Markov Chain Decomposition for Convergence Rate Analysis

Neal Madras* Dana Randall†

Abstract

In this paper we develop tools for analyzing the rate at which a reversible Markov chain converges to stationarity. Our techniques are useful when the Markov chain can be decomposed into pieces which are themselves easier to analyze. The main theorems relate the spectral gap of the original Markov chains to the spectral gap of the pieces. In the first case the pieces are restrictions of the Markov chain to subsets of the state space; the second case treats a Metropolis-Hastings chain whose equilibrium distribution is a weighted average of equilibrium distributions of other Metropolis-Hastings chains on the same state space.

1 Introduction and Main Results

Suppose you are studying a reversible Markov chain on a state space Ω, and you want to estimate its spectral gap (loosely, the rate it which the chain converges to equilibrium). The overall chain may be hard to analyze, but it may be made up of "pieces" that are easier to analyze. If the chain moves from piece to piece efficiently, and if each piece equilibriates rapidly, then one would expect the entire chain to equilibriate rapidly. This is the spirit of our main results.

Let Ω be the state space of our Markov chain. We want to consider discrete and general state spaces simultaneously. For the reader who is primarily interested in the discrete case, we shall try to limit our measure-theoretic notation. In particular, we shall say things like "B is a subset of Ω" when we really mean "B is a measurable subset of Ω."

To discuss probability densities on our state space, we need a reference measure λ on Ω. (For example, if Ω is discrete, then λ can be counting measure,

*Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada; madras@mathstat.yorku.ca. Research supported in part by a Research Grant from NSERC.

†School of Mathematics and College of Computing, Georgia Institute of Technology, Atlanta GA 30332-0160; randall@math.gatech.edu. Research supported in part by NSF Career Award No. CCR-970320.

while if $\Omega = \mathbf{R}^n$, then λ can be Lebesgue measure.) If ρ is a probability density function (with respect to λ) on Ω, then the associated probability of a set $B \subset \Omega$ is denoted

$$\rho[B] := \int_B \rho(x)\, \lambda(dx)\,.$$

(If λ is counting measure, then this says $\rho[B] := \sum_{x \in B} \rho(x)$.)

Let $P(x, dy)$ be the transition probability kernel of a Markov chain on Ω that is reversible with respect to a probability density π. (In many examples this kernel can be written in the form $p(x, y)\lambda(dy)$, where p is a transition density.)

To set up the framework for our first main theorem, we first describe the "pieces" of the chain P. Let $A_1, \ldots, A_m$ be subsets of Ω such that $\cup A_i = \Omega$. (In general, these subsets will not be pairwise disjoint.) For each $i = 1, \ldots, m$, we define a new Markov chain on A_i by rejecting any transition of P that would leave A_i. The transition kernel $P_{[A_i]}$ of the new chain is given by

$$P_{[A_i]}(x, B) \;=\; P(x, B) + 1_{\{x \in B\}} P(x, A_i^c) \qquad \text{for } x \in A_i,\ B \subset A_i. \tag{1}$$

It is easy to see that $P_{[A_i]}$ is reversible (on the state space A_i) with respect to the measure whose density is proportional to the restriction of π to A_i.

Next define

$$Z := \sum_{i=1}^{m} \pi[A_i], \tag{2}$$

and define the "maximum overlap" Θ of the covering $\{A_1, \ldots, A_m\}$ by

$$\Theta := \max_{x \in \Omega} |\{i : x \in A_i\}| \tag{3}$$

(where $|\cdot|$ denotes cardinality). Then we see that

$$1 \;\le\; Z \;\le\; \Theta \;\le\; m. \tag{4}$$

Next, we introduce a crude model of the movement of the original chain among the "pieces". We consider a state space $\{a_1, \ldots, a_m\}$ of m points representing our m pieces. We define the following transition probabilities for a discrete Markov chain on this finite state space:

$$P_H(a_i, a_j) := \frac{\pi[A_i \cap A_j]}{\Theta \pi[A_i]} \qquad \text{for } i \neq j \tag{5}$$

and $P_H(a_i, a_i) = 1 - \sum_{j \neq i} P_H(a_i, a_j)$.

To describe the rate of convergence to equilibrium, we shall use the spectral gap. Suppose R is a Markov chain that is reversible with respect to the probability measure ρ. Let E_ρ denote the expectation with respect to ρ:

$$E_\rho f \;=\; \int f(y)\, \rho(dy)\,. \tag{6}$$

The spectral gap of R, $\mathrm{Gap}(R)$, is defined by

$$\mathrm{Gap}(R) := \inf_f \frac{\int\int |f(x) - f(y)|^2 \, \rho(dx) R(x, dy)}{\int\int |f(x) - f(y)|^2 \, \rho(dx) \rho(dy)} \tag{7}$$

where the inf is over all non-constant functions f (i.e. functions that are not constant almost surely with respect to ρ) such that $E_\rho(f^2) < \infty$. Notice the denominator equals twice the variance of $f(X)$, where X is a random variable whose distribution is ρ.

The spectral gap is important because it can be viewed as determining the speed of convergence of the Markov chain to equilibrium. Very roughly, a chain is close to equilibrium after a few multiples of $1/\mathrm{Gap}(R)$ iterations. Alternatively, once a chain is in equilibrium, $k/\mathrm{Gap}(R)$ consecutive observations of the chain will be "statistically equivalent" to k independent samples from the equilibrium distribution of the chain. (See for example Sokal and Thomas (1989), Diaconis and Stroock (1991), Welsh (1993, Section 8.4), or Madras and Slade (1993, Section 9.2.3).)

