Decomposition Methods and Sampling Circuits in the Cartesian Lattice

Dana Randall*

College of Computing and School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0280
randall@math.gatech.edu

Abstract. Decomposition theorems are useful tools for bounding the convergence rates of Markov chains. The theorems relate the mixing rate of a Markov chain to smaller, derivative Markov chains, defined by a partition of the state space, and can be useful when standard, direct methods fail. Not only does this simplify the chain being analyzed, but it allows a hybrid approach whereby different techniques for bounding convergence rates can be used on different pieces. We demonstrate this approach by giving bounds on the mixing time of a chain on circuits of length $2n$ in $\mathbb{Z}^d$.

1 Introduction

Suppose that you want to sample from a large set of combinatorial objects. A popular method for doing this is to define a Markov chain whose state space Ω consists of the elements of the set, and use it to perform a random walk. We first define a graph H connecting pairs of states that are close under some metric. This underlying graph on the state space representing allowable transitions is known as the Markov kernel.

To define the transition probabilities of the Markov chain, we need to consider the desired stationary distribution π on Ω. A method known as the *Metropolis algorithm* assigns probabilities to the edges of H so that the resulting Markov chain will converge to this distribution. In particular, if Δ is the maximum degree of any vertex in H, and (x, y) is any edge,

$$P(x,y) = \frac{1}{2\Delta} \min\left(1, \frac{\pi(y)}{\pi(x)}\right).$$

We then assign self loops all remaining probability at each vertex, so $P(x, x) \geq 1/2$ for all $x \in \Omega$. If H is connected, π will be the unique stationary distribution of this Markov chain. We can see this by verifying that *detailed balance* is satsified on every edge (x, y), i.e., $\pi(x)P(x, y) = \pi(y)P(y, x)$.

As a result, if we start at any vertex in Ω and perform a random walk according to the transition probabilities defined by P, and we walk long enough, we will converge to the desired distribution. For this to be useful, we need that we are converging rapidly to π, so that after a small, polynomial number of steps, our samples will be chosen from a

* Supported in part by NSF Grant No. CCR-9703206.

distribution which is provably arbitrarily close to stationarity. A Markov chain with this property is *rapidly mixing*.

Consider, for example, the set of independent sets $\mathcal{I}$ of some graph G. Taking the Hamming metric, we can define H by connecting any two independent sets that differ by the addition or deletion of a single vertex. A popular stationary distribution is the *Gibbs distribution* which assigns weight $\pi(I) = \gamma^{|I|}/Z_\gamma$ to I, where $\gamma > 0$ is an input parameter of the system, $|I|$ is the size of the independent set I, and $Z_\gamma = \sum_{J \in \mathcal{I}} \gamma^{|J|}$ is the normalizing constant known as the *partition function*. In the Metropolis chain, we have $P(I, I') = \frac{1}{2n} \min(1, \gamma)$ if I' is formed by adding a vertex to I, and $P(I, I') = \frac{1}{2n} \min(1, \gamma^{-1})$ if I' is formed by deleting a vertex from I.

Recently there has been great progress in the design and analysis of Markov chains which are provably efficient. One of the most popular proof techniques is *coupling*. Informally, coupling says that if two copies of the Markov chain can be simultaneously simulated so that they end up in the same state very quickly, regardless of the starting states, then the chain is rapidly mixing. In many instances this is not hard to establish, which gives a very easy proof of fast convergence.

Despite the appeal of these simple coupling arguments, a major drawback is that many Markov chains which appear to be rapidly mixing do not seem to admit coupling proofs. In fact, the complexity of typical Markov chains often makes it difficult to use any of the standard techniques, which include bounding the conductance, the log Sobolev constant or the spectral gap, all closely related to the mixing rate.

The decomposition method offers a way to systematically simplify the Markov chain by breaking it into more manageable pieces. The idea is that it should be easier to apply some of these techniques to the simplified Markov chains and then infer a bound on the original Markov chain. In this survey we will concentrate on the *state decomposition theorem* which utilizes some partition of the state space. It says that if the Markov chain is rapidly mixing when restricted to each piece of the partition, and if there is sufficient flow between the pieces (defined by a *"projection"* of the chain), then the original Markov chain must be rapidly mixing as well. The allows us to take a top-down approach to mixing rate analysis, whereby we need only consider the mixing rate of the restrictions and the projection. In many cases it is easier to define good couplings on these simpler Markov chains, or to use one of the other known methods of analysis. We note, however, that using indirect methods such as the decomposition or comparison (defined later) invariably adds orders of magnitude the bounds on the running time of the algorithm. Hence it is wise to use these methods judiciously unless the goal is simply to establish a polynomial bound on the mixing rate.

