumerate both it and its complement in increasing order to obtain sequences $(s_i)_{i=1}^{n-k}$ and $(\overline{s}_i)_{i=1}^{k}$. We may also define matrices, $i_s = \begin{bmatrix} e_{s_1} & \cdots & e_{s_k} \end{bmatrix}$ and $\pi_s = i_s^t$, with corresponding definitions for $\overline{s}$. Then $P_s = \begin{bmatrix} i_{\overline{s}} & i_s \end{bmatrix}$ is a permutation matrix such that if $j = s(u)$, $P_s e_u = e_j$; likewise, if $j = \overline{s}(u)$, $P_s e_{n-k+u} = e_j$.

Given a matrix, M, and a set of indices s, we may form the sub-matrices: $\widetilde{M} = \pi_{\overline{s}} M i_{\overline{s}}$, $\overline{M} = \pi_s M i_s$, $\overline{N} = \pi_{\overline{s}} M i_s$, and $\widetilde{N} = \pi_s M i_{\overline{s}}$. Notice then that $\begin{pmatrix} \widetilde{M} & \overline{N} \\ \widetilde{N} & \overline{M} \end{pmatrix} = \begin{bmatrix} \pi_{\overline{s}} \\ \pi_s \end{bmatrix} M \begin{bmatrix} i_{\overline{s}} & i_s \end{bmatrix} = P_s^t M P_s$. We will refer to this collection of sub-matrices as a *partitioning of M with respect to s*.

We state without proof the following basic result characterizing open sets of indices of a Markov matrix, M, in terms of the corresponding partitioning.

Lemma 3.1 *Consider an $n \times n$ Markov matrix, M. If $\overline{M}$ is defined by the partitioning of M with respect to a set of indices, $s \subset S_n$, then s is open with respect to M iff $I - \overline{M}$ is invertible. In that case, $\left(I - \overline{M}\right)^{-1} = \lim_{i\to\infty} \sum_{j=0}^{i-1} \overline{M}^j$.*

If s is open, with $k = |\overline{s}|$, we may define the following: a $k \times n$-dimensional matrix

$$p = \begin{pmatrix} I & \overline{N}\left(I - \overline{M}\right)^{-1} \end{pmatrix} P_s^t \tag{2}$$

and an $n \times k$-dimensional matrix

$$i = P_s \begin{pmatrix} I \\ \left(I - \overline{M}\right)^{-1} \widetilde{N} \end{pmatrix} \tag{3}$$

Letting $\widehat{M} = p(M - I)i + I$, we will call the triple, $(\widehat{M}, p, i)$, the *quotient* of M with respect to s. We will often refer to p and i as the *projection* and *inclusion* operators of the quotient (since they are surjective and injective mappings, respectively), and call simply $\widehat{M}$ the quotient. Notice that, by multiplying out, $\widehat{M}$ may also be written as $\widetilde{M} + \overline{N}\left(I - \overline{M}\right)^{-1} \widetilde{N}$.

We will see in Section 5, if we consider a Markov process, X_*, with transition matrix, M, and any initial distribution, $\widehat{M}$ corresponds to another Markov process, $\widehat{X}_*$, which is just X_*, except that we pass through states of s without pause. In particular, $\widehat{M}$ is a Markov matrix. In addition, we may then identify the entries of $\widehat{M}$ as the probability of a random walk on $\mathcal{G}(M)$ traversing a path between vertices in $V_{\overline{s}}$, while only passing through vertices of V_s. We will likewise obtain a compelling probabilistic interpretation of p.

While there is no obvious such interpretation of i, it possesses the following important properties. As we showed in [14], this construction "preserves" the set of stable distributions, in the following sense.

Theorem 3.2 *Given a Markov matrix M, $\widehat{M}$ is i-equivalent to M. In particular, i induces a bijective mapping between the stable distributions of $\widehat{M}$ and those of M via i.*

We say that $w \in \mathbb{R}^n$ is an *extension* of a vector $v \in \mathbb{R}^k$ with respect to s iff $v = \pi_{\overline{s}} w$.

Theorem 3.3 *Given a Markov matrix M, the eigenvector $iv \in \ker(M - I)$ is the unique extension with respect to s of any eigenvector $v \in \ker\left(\widehat{M} - I\right)$.*

We now observe that this construction gives a simple and direct proof of the uniqueness of stable distributions in a very general setting, without restrictive assumptions of aperiodicity or ergodicity, etc.

Corollary 3.4 *Every regular Markov matrix M has a unique stable distribution v.*

Proof 3.4 Let $s = S_n - j$, where j is any element of the unique closed class of M, and let $\widehat{M}$ be the quotient of M with respect to s. Since $\widehat{M} \in \mathrm{Mat}_1(\mathbb{R}^+)$, $\widehat{M} = (1)$. By Theorem 3.2, $\dim\ker(M - I) = \dim\ker\left(\widehat{M} - I\right) = 1$. In particular, $|\mathrm{stab}(M)| = 1$. □

A useful variant of our quotient construction will be employed in Section 6. Give a quotient $(\widehat{M}, p, i)$, of M, let d be the diagonal matrix such that $Jd = Ji$: that is, the diagonal entries of d correspond to the column sums of i. We refer to d as the *normalizing matrix* of the quotient, and define the normalized inclusion operator $i^* = id^{-1}$, and the *normalized quotient*, $(\widehat{M}^*, p, i^*)$, with $\widehat{M}^* = \left(\widehat{M} - I\right)d^{-1} + I$. Note that $Ji^* = Jid^{-1} = Jdd^{-1} = J$, that is, the columns of i^* sum to 1.

By definition, the normalized quotient is the result of "scaling" the columns of the quotient by $D = d^{-1}$. Thus, the analog of Theorem 3.2 holds for normalized quotients. Moreover, if s is a maximal, open subset of indices, we can show that the columns of i^* give the vertices of $\mathrm{stab}(M)$. It then follows easily that the converse of Corollary 3.4 holds as well.

We conclude by stating a useful geometric property of the quotient construction. Intuitively, it says that the "quotient" of an open set is open.

Theorem 3.5 *If s is an open set of indices of M and $\widehat{M}$ is the quotient of M with respect to s, and $s \cup s'$ is open with respect to M, then $\overline{s}^{-1}(s')$ is open with respect to $\widehat{M}$. In particular, if M is regular, then $\widehat{M}$ is regular. Likewise, if M is irreducible, then $\widehat{M}$ is irreducible.*

4 Markov Chains

In this section, we will review the basic definitions regarding finite-state, stationary, Markov chains, assuming the reader is familiar with basic probability and measure theory. A *discrete-time stochastic process* (or *chain*) is a sequence of random variables, $\{X_t\}_{t=0}^{\infty}$, i.e., real-valued measurable functions on some shared probability space, (Ω, μ). As is common, we will write $\Pr[\omega]$ for the probability of a measurable subset $\omega \subset \Omega$. Likewise, given a random variable, X, we will write $\Pr[X \in \sigma]$ for $\Pr[X^{-1}(\sigma)]$, assuming that $\sigma \in \mathcal{B}$, the so-called Borel sets of $\mathbb{R}$. In this way, we avoid explicit reference to Ω and μ. Likewise, if $\{x\} \in \mathcal{B}$, we will write $\Pr[X = x]$ for $\Pr[X \in \{x\}]$. The *support*, supp_X, of a random variable, X, is the smallest Borel set, σ, such that $\Pr[X \in \sigma] = 1$. in this paper, we will restrict attention to those chains whose *state space*, $\mathcal{S} = \bigcup_i \mathrm{supp}_{X_i}$, is a finite set.

A chain, $\{X_t\}_{t=0}^{\infty}$, is *Markov* iff for all t and $s_0, \ldots, s_t, s_{t+1} \in \mathcal{S}$, such that $\Pr[X_t = s_t, \ldots, X_0 = s_0] \neq 0$, $\Pr[X_{t+1} = s_{t+1} \mid X_t = s_t, \ldots, X_0 = s_0] = \Pr[X_{t+1} = s_{t+1} \mid X_t = s_t]$. This so-called *Markov property* (sometimes called the memoryless property) implies that the probability of transitions to future states, such as s_{t+1}, depend only on the present state s_t, and so are independent of the remote past, namely $s_{t-1}, \ldots, s_0$.

A Markov chain is *stationary* iff $\Pr[X_{t+1} = s_{t+1} \mid X_t = s_t]$ is constant over $\{t \mid \Pr[X_t = s_t] > 0\}$. Since $\forall s \in \mathcal{S}$, $\exists t_s \geq 0$ such that $\Pr[X_{t_s} = s] > 0$, given a labeling of the state space, i.e., a bijection $\iota : S_n \to \mathcal{S}$, there is a unique matrix, M, such that $\Pr[X_{t+1} = \iota(i) \mid X_t = \iota(j)] = M_{i,j} = e_i^t M e_j$, whenever $\Pr[X_t = \iota(j)] > 0$. We will refer to M as the *transition matrix* of the chain *consistent with* ι.

Notice that if M_1 and M_2 are two transition matrices, consistent with ι_1 and ι_2, respectively, then $M_2 = P^{-1} M_1 P$, where P is the permutation matrix such that $P_{i,j} = 1$ iff $\iota_1(i) = \iota_2(j)$. In particular, there is a unique ι-consistent transition matrix for which ι is an increasing function. Notice also that for every sequence, i_*, of length $k+1$ taking values in S_n,

$$\Pr[X_j = \iota(i_0), \ldots X_{j+k} = \iota(i_k)] = M_{i_*} \Pr[X_j = \iota(i_0)] \tag{4}$$

where $M_{i_*} = \prod_{t=0}^{k-1} M_{i_{t+1}, i_t}$. We will refer to such a sequence as a *walk* of length k.

It is often helpful to view a stationary Markov chain with transition matrix M as a random walk on the weighted graph $\mathcal{G}(M)$, where the state, $\iota(i)$, corresponds to the vertex, e_i. We may sample from this random walk by first choosing an initial vertex according to the initial distribution (i.e., the distribution of X_0), and then choosing each successive vertex according to the distribution given by the weights of the edges originating at the current vertex. Such a random path of length N gives a sample from the joint distribution of $\{X_t\}_{t=0}^{N}$. This sequence of joint distributions is sufficient to uniquely characterize the chain, up to relabeling of the states. By Equation 4, these distributions in turn are uniquely characterized by its transition matrix and the distribution of X_0. Conversely, it is well-known that given any Markov matrix, $M \in \mathrm{Mat}_n(\mathbb{R})$, and initial distribution, $v_0 \in \mathbb{R}^n$, we may construct an associated chain, $\{X_t\}_{t=0}^{\infty}$ taking values in $\{1, \ldots, n\}$ (cf. [2, p. 231-3]), with transition matrix, M, such that $\Pr[X_0 = i] = v_i$.

As before, we may carry over the terminology of communicating classes, closed classes, invariant and transient sets of vertices in $\mathcal{G}(M)$ from Section 2 and apply it to sets of states of a stationary Markov process. Notice that a subset of states is invariant iff the probability of ever transitioning away from the set is 0. Likewise, any transient state has a positive probability of transitioning away from it without ever returning.

5 The Quotient Construction on Markov Chains

We now show how the construction of Section 3 corresponds to a *quotient construction* on finite-state, stationary Markov chains. Given a chain, $\{X_t\}_{t=0}^{\infty}$, and a Borel set, $\sigma \in \mathcal{B}$, we may define a new chain, $\left\{\tilde{X}_t\right\}_{t=0}^{\infty}$, where we "collapse" the time spent in σ. To make this precise, let $\kappa_\sigma(t,\omega) = |\{k \mid X_k(\omega) \notin \sigma\}|$, and, $\tau_{\sigma,k}(\omega) = \min\{t \mid \kappa_\sigma(t,\omega) > k\}$, where $\tau_{\sigma,k}(\omega) = \infty$ if this set is empty. In other words, $\tau_{\sigma,k}$ is the $k+1$st "hitting time" for $\overline{\sigma}$. This is a Markov time, since $\{\tau_{\sigma,k} = t\}$ may be expressed solely in terms of $X_0, \ldots, X_t$.

For completeness, we prove the following basic characterization of open sets of states.