2 Mixing machinery

In what follows, we assume that $\mathcal{M}$ is an ergodic (i.e. irreducible and aperiodic), reversible Markov chain with finite state space Ω, transition probability matrix P, and stationary distribution π.

The time a Markov chain takes to converge to its stationary distribution, i.e., the mixing time of the chain, is measured in terms of the distance between the distribution at time t and the stationary distribution. Letting $P^t(x, y)$ denote the t-step probability of going from x to y, the *total variation distance* at time t is

$$\|P^t, \pi\|_{tv} = \max_{x \in \Omega} \frac{1}{2} \sum_{y \in \Omega} |P^t(x,y) - \pi(y)|.$$

For $\varepsilon > 0$, the *mixing time* $\tau(\varepsilon)$ is

$$\tau(\varepsilon) = \min\{t : \|P^{t'}, \pi\|_{tv} \leq \varepsilon, \forall t' \geq t\}.$$

We say a Markov chain is *rapidly mixing* if the mixing time is bounded above by a polynomial in n and $\log \frac{1}{\varepsilon}$, where n is the size of each configuration in the state space.

It is well known that the mixing rate is related to the *spectral gap* of the transition matrix. For the transition matrix P, we let $Gap(P) = \lambda_0 - |\lambda_1|$ denote its spectral gap, where $\lambda_0, \lambda_1, \ldots, \lambda_{|\Omega|-1}$ are the eigenvalues of P and $1 = \lambda_0 > |\lambda_1| \geq |\lambda_i|$ for all $i \geq 2$. The following result the spectral gap and mixing times of a chain (see, e.g., [18]).

Theorem 1. *Let $\pi_* = \min_{x \in \Omega} \pi(x)$. For all $\varepsilon > 0$ we have*

(a) $\tau(\varepsilon) \leq \frac{1}{Gap(P)} \log(\frac{1}{\pi_* \varepsilon})$
(b) $\tau(\varepsilon) \geq \frac{|\lambda_1|}{2Gap(P)} \log(\frac{1}{2\varepsilon})$.

Hence, if $1/Gap(P)$ is bounded above by a polynomial, we are guaranteed fast (polynomial time) convergence. For most of what follows we will rely on the spectral gap bound on mixing. Theorem 1 is useful for deriving a bound on the spectral gap from a coupling proof, which provides bounds on the mixing rate.

We now review of some of the main techniques used to bound the mixing rate of a chain, including the decomposition theorem.

2.1 Path Coupling

One of the most popular methods for bounding mixing times has been the coupling method. A *coupling* is a Markov chain on $\Omega \times \Omega$ with the following properties. Instead of updating the pair of configurations independently, the coupling updates them so that i) the two processes will tend to coalesce, or "move together" under some measure of distance, yet ii) each process, viewed in isolation, is performing transitions exactly according to the original Markov chain. A valid coupling ensures that once the pair of configurations coalesce, they agree from that time forward. The mixing time can be bounded by the expected time for configurations to coalesce under any valid coupling.

The method of path coupling simplifies our goal by letting us bound the mixing rate of a Markov chain by considering only a small subset of $\Omega \times \Omega$ [3, 6].

Theorem 2. **(Dyer and Greenhill [6])** *Let Φ be an integer valued metric defined on $\Omega \times \Omega$ which takes values in $\{0, \ldots, B\}$. Let U be a subset of $\Omega \times \Omega$ such that for all $(x, y) \in \Omega \times \Omega$ there exists a path $x = z_0, z_1, \ldots, z_r = y$ between x and y such that $(z_i, z_{i+1}) \in U$ for $0 \leq i < r$ and*

$$\sum_{i=0}^{r-1} \Phi(z_i, z_{i+1}) = \Phi(x, y).$$

Define a coupling $(x, y) \to (x', y')$ of the Markov chain $\mathcal{M}$ on all pairs $(x, y) \in U$. Suppose that there exists $\alpha < 1$ such that $\mathbf{E}[\Phi(x', y')] \leq \alpha\,\Phi(x, y)$ for all $(x_t, y_t) \in U$, Then the mixing time of $\mathcal{M}$ satisfies

$$\tau(\epsilon) \leq \frac{\log(B\epsilon^{-1})}{1-\alpha}.$$

Useful bounds can also be derived in the case that $\alpha = 1$ in the theorem (see [6]).

2.2 The disjoint decomposition method

Madras and Randall [12] introduced two decomposition theorems which relate the mixing rate of a Markov chain to the mixing rates of related Markov chains. The *state decomposition theorem* allows the state space to be decomposed into overlapping subsets; the mixing rate of the original chain can be bounded by the mixing rates of the *restricted Markov chains*, which are forced to stay within the pieces, and the ergodic flow between these sets. The *density decomposition theorem* is of a similar flavor, but relates a Markov chain to a family of other Markov chains with the same Markov kernel, where the transition probabilities of the original chain can be described as a weighted average of the transition probabilities of the chains in the family.

We will concentrate on the state decomposition theorem, and will present a newer version of the theorem due to Martin and Randall [15] which allows the decomposition of the state to be a partition, rather than requiring that the pieces overlap.

Suppose that the state space is partitioned into m disjoint pieces $\Omega_1, \ldots, \Omega_m$. For each $i = 1, \ldots, m$, define $P_i = P\{\Omega_i\}$ as the restriction of P to Ω_i which rejects moves that leave Ω_i. In particular, the restriction to Ω_i is a Markov chain, $\mathcal{M}_i$, where the transition matrix P_i is defined as follows: If $x \neq y$ and $x, y \in \Omega_i$ then $P_i(x, y) = P(x, y)$; if $x \in \Omega_i$ then $P_i(x, x) = 1 - \sum_{y \in \Omega_i, y \neq x} P_i(x, y)$. Let π_i be the normalized restriction of π to Ω_i, i.e., $\pi_i(A) = \frac{\pi(A \cap \Omega_i)}{\pi(\Omega_i)}$. Notice that if Ω_i is connected then π_i is the stationary distribution of P_i.

Next, define $\overline{P}$ to be the following aggregated transition matrix on the state space $[m]$:

$$\overline{P}(i, j) = \frac{1}{\pi(\Omega_i)} \sum_{\substack{x \in \Omega_i, \\ y \in \Omega_j}} \pi(x) P(x, y).$$

Theorem 3. (Martin and Randall [15]) *Let P_i and $\overline{P}$ be as above. Then the spectral gaps satisfy*

$$Gap(P) \geq \frac{1}{2} Gap(\overline{P}) \min_{i \in [m]} Gap(P_i).$$

A useful corollary allows us to replace $\overline{P}$ in the theorem with the Metropolis chain defined on the same Markov kernel, provided some simple conditions are satisfied. Since the transitions of the Metropolis chain are fully defined by the stationary distribution $\overline{\pi}$, this is often easier to analyze than the true projection.

Define P_M on the set $[m]$, with Metropolis transitions $P_M(i, j) = \min\{1, \frac{\pi(\Omega_j)}{\pi(\Omega_i)}\}$. Let $\partial_i(\Omega_j) = \{y \in \Omega_j : \exists\, x \in \Omega_i \text{ with } P(x, y) > 0\}$.

Corollary 1. **[15]** *With P_M as above, suppose there exists $\beta > 0$ and $\gamma > 0$ such that*

(a) $P(x,y) \geq \beta$ *whenever* $P(x,y) > 0$*;*
(b) $\pi(\partial_i(\Omega_j)) \geq \gamma\pi(\Omega_j)$ *whenever* $\overline{P}(i,j) > 0$.

Then

$$Gap(P) \geq \frac{1}{2}\beta\gamma\, Gap(P_M) \min_{i=1,\ldots,m} Gap(P_i).$$

2.3 The comparison method

When applying the decomposition theorem, we reduce the analysis of a Markov chain to bounding the convergence times of smaller related chains. In many cases it will be much simpler to analyze variants of these auxiliary Markov chains instead of the true restrictions and projections. The comparison method tells us ways in which we can slightly modify one of these Markov chains without qualitatively changing the mixing time. For instance, it allows us to add additional transition edges or to amplify some of the transition probabilities, which can be useful tricks for simplifying the analysis of a chain.

Let $\tilde{P}$ and P be two reversible Markov chains on the same state space Ω with the same stationary distribution π. The comparison method allows us to relate the mixing times of these two chains (see [4] and [17]). In what follows, suppose that $Gap(\tilde{P})$, the spectral gap of $\tilde{P}$, is known (or suitably bounded) and we desire a bound on $Gap(P)$, the spectral gap of P, which is unknown.

Following [4], we let $E(P) = \{(x,y) : P(x,y) > 0\}$ and $E(\tilde{P}) = \{(x,y) : \tilde{P}(x,y) > 0\}$ denote the sets of edges of the two chains, viewed as directed graphs. For each $(x,y) \in E(\tilde{P})$, define a *path* γ_{xy} using a sequence of states $x = x_0, x_1, \ldots, x_k = y$ with $(x_i, x_{i+1}) \in E(P)$, and let $|\gamma_{xy}|$ denote the length of the path. Let $\Gamma(z,w) = \{(x,y) \in E(\tilde{P}) : (z,w) \in \gamma_{xy}\}$ be the set of paths that use the transition (z,w) of P. Finally, define

$$A = \max_{(z,w)\in E(P)} \left\{ \frac{1}{\pi(z)P(z,w)} \sum_{\Gamma(z,w)} |\gamma_{xy}|\pi(x)\tilde{P}(x,y) \right\}.$$

Theorem 4. **(Diaconis and Saloff-Coste [4])** *With the above notation, the spectral gaps satisfy* $Gap(P) \geq \frac{1}{A} Gap(\tilde{P})$

It is worthwhile to note that there are several other comparison theorems which turn out to be useful, especially when applying decomposition techniques. The following lemma helps us reason about a Markov chain by slightly modifying the transition probabilities (see, e.g., [10]). We use this trick in our main application, sampling circuits.