Lemma 5.1 *A set of states, σ, of a finite-state, stationary Markov process, $\{X_t\}_{t=0}^{\infty}$, is open iff $\Pr\left[\bigcap_{r\geq j} X_r^{-1}(\sigma)\right] = 0, \forall j$. In particular, if σ is open, $\tau_{\sigma,0} < \tau_{\sigma,1} < \cdots < \infty$ are stopping times.*

Proof 5.1 Assume that the process is ι-consistent with a matrix, M with state space, $\mathcal{S}$, and let $s = \iota^{-1}(\sigma)$. Let $\mathcal{I}$ denote the set of all sequences of length $q+1$ taking values in $S_{|s|}$. Partition this set according to the starting and ending values of each sequence, so that $\mathcal{I}_{u,v} = \{i_* \in \mathcal{I} \mid i_0 = u,\ i_q = v\}$. Then

$$\begin{aligned}
\Pr\left[\bigcap_{r=j}^{j+q} X_r^{-1}(\sigma)\right] &= \Pr[X_j \in \sigma, \ldots, X_{j+q} \in \sigma] = \Pr[X_j \in \sigma \cap \mathcal{S}, \ldots, X_{j+q} \in \sigma \cap \mathcal{S}] \\
&= \sum_{i_* \in \mathcal{I}} \Pr[X_j = \iota(s_{i_0}), \ldots X_{j+q} = \iota(s_{i_q})] = \sum_{i_* \in \mathcal{I}} M_{i_*} \Pr[X_j = \iota(s_{i_0})] \\
&= \sum_{u,v} \sum_{i_* \in \mathcal{I}_{u,v}} M_{s_{i_*}} \Pr[X_j = \iota(s_{i_0})] = \sum_{u,v} \sum_{i_* \in \mathcal{I}_{u,v}} \overline{M}_{i_*} \Pr[X_j = \iota(s_{i_0})] \\
&= \sum_{u,v} e_v^t \overline{M}^q e_u \Pr[X_j = \iota(s_u)] = \sum_u J\overline{M}^q e_u \Pr[X_j = \iota(s_u)]
\end{aligned}$$

which is a sum of positive terms. For each u, there is some j for which $\Pr[X_j = \iota(s_u)] > 0$. In particular, if σ is open, we must have $\lim_{q\to\infty} J\overline{M}^q e_u = 0$ for all u, so that $\lim_{q\to\infty} \overline{M}^q = 0$, and conversely.

By definition, $\tau_{\sigma,t}(\omega) < \tau_{\sigma,t+1}(\omega)$, unless both equal ∞. However, if σ is open, this occurs with probability 0. In particular, $\Pr[\tau_{\sigma,t} < \infty] = 1$, so that $\tau_{\sigma,t}$ is a stopping time. □

It is well-known that evaluating a Markov chain at a stopping time is a random variable (cf. [12]). Thus, if σ is open, we may define $\tilde{X}_t = \pi_{\sigma,t}(X_*) = X_{\tau_{\sigma,t}}$ (where we define $\tilde{X}_t$ arbitrarily when $\tau_{\sigma,t} = \infty$). We will show that $\pi_{\sigma,*}$ is an operator on Markov chains and corresponds directly to our earlier quotient construction on Markov matrices. It is easy to see that, for almost every element, $\omega \in \Omega$, the sequence $\tilde{X}_*(\omega)$ is the result of deleting elements in σ from $X_*(\omega)$. This implies that this quotient operator is "natural" in the following sense:

Theorem 5.2 *Given a Markov chain, $\{X_t\}_{t=0}^{\infty}$, and open Borel sets, $\sigma = \sigma_1 \cup \sigma_2$, $\pi_{\sigma,t}(X_*) = \pi_{\sigma_1,t}\pi_{\sigma_2}(X_*)$ almost everywhere.*

In order to motivate our main Theorem, consider a transition matrix, M, which is ι-consistent with a stationary, Markov chain, any set of states, $\sigma \subset \mathcal{S}$, corresponds to a set of indices, $s = \iota^{-1}(\sigma)$. If we partition M with respect to s, as in Section 3,

- $\widetilde{M}$ corresponds to the transitions among the states in $\overline{\sigma}$;
- $\overline{M}$ corresponds to the transitions among the states in σ;

- $\overline{N}$ corresponds to the transitions out of the states in σ into the states in $\overline{\sigma}$; and
- $\widetilde{N}$ corresponds to the transitions out of the states in $\overline{\sigma}$ into the states in σ.

For example, if $1 \leq i \leq k$ and $1 \leq j \leq n-k$, then $\iota(\overline{s}_j) \in \overline{\sigma}$, $\iota(s_i) \in \sigma$, and $\widetilde{N}_{i,j} = e_i^t \widetilde{N} e_j = e_i^t \pi_s M i_{\overline{s}} e_j = e_{s_i} M e_{\overline{s}_j} = M_{s_i,\overline{s}_j} = \Pr[X_{t+1} = \iota(s_i) \mid X_t = \iota(\overline{s}_j)]$. Notice that we have employed the enumerations, s_*, and $\overline{s}_*$ of the sets s and $\overline{s}$, respectively.

Now observe that we may interpret the equation $\widehat{M} = \widetilde{M} + \overline{N}\left(I - \overline{M}\right)^{-1} \widetilde{N}$ as saying the entries of $\widehat{M}$ correspond to paths between states of $\overline{\sigma}$ passing through an arbitrary number of states of σ. This suggests the following theorem.

Theorem 5.3 *Using the notation introduced above, if $\{X_t\}_{t=0}^{\infty}$ is a finite state, stationary Markov chain which is ι-consistent with transition matrix, M, and σ is open, then $\{\pi_{\sigma,t}(X_*)\}_{t=0}^{\infty}$ is a stationary Markov chain ι'-consistent with transition matrix, $\widehat{M} = p(M-I)i + I$, where $\iota'(k) = \iota(\overline{s}_k)$. Moreover, the projection, p, maps the distribution of X_0 to that of $\pi_{\sigma,0}(X_*)$, i.e.,*

$$\Pr\left[\pi_{\sigma,0}(X_*) = \iota'(k)\right] = \sum_j p_{k,j} \Pr\left[X_0 = \iota(j)\right]$$

Proof 5.3 For convenience, define $\widetilde{X}_t = \pi_{\sigma,t}(X_*)$. By construction, the state space for $\{\widetilde{X}_t\}_{t=0}^{\infty}$ is contained in $\widetilde{\mathcal{S}} = \mathcal{S} - \sigma$. By our notational conventions introduced earlier, s and $\overline{s}$ are increasing sequences so that $\iota(s_*)$ enumerates σ and $\iota(\overline{s}_*)$ enumerates $\widetilde{\mathcal{S}}$.

Define $\mathcal{I}_{m,l,k,j}$ as the set of all walks of length $m+l$, such that exactly $m+1$ values coincide with values of $\overline{s}$ (and l values are thus in s) with final value, $\overline{s}_k$ and initial value j, $\mathcal{I}_{m,l,k} = \bigcup_j \mathcal{I}_{m,l,k,j}$, and $\mathcal{I}_{m,k} = \bigcup_l \mathcal{I}_{m,l,k}$. Then

$$\begin{aligned}
\Pr\left[\widetilde{X}_t = \iota'(u)\right] &= \sum_l \sum_{w_* \in \mathcal{I}_{t,l,u}} \Pr\left[X_0 = \iota(w_0), \ldots X_{t+l} = \iota(w_{t+l})\right] \\
&= \sum_l \sum_{w_* \in \mathcal{I}_{t,l,u}} M_{w_*} \Pr\left[X_0 = \iota(w_0)\right] = \sum_j \sum_l \sum_{w_* \in \mathcal{I}_{t,l,u,j}} M_{w_*} \Pr\left[X_0 = \iota(j)\right]
\end{aligned}$$

Notice that for any state $\iota'(u) \in \widetilde{\mathcal{S}}$, there is some t for which $\Pr\left[X_t = \iota'(u)\right] > 0$, and hence there is some $k \leq t$ and $w_* \in \mathcal{I}_{k,t-k,u}$ such that $\Pr\left[X_0 = \iota(w_0), \ldots X_t = \iota(w_t)\right] > 0$, so that $\Pr\left[\widetilde{X}_k = \iota'(u)\right] > 0$. In particular, $\widetilde{\mathcal{S}}$ is the state space for $\{\widetilde{X}_t\}_{t=0}^{\infty}$.

If $j \in \overline{s}$, $\mathcal{I}_{0,l,u,j} = \emptyset$, unless $l = 0$ and $j = \overline{s}_u$, in which case $\mathcal{I}_{0,l,u,j}$ contains the single walk of length 0 with $i_0 = j$. Then $P_s e_u = e_j$ and

$$\begin{aligned}
\sum_l \sum_{w_* \in \mathcal{I}_{0,l,u,j}} M_{w_*} \Pr\left[X_0 = \iota(j)\right] &= \Pr\left[X_0 = \iota(j)\right] = e_u^t e_u \Pr\left[X_0 = \iota(j)\right] = e_u^t P_s^t e_j \Pr\left[X_0 = \iota(j)\right] \\
&= e_u^t p e_j \Pr\left[X_0 = \iota(j)\right] = p_{u,j} \Pr\left[X_0 = \iota(j)\right]
\end{aligned}$$

Otherwise, if $j = s_v$,

$$\begin{aligned}
\sum_l \sum_{w_* \in \mathcal{I}_{0,l,u,j}} M_{w_*} \Pr\left[X_0 = \iota(j)\right] &= \sum_l \left(\overline{N}\,\overline{M}^l\right)_{u,v} \Pr\left[X_0 = \iota(j)\right] = \sum_l e_u^t \overline{N}\,\overline{M}^l e_v \Pr\left[X_0 = \iota(j)\right] \\
&= e_u^t \overline{N} \left(\sum_l \overline{M}^l\right) e_v \Pr\left[X_0 = \iota(j)\right] \\
&= e_u^t \overline{N} \left(I - \overline{M}\right)^{-1} e_v \Pr\left[X_0 = \iota(j)\right] \\
&= e_u^t p P_s e_{n-k+v} \Pr\left[X_0 = \iota(j)\right] = e_u^t p e_j \Pr\left[X_0 = \iota(j)\right] \\
&= p_{u,j} \Pr\left[X_0 = \iota(j)\right]
\end{aligned}$$

Thus,

$$\Pr\left[\widetilde{X}_0 = \iota'(u)\right] = \sum_j \sum_l \sum_{w_* \in \mathcal{I}_{0,l,u,j}} M_{w_*} \Pr\left[X_0 = \iota(j)\right] = \sum_j p_{u,j} \Pr\left[X_0 = \iota(j)\right]$$

as desired.

More generally, if $k > 0$, partition $\mathcal{I}_{k,l,u,j}$ according to the the kth value in $\overline{s}$; specifically, let $\mathcal{I}_{k,l,u,j,v}$ be the set of all walks in $\mathcal{I}_{k,l,u,j}$ whose kth value in $\overline{s}$ is $\overline{s}_v$. Then if $t > 0$, then

$$\begin{aligned}
\Pr\left[\widetilde{X}_t = \iota'(u), \widetilde{X}_{t-1} = \iota'(v)\right] &= \sum_{j,l} \sum_{w_* \in \mathcal{I}_{t,l,u,j,v}} M_{w_*} \Pr\left[X_0 = \iota(j)\right] \\
&= \sum_{j,l} \sum_{m \le l} \sum_{w'_* \in \mathcal{I}_{1,m,u,\overline{s}_v}} \sum_{w''_* \in \mathcal{I}_{t-1,l-m,v,j}} M_{w'_*} M_{w''_*} \Pr\left[X_0 = \iota(j)\right] \\
&= \sum_{j,q,m} \sum_{w'_* \in \mathcal{I}_{1,m,u,\overline{s}_v}} \sum_{w''_* \in \mathcal{I}_{t-1,q,v,j}} M_{w'_*} M_{i''_*} \Pr\left[X_0 = \iota(j)\right] \\
&= \sum_m \sum_{w'_* \in \mathcal{I}_{1,m,u,\overline{s}_v}} M_{w'_*} \sum_{j,q} \sum_{w''_* \in \mathcal{I}_{t-1,q,v,j}} M_{w''_*} \Pr\left[X_0 = \iota(j)\right] \\
&= \sum_m \sum_{w'_* \in \mathcal{I}_{1,m,u,\overline{s}_v}} M_{w'_*} \Pr\left[\widetilde{X}_{t-1} = \iota'(v)\right]
\end{aligned}$$

Thus, if $\Pr\left[\widetilde{X}_{t-1} = \iota'(v)\right] > 0$,

$$\begin{aligned}
\Pr\left[\widetilde{X}_t = \iota'(u) \mid \widetilde{X}_{t-1} = \iota'(v)\right] &= \sum_m \sum_{w'_* \in \mathcal{I}_{1,m,u,\overline{s}_v}} M_{w'_*} \\
&= \sum_{w'_* \in \mathcal{I}_{1,0,u,\overline{s}_v}} M_{w'_*} + \sum_{m \ge 1} \sum_{w'_* \in \mathcal{I}_{1,m,u,\overline{s}_v}} M_{w'_*} \\
&= e_u^t \widetilde{M} e_v + \sum_{m \ge 1} e_u^t \overline{N} \, \overline{M}^{m-1} \widetilde{N} e_v \\
&= e_u^t \widetilde{M} e_v + \sum_{m \ge 1} e_u^t \overline{N} \left(I - \overline{M}\right)^{-1} \widetilde{N} e_v = e_u^t \widehat{M} e_v
\end{aligned}$$

Thus, $\left\{\widetilde{X}_t\right\}_{t=0}^{\infty}$ is a stationary Markov chain ι'-consistent with transition matrix, $\widehat{M}$. □

Theorem 5.3 allows us to easily show that the quotient construction on matrices of Section 3 is "natural", as well.