Lemma 1. *Suppose P and P' are Markov chains on the same state space, each reversible with respect to the distribution π. Suppose there are constants c_1 and c_2 such that $c_1 P(x,y) \leq P'(x,y) \leq c_2 P(x,y)$ for all $x \neq y$. Then $c_1 Gap(P) \leq Gap(P') \leq c_2 Gap(P)$.*

3 Sampling circuits in the Cartesian lattice

A *circuit* in $\mathbb{Z}^d$ is a walk along lattice edges which starts and ends at the origin. Our goal is to sample from $\mathcal{C}$, the set of circuits of length $2n$. It is useful to represent each walk as a string of $2n$ letters using $\{a_1, ..., a_d\}$ and their inverses $\{a_1^{-1}, ..., a_d^{-1}\}$, where a_i represents a positive step in the ith direction, and a_i^{-1} represents a negative step. Since these are closed circuits, the number of times a_i appears must equal the number of times a_i^{-1} appears, for all i. We will show how to uniformly sample from the set of all circuits of length $2n$ using an efficient Markov chain. The primary tool will be finding an appropriate decomposition of the state space. We outline the proof here and refer the reader to [16] for complete details.

Using a similar strategy, Martin and Randall showed how to use a Marko chain to sample circuits in regular d-ary trees, i.e., paths of length $2n$ which trace edges of the tree starting and ending at the origin [15]. This problem generalizes to sampling Dyke paths according to a distribution which favors walks that hit the x-axis a large number of times, known in the statistical physics community as "adsorbing staircase walks." Here too the decomposition method was the basis of the analysis. We note that there are other simple algorithms for sampling circuits on trees which do not require Markov chains. In contrast, to our knowledge, the Markov chain based algorithm discussed in this paper is the first efficient method for sampling circuits on $\mathbb{Z}^d$.

3.1 The Markov chain on circuits

The Markov chain on $\mathcal{C}$ is based on two types of moves: *transpositions* of neighboring letters in the word (which keep the numbers of each letter fixed) and *rotations*, which replace an adjacent (a_i, a_i^{-1}) with (a_j, a_j^{-1}), for some pair of letters a_i and a_j.

We now define the transition probabilities $\mathcal{P}$ of $\mathcal{M}$, where we say $x \in_u X$ to mean that we choose x from set X uniformly. Starting at σ, do the following. With probability 1/2, pick $i \in_u [n-1]$ and transpose σ_i and σ_{i+1}. With probability 1/2, pick $i \in_u [n-1]$ and $k \in_u [d]$ and if σ_i and σ_{i+1} are inverses (where σ_i is a step in the positive direction), then replace them with (a_k, a_k^{-1}). Otherwise keep σ unchanged.

The chain is aperiodic, ergodic and reversible, and the transitions are symmetric, so the stationary distribution of this Markov chain is the uniform distribution on $\mathcal{C}$.

3.2 Bounding the mixing rate of the circuits Markov chain

We bound the mixing rate of $\mathcal{M}$ by appealing to the decomposition theorem. Let $\sigma \in \mathcal{C}$ and let x_i equal the number of occurrences of a_i, and hence a_i^{-1} in σ, for all i. Define the *trace* $\mathrm{Tr}(\sigma)$ to be the vector $X = (x_1, ..., x_d)$. This defines a partition of the state space into

$$\mathcal{C} = \cup\, \mathcal{C}_X,$$

where the union is over all partitions of n into d pieces and $\mathcal{C}_X$ is the set of words $\sigma \in \mathcal{C}$ such that $\mathrm{Tr}(\sigma) = \mathrm{X} = (\mathrm{x}_1, ..., \mathrm{x}_\mathrm{d})$. The cardinality of the set $\mathcal{C}_X$ is $\binom{2n}{x_1, x_1, ..., x_d, x_d}$, the number of distinct words (or permutations) of length $2n$ using the letters with these prescribed multiplicities. The number of sets in the partition of the state space is exactly the number of partitions of n into d pieces, $D = \binom{n+d-1}{d-1}$.