Theorem 5.4 *If $M \in \mathrm{Mat}_n(\mathbb{R})$ is Markov, $s = s_1 \cup s_2$ is open with respect to M, (M_1, p_1, i_1) is the quotient of M with respect to s_1, (M_2, p_2, i_2) is the quotient of M_1 with respect to $s'_2 = \overline{s}_1^{-1}(s_2)$, and $\left(\widehat{M}, p, i\right)$ is the quotient of M with respect to s, then $M_2 = \widehat{M}$, $p = p_2 p_1$, and $i = i_1 i_2$.*

Proof 5.4 First, notice that, by Theorem 3.5, s'_2 is open with respect to M_1, so that the statement of the Theorem makes sense. If ι is the identity on S_n, and v is some n-dimensional distribution, we may define a chain, $\{X_t\}_{t=0}^{\infty}$, which is ι-consistent with M such that $\Pr[X_0^M = j] = v_j, \forall j \in S_n$. By Theorem 5.3, M_1 is ι_1-consistent with $\overline{X}_* = \pi_{s_1}(X_*)$ and $\Pr\left[\overline{X}_0 = \iota_1(j)\right] = (p_1 v)_j$, where $\iota_1 = \overline{s}_1$. Likewise, M_2 is ι_2-consistent with $\widetilde{X}_* = \pi_{\iota_1(s'_2)}\left(\overline{X}_*\right) = \pi_{\overline{s}_1 \cap s_2}\left(\overline{X}_*\right)$ and $\Pr\left[\widetilde{X}_0 = \iota_2(j)\right] = (p_2 p_1 v)_j$, where $\iota_2 = \iota_1 \overline{s}'_2 = \overline{s}$. Thus, if we let $\widehat{X}_* = \pi_s(X_*)$, by Theorem 5.2,

$\widetilde{X}_* = \pi_{\overline{s}_1 \cap s_2}(\pi_{s_1}(X_*)) = \pi_s(X_*) = \widehat{X}_*$ almost everywhere. Since $\widetilde{X}_*$ is ι_2-consistent with M_2, and $\widehat{X}_*$ is ι_2-consistent with $\widehat{M}$, $\widehat{M} = M_2$.

For any $k \in S_n$, if $v = e_k$, then $(p_2 p_1 e_k)_j = \Pr\left[\widetilde{X}_0 = \overline{s}(j)\right] = \Pr\left[\widehat{X}_0 = \overline{s}(j)\right] = (pe_k)_j$, so that $p = p_2 p_1$. Finally, by Theorem 3.2, $i(v)$ is the unique extension of an eigenvector $v \in \ker\left(\widehat{M} - I\right)$ to an eigenvector in $\ker(M - I)$. Likewise, $i_1 i_2(v)$ is an extension of an eigenvector $v \in \ker(M_2 - I)$ first to an eigenvector in $\ker(M_1 - I)$, and then to an eigenvector in $\ker(M - I)$. Therefore, since $\widehat{M} = M_2$, by uniqueness, $i_1 i_2(v) = i(v)$. □

6 Calculating the Energy of a Generalized Simulated Annealing

In this section, we introduce the notion of a perturbed Markov matrix and show how this includes that of a generalized simulated annealing. We then present a high-level view of our algorithm to compute the energy level of each state in a GSA, followed by a discussion of two crucial subroutines. We conclude by discussing some subtle points regarding the implementation.

6.1 Perturbed Markov Matrices

We will say that f *converges exponentially* (cf. [15]) iff $f(\epsilon) = \epsilon^{r(f)} c(\epsilon)$ for some positive constant, $r(f)$, and some function, $c(\epsilon)$, which is continuous for $\epsilon \geq 0$ with $c(0) \neq 0$. This constant, $r(f)$, is called the *resistance* of f. By convention $\epsilon^\infty = 0$, so that $r(0) = \infty$. We define a *perturbed matrix*, M_ϵ, as a matrix whose entries are positive, exponentially convergent functions of ϵ. For any perturbed matrix M_ϵ, we may define the associated *resistance* matrix, $R(M_\epsilon)$, where $R(M_\epsilon)_{i,j} = r\left((M_\epsilon)_{ij}\right)$. When the perturbed matrix is clear from the context, we will simply write R for $R(M_\epsilon)$. For example, $M_\epsilon = \begin{pmatrix} 4\epsilon^2 - 3\epsilon^5 & -\sin(\epsilon) \\ 5\cos(\epsilon) & 0 \end{pmatrix}$ is a perturbed matrix with associated resistance matrix, $R = \begin{pmatrix} 2 & 1 \\ 0 & \infty \end{pmatrix}$, where we use the Taylor expansion of each entry of M_ϵ to identify its resistance as the exponent of its most significant term.

A *perturbed Markov matrix* is a perturbed matrix M_ϵ such that, for $\epsilon \geq 0$, M_ϵ is a Markov matrix and is regular for $\epsilon > 0$. Under the change of variable $\epsilon = e^{-\frac{1}{T}}$, an *irreducible*, perturbed Markov matrix corresponds precisely to an *admissible Markov kernel* of [13]. We will define the graph of the perturbed matrix to be $\mathcal{G}(R)$. Notice that for each $\epsilon > 0$, $\mathcal{G}(R) = \mathcal{G}(M_\epsilon)$, as unweighted graphs. That is, the definition of a perturbed matrix fixes the geometry of its graph independently of ϵ. When M_ϵ is a perturbed Markov matrix, $R_{i,j}$ is also called the *resistance of edge* (i, j) in $\mathcal{G}(M_\epsilon)$. In the GSA literature, this is referred to as the "communication cost" of the transition.

For each $\epsilon > 0$, let $v_\epsilon = \mathrm{stab}(M_\epsilon)$ denote the unique stable distribution of M_ϵ. We now state a generalization of a result proven by Freidlin and Wentzell [5] for perturbed Markov matrices. For $1 \leq i \leq n$ define

$$T_i = \{\sigma : S_n \to S_n \mid \sigma(i) = i \text{ and } \sigma(j) \neq j, \forall j \neq i\}$$

Each such mapping may be viewed as the successor relation of 1-regular[1] graph on n vertices with exactly one self-loop at i. By dropping the self-loop, such a graph is just a directed spanning tree rooted at i. For any $\sigma \in T_i$, we then define the *resistance of* σ *in* M_ϵ as $r(\sigma, M_\epsilon) = \sum_{j \neq i} R_{\sigma(j),j}$, i.e., the weight of the spanning subtree of $\mathcal{G}(M_\epsilon)$ corresponding to σ.

[1] That is, all vertices have out-degree 1.

While [5] proves Theorem 6.1 for *irreducible* perturbed Markov matrices, the following more general result holds, as well.

Theorem 6.1 *If M_ϵ is a perturbed Markov matrix, then the entries of v_ϵ converge exponentially. If $r_i = \min_{\sigma \in T_i} r(\sigma, M_\epsilon)$, we may calculate the resistance of eahc coordinate of v_ϵ as $r((v_\epsilon)_i) = r_i - \min_j r_j$.*

The *virtual energy*[2] of each index of M_ϵ is the resistance of the corresponding entry of v_ϵ, which by Theorem 6.1 is well-defined. The indices with virtual energy 0 are called the *ground states* of M_ϵ. By restricting attention to a sequence $\epsilon_t = e^{-\frac{1}{T_t}}$ for some sequence $T_t \to 0$, for any given initial distribution, a perturbed Markov matrix defines an *inhomogenous* Markov chain, $\{X_t\}_{t=0}^{\infty}$, of a generalized simulated annealing (GSA) (cf. [13] and [4]) with transition matrix, M_{ϵ_t}, at time t.

If $T_t \to 0$ slowly enough, Desai, et al. [4] describe the resulting process as "quasi-statically cooled." This is intended to connote that, in some sense, the limiting distribution of this process equals the limit of the stable distributions, v_{ϵ_t}, as $t \to \infty$. Specifically, assume that the resistance matrix corresponds to the energy differences of a potential function, $U(i)$, so that $R_{i,j} = (U(i) - U(j))^+$. Such a potential is defined up to an additive constant and there is a unique choice with minimum value 0. Under certain conditions, the virtual energies defined above will agree with such a potential. To be precise, Trouvé [13] shows that this holds iff Hajek's "weak reversability" condition [7] is satisfied. For example, it suffices for the unweighted resistance graph to be undirected; i.e., $R_{i,j} < \infty$ iff $R_{j,i} < \infty$.

We may generalize the notion of equivalence from Section 2. We will say that two matrices, M_ϵ and M'_ϵ, are *D_ϵ-equivalent* and write $M_\epsilon \sim_{D_\epsilon} M'_\epsilon$ iff

- D_ϵ has exponentially convergent entries,
- $f(\epsilon)\,\mathrm{stab}(M_\epsilon) = D_\epsilon\,\mathrm{stab}(M'_\epsilon)$ with $f(\epsilon) \neq 0$ for $\epsilon > 0$, and
- $r(f(\epsilon)) = 0$.

Because the geometry of $\mathcal{G}(M_\epsilon)$ is independent of ϵ, if a set of states is open for any ϵ, it is true for all $\epsilon > 0$. Thus, we may also apply the quotient construction for each $\epsilon > 0$ to a perturbed Markov process. Moreover, Theorem 3.2 generalizes so that $M_\epsilon \sim_{i_\epsilon} \widehat{M}_\epsilon$. This means that we may recover the virtual energies of the original matrix from any quotient.

6.2 Computing Virtual Energies

By Theorem 6.1, we see that virtual energy levels of a GSA depends only on the collection of minimal directed subtrees of its resistance graph. One might suspect that, as in Kruskal's algorithm for undirected graphs, one may proceed inductively by starting with the edges of minimal weight. Our algorithm, Algorithm 1, may be viewed in this light. We proceed by analyzing the unperturbed matrix, given by M_0, which corresponds to the subgraph of 0-resistance edges. By applying our quotient construction, we may then "factor out" these edges to focus attention on the contribution of higher resistance edges. While we never actually calculate such subtrees, the normalized quotient construction does this is implicitly, since the columns of the normalized inclusion operator are stable distributions, which by Theorem 6.1 encode a summary of certain minimal resistance subtrees.

Because we are working with *directed* graphs, we must be a bit careful. Fortunately, if we operate on one communicating class at a time, this approach works. Specifically, one can show that the

[2] Desai, et. al. [4] call this the "stationary order".

intersection of a minimal spanning tree with a communicating class, $\mathcal{C}$, of the minimum-edge-weight subgraph must be a spanning tree of the subgraph on $\mathcal{C}$.

Using naturality of the quotient construction, we may in fact analyze all communicating classes simultaneously. More precisely, we examine all non-trivial (i.e., containing more than one element) communicating classes of M_0. While this intuition suggests Algorithm 1, its correctness is based on the two theorems given below.

Algorithm 1 To Compute the Virtual Energies of a GSA.

```
1: function v_0 = virtualEnergy (M_ε)
2:     if (dim M_ε == 1)
3:         return(1);
4:     /*  Calculate the communicating classes of M_0,
5:         marking each as closed, transient, and/or trivial. */
6:     C = commClasses(M_0);
7:     D = I;
8:     if (C.nonTrivial == 0) {
9:         (M_ε, D) = nonUniformScale (M_ε, C);
10:        return(D (virtualEnergy (M_ε)));
11:    }
12:    (M̂*_ε, i*) = quotient (M_ε, C);
13:    return(Di* (virtualEnergy (M̂*_ε)));
```

For this approach to make progress, the unperturbed component, M_0, must possess at least one non-trivial communicating class, which is not always the case for an arbitrary perturbed Markov matrix, M_ϵ. For example, we could have $M_\epsilon = \begin{pmatrix} 1-2\epsilon-3\epsilon^2 & 2\epsilon^2 & 1/2 \\ 3\epsilon^2 & 1-\epsilon-2\epsilon^2 & \epsilon \\ 2\epsilon & \epsilon & 1/2-\epsilon \end{pmatrix}$. However, in this case we may transform M_ϵ to a closely related perturbed Markov matrix. By the following Theorem, we may always guarantee that M_ϵ always possesses a non-trivial communicating class, as long as we keep track of the corresponding shift in virtual energies.