Each restricted Markov chain consists of all the words which have a fixed trace. Hence, transitions in the restricted chains consist of only transpositions, as rotations would change the trace. The projection $\overline{P}$ consists of a simplex containing D vertices, each representing a distinct partition of $2n$. Letting

$$\Phi(X,Y) = \frac{1}{2}||X-Y||_1 \quad ,$$

two points X and Y are connected by an edge of $\overline{P}$ iff $\Phi(X,Y) = 1$, where $||\cdot||_1$ denotes the ℓ_1 metric. In the following we make no attempt to optimize the running time, and instead simply provide polynomial bounds on the convergence rates.

• Step 1 – The restricted Markov chains: Consider any of the restricted chains P_X on the set of configurations with trace X. We need to show that this simpler chain, connecting pairs of words differing by a transposition of adjacent letters, converges quickly for any fixed trace.

We can analyze the transposition moves on this set by mapping C_X to the set of linear extensions of a particular partial order. Consider the alphabet $\cup_i\{\{a_{i,1},...,a_{i,x_i}\} \cup \{A_{i,1},...,A_{i,x_i}\}\}$, and the partial order defined by the relations $a_{i,1} \prec a_{i,2} \prec ... \prec a_{i,x_i}$ and $A_{i,1} \prec A_{i,2} \prec ... \prec A_{i,x_i}$, for all i. It is straightforward to see that there is a bijection between the set of circuits in $\mathcal{C}_X$ and the set of linear extensions to this partial order (mapping a^{-1} to A). Furthermore, this bijection preserves transpositions. We appeal to the following theorem due to Bubley and Dyer [3]:

Theorem 5. *The transposition Markov chain on the set of linear extensions to a partial order on n elements has mixing time $O(n^4(\log^2 n + \log\epsilon^{-1}))$.*

Referring to theorem 1, we can derive the following bound.

Corollary 2. *The Markov chain P_X has spectral gap $Gap(P_X) \geq 1/(cn^4\log^2 n)$ for some constant c.*

• Step 2 – The projection of the Markov chain: The states $\overline{\Omega}$ of the projection consist of partitions of n into d pieces, so $|\overline{\Omega}| = D$. The stationary probability of $X = (x_1,...,x_d)$ is $\overline{\pi}(X) = \binom{2n}{x_1,x_1,...,x_d,x_d}$, the number of words with these multiplicities.

The Markov kernel is defined by connecting two partitions X and Y if the distance $\Phi(X,Y) = (||x-y||_1)/2 = 1$. Before applying corollary 1 we first need to bound the mixing rate of the Markov chain defined by Metropolis probabilities. In particular, if $X = (x_1,...,x_d)$ and $Y = (x_1,...,x_i+1,...,x_j-1,...,x_d)$, then

$$\begin{aligned} P_M(X,Y) &= \frac{1}{2n^2}\cdot\min\left(1,\frac{\overline{\pi}(Y)}{\overline{\pi}(X)}\right) \\ &= \frac{1}{2n^2}\cdot\min\left(1,\frac{x_j^2}{(x_i+1)^2}\right) . \end{aligned}$$

We analyze this Metropolis chain indirectly by first considering a variant P'_M which admits a simpler path coupling proof. Using the same Markov kernel, define the transitions

$$P'_M(X,Y) = \frac{1}{2n^2(x_i+1)^2}.$$

In particular, the x_i in the denominator is the value which would be increased by the rotation. Notice that detailed balance is satisfied:

$$\frac{\overline{\pi}(X)}{\overline{\pi}(Y)} = \frac{(x_i+1)^2}{x_j^2} = \frac{P'_M(Y,X)}{P'_M(X,Y)}.$$

This guarantees that P'_M has the same stationary distribution as P_M, namely $\overline{\pi}$.

The mixing rate of this chain can be bounded directly using path coupling. Let $U \subseteq \overline{\Omega} \times \overline{\Omega}$ be pairs of states X and Y such that $\Phi(X,Y) = 1$. We couple by choosing the same pair of indices i and j, and the same bit $b \in \{-1, 1\}$ to update each of X and Y, where the probability for accepting each of these moves is dictated by the transitions of P'_M.

Lemma 2. *For any pair $(X_t, Y_t) \in U$, the expected change in distance after one step of the coupled chain is* $\mathrm{E}[\Phi(X_{t+1}, Y_{t+1})] \leq (1 - \frac{1}{n^6})\,\Phi(X_t, Y_t)$.

Proof. If $(X_t, Y_t) \in U$, then there exist coordinates k and k' such that $y_k = x_k + 1$ and $y_{k'} = x_{k'} - 1$. Without loss of generality, assume that $k = 1$ and $k' = 2$. We need to determine the expected change in distance after one step of the coupled chain. Suppose that in this move we try to add 1 to x_i and y_i and subtract 1 from x_j and y_j. We consider three cases.

Case 1: If $|\{i,j\} \cap \{1,2\}| = 0$, then both processes accept the move with the same probability and $\Phi(X_{t+1}, Y_{t+1}) = 1$.

Case 2: If $|\{i,j\} \cap \{1,2\}| = 1$, then we shall see that the expected change is also zero. Assume without loss of generality that $i = 1$ and $j = 3$, and first consider the case $b = 1$. Then we move from X to $X' = (x_1+1, x_2, x_3-1, ..., x_d)$ with probability $\frac{1}{2n^2(x_1+1)^2}$ and from Y to $Y' = (x_1+2, x_2-1, x_3-1, ..., x_d)$ with probability $\frac{1}{2n^2(x_1+2)^2}$. Since $P'_M(X,X') > P'_M(Y,Y')$, with probability $P'_M(Y,Y')$ we update both X and Y; with probability $P'_M(X,X') - P'_M(Y,Y')$ we update just X; and with all remaining probability we update neither. In the first case we end up with X' and Y', in the second we end up with X' and Y and in the final case we stay at X and Y. All of these pairs are unit distance apart, so the expected change in distance is zero. If $b = -1$, then $P'_M(X,X') = P'_M(Y,Y') = \frac{1}{2n^2(x_3+1)^2}$ and again the coupling keeps the configurations unit distance apart.

Case 3: If $|\{i,j\} \cap \{1,2\}| = 2$, then we shall see that the expected change is at most zero. Assume without loss of generality that $i = 1$, $j = 2$ and $b = 1$. The probability of moving from X to $X'' = (x_1+1, x_2-1, ..., x_d) = Y$ is $P'_M(X,X'') = \frac{1}{2n^2(x_1+1)^2}$. The probability of moving from Y to $Y'' = (x_1+2, x_2-2, ..., x_d)$ is $P'_M(Y,Y'') = \frac{1}{2n^2(x_1+2)^2}$. So with probability $P'_M(Y,Y'')$ we update both configurations, keeping them unit distance apart, and with probability $P'_M(X,X'') - P'_M(Y,Y'') \geq \frac{1}{2n^6}$ we update just X, decreasing the distance to zero. When $b = -1$ the symmetric argument shows that we again have a small chance of decreasing the distance.

Summing over all of these possibilities yields the lemma. $\square$

The path coupling theorem implies that the mixing time is bounded by $\tau(\epsilon) \leq O(n^6 \log n)$. Furthermore, we get the following bound on the spectral gap.

Theorem 6. *The Markov chain P'_M on $\overline{\Omega}$ has spectral gap $Gap(P'_M) \geq c'/(n^6 \log n)$ for some constant c'.*

This bounds the spectral gap of the modified Metropolis chain P'_M, but we can readily compare the spectral gaps of P'_M and P_M using lemma 1. Since all the transitions of P_M are at least as large as those of P'_M, we find

Corollary 3. *The Markov chain P_M on $\overline{\Omega}$ has spectral gap $Gap(P_M) \geq c'/(n^6 \log n)$.*

• Step 3 – Putting the pieces together: These bounds on the spectral gaps of the restrictions P_i and the Metropolis projection P'_M enable us to apply the decomposition theorem to derive a bound on the spectral gap of P, the original chain.

Theorem 7. *The Markov chain P is rapidly mixing on $\mathcal{C}$ and the spectral gap satisfies $Gap(P) \geq c''/(n^{12} d \log^3 n)$, for some constant c''.*

Proof. To apply corollary 1, we need to bound the parameters β and γ. We find that $\beta \geq \frac{1}{4nd}$, the minimum probability of a transition. To bound γ we need to determine what fraction of the words in C_X are neighbors of a word in C_Y if $\Phi(X,Y) = 1$ (since π is uniform within each of these sets). If $X = (x_1, ..., x_d)$ and $Y = (x_1, ..., x_i + 1, ..., x_j - 1, ..., x_n)$, this fraction is exactly the likelihood that a word in C_X has an a_i followed by an a_i^{-1}, and this is easily determined to be at least $1/n$.

Combining $\beta \geq \frac{1}{4nd}$, $\gamma \geq \frac{1}{n}$ with our bounds from lemmas 2 and 3, corollary 1 gives the claimed lower bound on the spectral gap. $\square$

4 Other applications of decomposition

The key step to applying the decomposition theorem is finding an appropriate partition of the state space. In most examples a natural choice seems to be to cluster configurations of equal probability together so that the distribution for each of the restricted chains is uniform, or so that the restrictions share some essential feature which will make it easy to bound the mixing rate.

In the example of section 3, the state space is divided into subsets, each representing a partition of n into d parts. It followed that the vertices of the projection formed a d-dimensional simplex, where the Markov kernel was formed by connecting vertices which are neighbors in the simplex. We briefly outline two other recent applications of the decomposition theorem where we get other natural graphs for the projection. In the first case graph defining the Markov kernel of the projection is one-dimensional and in the second it is a hypercube.