Theorem 6.2 *Given any $n \times n$ irreducible, perturbed Markov matrix, M_ϵ, if M_0 possesses more than one closed class and $n > 1$, there is an i_ϵ-equivalent perturbed Markov matrix, M'_ϵ, where $i_\epsilon \in \mathrm{Mat}_n(\mathcal{C}^+)$ is a diagonal matrix, M'_0 possesses a non-trivial communicating class, and M'_ϵ is irreducible.*

This Theorem forms the basis for Algorithm 2.

Applying Theorem 6.2 to the example given above would yield

$$M'_\epsilon = \begin{pmatrix} 1/2-3\epsilon/4 & \epsilon/2 & 1/2 \\ 3\epsilon/4 & 3/4-\epsilon/2 & \epsilon \\ 1/2 & 1/4 & 1/2-\epsilon \end{pmatrix} \qquad \text{with} \qquad i_\epsilon = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4\epsilon \end{pmatrix}$$

Notice that $M'_\epsilon = (M_\epsilon - I)\, D_\epsilon + I$, where $D_\epsilon = (4\epsilon)^{-1} i_\epsilon$. In effect, we have divided through by the greatest common factor (i.e., ϵ) of the off-diagonal entries in the non-transient columns (i.e., columns 1 and 2), with an additional factor (in this case, 4) to guarantee that the result is Markov. Notice that every non-transient index is closed and the exponent of this greatest common factor is just the minimum resistance. This is why the key function in Algorithm 2 is called **minClosedResistance**.

Algorithm 2 To Perform a Non-Uniform Scaling Transformation of a GSA.

function $(M_\epsilon, i) = \text{nonUniformScale}\,(M_\epsilon, C)$
  $(D_\epsilon, i_\epsilon) = \text{minClosedResistance}\,(M_\epsilon, C)$;
  $M'_\epsilon = (M_\epsilon - I)\, D_\epsilon + I$;
  return$(M'_\epsilon, i_\epsilon)$;

One should be concerned, however, about the computational complexity of applying the quotient construction directly to a perturbed matrix, since this would involve inverting a matrix whose entries are functions. Fortunately, by exploiting the description of our quotient construction in terms of projection and inclusion operators, we may avoid this difficulty. If we choose s_0 to be the complement of a set of representatives of the communicating classes of the *unperturbed* matrix, M_0, we may apply the corresponding projection and inclusion operators of the associated normalized quotient of M_0 to M_ϵ. Using the path interpretation of the quotient from Theorem 5.3, we may show that this yields a result equivalent to the quotient construction applied directly to M_ϵ. Specifically,

Theorem 6.3 *If M_ϵ is a perturbed Markov matrix and s_0 contains all indices but one representative of each communicating class in M_0, let $\left(\widehat{M}_0^*, p_0, i_0^*\right)$ be the normalized quotient with respect to s_0 of M_0. Then $M_\epsilon \sim_{i_0^*} p_0\,(M_\epsilon - I)\, i_0^* + I$.*

This Theorem forms the basis for Algorithm 3.

Algorithm 3 To Compute the Quotient of a GSA.

function $(M_\epsilon, i^*) = \text{quotient}\,(M_\epsilon, C)$
  /* Let s_0 be the union of all but one representative from each communicating
    class, where the representative is taken to be the first element of each class */
  $s_0 = C[0].\text{rest}()$;
  for $(j = 1;\ j < C.\text{nonTrivial};\ j{+}{+})$
    $s_0 = \text{append}(s_0, C[j].\text{rest}())$;
  $\left(\widetilde{M}, \overline{N}, \overline{M}, \widetilde{N}, P\right) = \text{partition}\,(M_0, s_0)$;
  $i^* = \text{normalize}\left(P \begin{pmatrix} I \\ \left(I - \overline{M}\right)^{-1} \widetilde{N} \end{pmatrix}\right)$;
  $p = \begin{pmatrix} I & \overline{N}\left(I - \overline{M}\right)^{-1} \end{pmatrix} P.\text{transpose}()$;
  $M_\epsilon = p\,(M_\epsilon - I)\, i^* + I$;
  return(M_ϵ, i^*);

6.3 Representing Perturbed Matrices

Even avoiding the problem of matrix inversion, storing and manipulating matrices whose entries are functions is computationally intensive. However, in principle, Theorem 6.1 suggests that we should only need to work with the associated resistance matrix and not the full transition matrix of a GSA. Thus, we will represent such transition matrices simply by their associated resistance matrices. Specifically, we will introduce an equivalence relation on perturbed matrices and observe that all necessary operations preserve this relation. **Note**: We refered to this equivalence relation implicitly in the comments preceding Theorem 6.3.

We will say that two perturbed matrices, M_ϵ and $\overline{M}_\epsilon$ are *weakly equivalent* and write $M_\epsilon \backsim \overline{M}_\epsilon$ iff $R(M_\epsilon) = R\left(\overline{M}_\epsilon\right)$. By Theorem 6.1, the virtual energy of a perturbed Markov matrix only depends on its weak equivalence class. Moreover, it is easy to show that matrix addition and multiplication factor to operations on such equivalence classes.

Lemma 6.4 *Assume that M_ϵ, $\overline{M}_\epsilon$, and $\widetilde{M}_\epsilon$ are pertubed matrices, where the dimensions of the first two are $n \times m$ and the last is $m \times p$.*

1. $R\left(M_\epsilon + \overline{M}_\epsilon\right)_{i,j} = \min\left\{R(M_\epsilon)_{i,j}, R\left(\overline{M}_\epsilon\right)_{i,j}\right\}$.
2. $R\left(M_\epsilon \widetilde{M}_\epsilon\right)_{i,j} = \min_k\left\{R(M_\epsilon)_{i,k} + R\left(\widetilde{M}_\epsilon\right)_{k,j}\right\}$.

In particular, addition and multiplication of perturbed matrices preserve weak equivalence.

With some care, we can show that the quotient construction of Section 5 generalizes so that we may apply it to weak-equivalence classes of perturbed Markov matrices. Specifically, if M_ϵ is a perturbed Markov matrix:

- if for each ϵ, $\widehat{M}_\epsilon$ is the quotient of M_ϵ with respect to s, where s is an open subset of indices, $\widehat{M}_\epsilon$ is a perturbed Markov matrix;
- if $M_\epsilon \backsim M'_\epsilon$, then $\widehat{M}_\epsilon \backsim \widehat{M}'_\epsilon$.

Thus, while it is conceptually useful to think of the algorithm in terms of operations on perturbed matrices and vectors, for efficiency they should be represented internally simply by the corresponding resistance matrices and, by Lemma 6.4, computations may be performed directly on such resistances.

This simplifies a number of the calculations. For example, in the scaling operation of Algorithm 2, we do not need to bother with the constant factor, since this does not affect the resistance. Likewise, in Algorithm 3, we do not need to actually invert $I - \overline{M}$. By construction, all entries of $I - \overline{M}$ have either resistance 0 or ∞ (corresponding to non-zero or zero entries, respectively). Thus, $\left(I - \overline{M}\right)^{-1} \backsim \sum_{t=0}^{n-1} \overline{M}^t$; that is, we only need to determine which pairs of vertices are path-connected.

7 Summary and Future Work

In this paper, we have presented a novel quotient construction on Markov chains. We have shown that it generalizes the quotient construction on Markov matrices of [14]. We have pointed out how this construction may be applied to perturbed Markov matrices both to compute the stochastically stable distribution of a PMP as well as the virtual energy levels in a GSA. We also indicated how many basic facts about Markov matrices may be deduced using this construction, such as the necessary and sufficient conditions for a Markov matrix to have a unique stable distribution. We are currently exploring other applications of our quotient construction, in areas ranging from Decision Theory to Information Retrieval.

8 Appendix

For completeness, we include the details of the algorithm from [14]. As with Algorithm 1, all computations are performed on equivalence classes of perturbed Markov matrices. However, because the stochastically stable distribution is a more detailed invariant than the virtual energy levels, we must utilize a weaker notion of equivalence which takes into account the leading coefficients of the transition matrices, as well. This requires a bit more care when computing quotients; in particular, we must work one communicating class at a time. Conceptually, however, the algorithms are quite similar.

Algorithm 4 To Compute the Stochastically Stable Distribution of a PMP.

function $v_0 = \text{stochasticallyStableDistribution}\ (M_\epsilon)$
  $C = \text{commClasses}\,(M_0)$;
  $D = I$;
  if $((C.\text{nonTrivialClosed} == 0)\ \&\&\ (C.\text{numClosed} > 1))$
    $(M_\epsilon, D) = \text{nonUniformScale}\,(M_\epsilon, C)$;
  if $(C.\text{numClosed} == 1)$
    return$(\text{normalize}\,(D\,\text{stab}\,(M_0)))$;
  $i = I$;
  while $(C.\text{nonTrivialClosed} > 0)$
    $(M_\epsilon, i, C) = \text{quotient}\,(M_\epsilon, i, C)$;
  return$(\text{normalize}\,(Di\,(\text{stochasticallyStableDistribution}\,(M_\epsilon))))$;

Algorithm 5 To Compute the Quotient of a PMP.

#define uniformScale(M) $(\text{hasZeroOnDiagonalP}\,(M)\ :\ (I+M)/2\ ?\ M)$
function $(M_\epsilon, i, C) = \text{quotient}\ (M_\epsilon, i, C)$
  $M_\epsilon = \text{uniformScale}\,(M_\epsilon)$;
  /* Select the largest closed class, $c \in C$;
    return $s = c.\text{rest}()$ and C with c replaced by $\{c.\text{first}()\}$ */
  $(C, s) = \text{pickLargestClosed}\,(C)$;
  $\left(\widetilde{M}_\epsilon, \overline{N}_\epsilon, \overline{M}_\epsilon, \widetilde{N}_\epsilon, P\right) = \text{partition}\,(M_\epsilon, s)$;
  $M_\epsilon = \widetilde{M}_\epsilon + \overline{N}_\epsilon \left(I - \overline{M}_0\right)^{-1} \widetilde{N}_\epsilon$;
  $i = iP \begin{pmatrix} I \\ \left(I - \overline{M}_0\right)^{-1} \widetilde{N}_0 \end{pmatrix}$;
  return(M_ϵ, i, C);

References

[1] V. Anantharam and P. Tsoucas. A proof of the markov chain tree theorem. Technical report, University of Maryland, 1988.

[2] Robert B. Ash. *Real Analysis and Probability.* Academic Press, New York, 1972.

[3] Daniel P. Connors and P. R. Kumar. Simulated annealing type markov chains and their order balance equations. *SIAM Journal of Control and Optimization*, 27(6):1440–1461, 1989.

[4] Madhav Desai, Sunil Kumar, and P. R. Kumar. Quasi-statically cooled markov chains. *Probability in the Engineering and Informational Sciences*, 8:1–19, 1994.

[5] Mark Friedlin and Alexander Wentzell. *Random Perturbations of Dynamical Systems*. Springer–Verlag, Berlin, 1984.

[6] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[7] B. Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13(2):311–329, 1988.

[8] Roger Horn and Charles Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[9] Chii-Ruey Hwang and Shuenn Jyi Sheu. Singular perturbed Markov chains and exact behaviors of simulated annealing processes. *J. Theoret. Probab.*, 5(2):223–249, 1992.

[10] S. Kirkpatrick and C. D. Gelatt; M. P. Vecchi. Optimization by simulated annealing. *Science*, 220 (4598):671–680, 1983.

[11] Frank Thomson Leighton and Ronald L. Rivest. Estimating a probability using finite memory. *IEEE Transactions on Information Theory*, 32(6):733–742, 1986.

[12] A. N. Shiryaev. *Probability*. Springer-Verlag, New York, 1996.

[13] Alain Trouvé. Cycle decompositions and simulated annealing. *SIAM J. Control Optim.*, 34:966–986, 1996.

[14] John R. Wicks and Amy Greenwald. An algorith for computing stochastically stable distributions with applications to multiagent learning in replicated games. In *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-First Conference*, pages 623–632. AUAI Press, 2005.

[15] H. Peyton Young. The evolution of conventions. *Econometrica*, 61:57–84, 1993.

[16] H. Peyton Young. *Individual strategy and social structure : an evolutionary theory of institutions*. Princeton University Press, Princeton, 1998.