4.1 Independent sets

Our first example is sampling independent sets of a graph according to the Gibbs measure. Recall that $\pi(I) = \gamma^{|I|}/Z_\gamma$, where $\gamma > 0$ is an input parameter and Z_γ normalizes the distribution. There has been much activity in studying how to sample independent sets for various values of γ using a simple, natural Markov chain based on inserting, deleting or exchanging vertices at each step. Works of Luby and Vigoda [9] and Dyer and Greenhill [5] imply that this chain is rapidly mixing if $\gamma \le 2/(\Delta - 2)$, where Δ is the maximum number of neighbors of any vertex in G. It was shown by Borgs et al. [1] that this chain is slowly mixing on some graphs for γ sufficiently large.

Alternatively, Madras and Randall [12] showed that this algorithm is fast for *every* value of γ, provided we restrict the state space to independent sets of size at most $n^* = \lfloor |V|/2(\Delta+1) \rfloor$. This relies heavily on earlier work of Dyer and Greenhill [6] showing that a Markov chain defined by exchanges is rapidly mixing on the set of independent sets of fixed size k, whenever $k \le n^*$. The decomposition here is quite natural: We partition $\mathcal{I}$, the set of independent sets of G, into pieces $\mathcal{I}_k$ according to their size. The restrictions arising from this partition permit exchanges, but disallow insertions or deletions, as they exit the state space of the restricted Markov chain. These are now exactly the Markov chains proven to be rapidly mixing by Dyer and Greenhill (but with slightly greater self-loop probabilities) and hence can also be seen to be rapidly mixing. Consequently, we need only bound the mixing rate of the projection.

Here the projection is a one-dimensional graph on $\{0, ..., n^*\}$. Further calculation determines that the stationary distribution $\overline{\pi}(k)$ of the projection is unimodal in k, implying that the projection is also rapidly mixing. We refer the reader to [12] for details.

4.2 The swapping algorithm

To further demonstrate the versatility and potential of the decomposition method, we review an application of a very different flavor. In recent work, Madras and Zheng [13] show that the *swapping algorithm* is rapidly mixing for the *mean field* Ising model (i.e., the Ising model on the complete graph), as well as for a simpler toy model.

Given a graph $G = (V, E)$, the *ferromagnetic Ising model* consists of a graph G whose vertices represent particles and whose edges represent interactions between particles. A spin configuration is an assignment of *spins*, either $+$ or $-$, to each of the vertices, where adjacent vertices prefer to have the same spin. Let $J_{x,y} > 0$ be the interaction energy between vertices x and y, where $(x, y) \in E$. Let $\sigma \in \Omega = \{+, -\}^{|V|}$ be any assignment of $\{+, -\}$ to each of the vertices. The *Hamiltonian* of σ is

$$H(\sigma) = \sum_{(x,y)\in E} J_{x,y} 1_{\sigma_x \neq \sigma_y},$$

where 1_A is the indicator function which is 1 when the event A is true and 0 otherwise. The probability that the Ising spin state is σ is given by the *Gibbs distribution*:

$$\pi(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z(G)},$$

where β is inverse temperature and

$$Z(G) = \sum_{\sigma} e^{-\beta H(\sigma)}.$$

It is well known that at sufficiently low temperatures the distribution is bimodal (as a function the number of vertices assigned $+$), and any local dynamics will be slowly mixing. The simplest local dynamics, *Glauber dynamics*, is the Markov chain defined by choosing a vertex at random and flipping the spin at that vertex with the appropriate Metropolis probability.

Simulated tempering, which varies the temperature during the runtime of an algorithm, appears to be a useful way to circumvent this difficulty [8, 14]. The chain moves between m close temperatures that interpolate between the temperature of interest and very high temperature, where the local dynamics converges rapidly. The *swapping algorithm* is a variant of tempering, introduced by Geyer [7], where the state space is Ω^m and each configuration $S = (\sigma_1, ..., \sigma_m) \in \Omega^m$ consists of one sample at each temperature. The stationary distribution distribution is $\pi(S) = \Pi_{i=1}^m \pi_i(\sigma_i)$, where π_i is the distribution at temperature i. The transitions of the swapping algorithm consist of two types of moves: with probability 1/2 choose $i \in [m]$ and perform a local update of σ_i (using Glauber dynamics at this fixed temperature); with probability 1/2 choose $i \in [m-1]$ and move from $S = (\sigma_1, ..., \sigma_m)$ to $S' = (\sigma_1, ..., \sigma_{i+1}, \sigma_i, ..., \sigma_m)$, i.e., swap configurations i and $i+1$, with the appropriate Metropolis probability.

The idea behind the swapping algorithm, and other versions of tempering, is that, in the long run, the trajectory of each Ising configuration will spend equal time at each temperature, potentially greatly speeding up mixing. Experimentally, this appears to overcome obstacles to sampling at low temperatures.

Madras and Zheng show that the swapping algorithm is rapidly mixing on the mean-field Ising model at all temperatures. Let $\Omega^+ \subset \Omega$ be the set of configurations that are predominantly +, and similarly Ω^-. Define the trace of a configuration S to be Tr$(S) = (v_1, ..., v_m) \in \{+, -\}^m$ where $v_i = +$ if $\sigma_i \in \Omega^+$ and $v_i = -$ if $\sigma_i \in \Omega^-$. The analysis of the swapping chain uses decomposition by partitioning the state space according to the trace.

The projection for this decomposition is the m-dimensional hypercube where each vertex represents a distinct trace. The stationary distribution is uniform on the hypercube because, at each temperature, the likelihood of being in Ω^+ and Ω^- are equal due to symmetry. Relying on the comparison method, it suffices to analyze the following simplification of the projection: Starting at any vertex $V = (v_1, ..., v_m)$ in the hypercube, pick $i \in_u [m]$. If $i = 1$, then with probability 1/2 flip the first bit; if $i > 1$, then with probability 1/2 transpose the v_{i-1} and v_i; and with all remaining probability do nothing. This chain is easily seen to be rapidly mixing on the hypercube and can be used to infer a bound on the spectral gap of the projection chain.

To analyze the restrictions, Madras and Zheng first prove that the simple, single flip dynamics on Ω^+ is rapidly mixing at any temperature; this result is analytical, relying on the fact that the underlying graph is complete for the mean-field model. Using simple facts about Markov chains on product spaces, it can be shown that the each of the restricted chains must also be rapidly mixing (even without including any swap moves). Once again decomposition completes the proof of rapid mixing, and we can conclude that the swapping algorithm is efficient on the complete graph.

Acknowledgements

I wish to thank Russell Martin for many useful discussions, especially regarding the technical details of section 3, and Dimitris Achlioptas for improving an earlier draft of this paper.

References

1. C. Borgs, J.T. Chayes, A. Frieze, J.H. Kim, P. Tetali, E. Vigoda, and V.H. Vu. Torpid mixing of some MCMC algorithms in statistical physics. *Proc. 40th IEEE Symposium on Foundations of Computer Science*, 218–229, 1999.
2. R. Bubley and M. Dyer. Faster random generation of linear extensions. *Discrete Mathematics*, **201**:81–88, 1999.
3. R. Bubley and M. Dyer. Path coupling: A technique for proving rapid mixing in Markov chains. *Proc. 38th Annual IEEE Symposium on Foundations of Computer Science* 223–231, 1997.
4. P. Diaconis and L. Saloff-Coste. Comparison theorems for reversible Markov chains. *Annals of Applied Probability*, **3**:696–730, 1993.
5. M. Dyer and C. Greenhill. On Markov chains for independent sets. *Journal of Algorithms*, **35**: 17–49, 2000.
6. M. Dyer and C. Greenhill. A more rapidly mixing Markov chain for graph colorings. *Random Structures and Algorithms*, **13**:285–317, 1998.
7. C.J. Geyer Markov Chain Monte Carlo Maximum Likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E.M. Keramidas, ed.), 156-163. Interface Foundation, Fairfax Sta tion, 1991.
8. C.J. Geyer and E.A. Thompson. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *J. Amer. Statist. Assoc.* **90** 909–920, 1995.
9. M. Luby and E. Vigoda. Fast Convergence of the Glauber dynamics for sampling independent sets. *Random Structures and Algorithms* **15**: 229–241, 1999.
10. N. Madras and M. Piccioni. Importance sampling for families of distributions. *Ann. Appl. Probab.* **9**: 1202–1225, 1999.
11. N. Madras and D. Randall. Factoring graphs to bound mixing rates. *Proc. 37th Annual IEEE Symposium on Foundations of Computer Science*, 194–203, 1996.
12. N. Madras and D. Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, (to appear), 2001.
13. N. Madras and Z. Zheng. On the swapping algorithm. Preprint, 2001.
14. E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19** 451–458, 1992.
15. R.A. Martin and D. Randall Sampling adsorbing staircase walks using a new Markov chain decomposition method. *Proceedings of the 41st Symposium on the Foundations of Computer Science (FOCS 2000)* , 492–502, 2000.
16. R.A. Martin and D. Randall Disjoint decomposition with applications to sampling circuits in some Cayley graphs. Preprint, 2001.
17. D. Randall and P. Tetali. Analyzing Glauber dynamics by comparison of Markov chains. *Journal of Mathematical Physics*, **41**:1598–1615, 2000.
18. A.J. Sinclair. *Algorithms for random generation & counting: a Markov chain approach.* Birkhäuser, Boston, 1993.