To make the preceding intuitive descriptions more precise, we review the following well-known properties, although they will not be required for the rest of the paper. For measures μ_1 and μ_2 on the state space Ω, let $||\mu_1 - \mu_2||_2$ be the norm defined by

$$||\mu_1 - \mu_2||_2^2 \;=\; \int_\Omega |f_1(x) - f_2(x)|^2 \, \rho(dx)$$

where f_i is the density of μ_i with respect to ρ (i.e., f_i is the Radon-Nikodym derivative $d\mu_i/d\rho$). For a probability measure μ and nonnegative integer n, let μR^n be the distribution of X_n, where $X_0, X_1, \ldots$ is a Markov chain with transition kernel R and the distribution of the initial state X_0 is μ:

$$(\mu R^n)(A) \;=\; \int_\Omega R^n(x, A) \, \mu(dx) \qquad (A \subset \Omega).$$

Let $\Gamma = 1 - \mathrm{Gap}(R)$. Then for every nonnegative integer n,

$$||\mu R^n - \rho||_2 \;\le\; \Gamma^n ||\mu - \rho||_2 \, .$$

This says that the sequence of measures μR^n converges exponentially rapidly to the equilibrium measure ρ. Moreover, $\Gamma^n \le \exp(-n\mathrm{Gap}(R))$, with approximate equality in the usual case that $\mathrm{Gap}(R)$ is small. These assertions formalize the first assertion in the preceding paragraph. We note that convergence in the $||\cdot||_2$ metric implies convergence at the same rate in the "total variation" norm

$$||\mu_1 - \mu_2||_{\text{Total Variation}} \;=\; \sup_{A \subset \Omega} |\mu_1(A) - \mu_2(A)|$$

(Roberts and Rosenthal (1997), Roberts and Tweedie (2000)). The second assertion of the preceding paragraph corresponds to the fact that for every function f such that $E_\rho(f^2) < \infty$, we have

$$|\mathrm{Cov}(f(X'_n), f(X'_0))| \leq \Gamma^n \mathrm{Var} f(X'_0)),$$

where $X'_0, X'_1, \ldots$ is the Markov chain given by R and started *in equilibrium* (i.e. X'_0 has distribution ρ). See Section 9.2.3 of Madras and Slade (1993) for further discussion of this correspondence.

In this paper, we shall use the size of the spectral gap as our sole measure of speed of convergence to equilibrium. Our concern is not so much with proving whether or not a spectral gap is nonzero (i.e.. whether or not a chain converges exponentially or not). Indeed, in many interesting discrete problems the Markov chains are finite and then positivity of the spectral gap follows immediately from irreducibility. In such situations one is more concerned with estimating the size of the spectral gap. In a typical discrete problem, the size of the state space Ω is exponentially large in the size of the problem description (e.g., the state space may be a class of subsets of a given set), and the goal would be to choose an element of the state space at random according to a given distribution ρ. A Markov chain would be constructed whose equilibrium distribution is ρ, knowing that running the chain for "long enough" would result in a state that was "almost" distributed as ρ. The spectral gap determines how long "long enough" would be. If the spectral gap is exponentially small in the size of the problem, then we would have to run the chain for an exponentially long time; such a chain is said to be "slowly mixing". If the Markov chain approach is to be feasible, then we generally would like a spectral gap whose size is polynomial in the size of the problem. Such a chain is said to be "rapidly mixing". Sinclair (1993) gives a more careful description of these terms but we shall be content with less formality as our main results will be stated without using this terminology.

Our first main theorem shows that if our crude model P_H approaches equilibrium rapidly, and if the restrictions of P to each piece A_i approach equilibrium rapidly, then the original chain approaches equilibrium rapidly.

Theorem 1.1 (State Decomposition Theorem) *In the preceding framework, as given by Equations (1)–(5), we have*

$$Gap(P) \geq \frac{1}{\Theta^2} Gap(P_H) \left(\min_{i=1,\ldots,m} Gap(P_{[A_i]}) \right). \tag{8}$$

This theorem is useful when a Markov chain appears too difficult to analyze directly, but there is a natural decomposition of the state space into pieces for which the analysis is more tractable. Moreover, the decomposition allows a hybrid approach to showing rapid convergence of a Markov chain, using different techniques to bound the mixing rates of different pieces.

Theorem 1.1 decomposes the state space of the Markov chain, while our second result decomposes its equilibrium distribution. It applies specifically to reversible Metropolis-Hastings chains, which we now define. Let $R(x, dy)$ be the transition kernel of a Markov chain on Ω that is reversible with respect to a probability density ρ. Let ζ be another probability density whose support is contained in the support of ρ. Then the "Metropolis-Hastings chain for R with respect to ζ" is the new Markov chain whose transition kernel $R^{[\zeta]}$ is defined by

$$\begin{aligned} R^{[\zeta]}(x, dy) &= R(x, dy) \min\left\{1, \frac{\zeta(y)\rho(x)}{\zeta(x)\rho(y)}\right\} \quad \text{if } y \neq x \\ R^{[\zeta]}(x, \{x\}) &= 1 - \int_{\Omega \setminus \{x\}} R^{[\zeta]}(x, dy) \end{aligned} \tag{9}$$

(If the denominator $\zeta(x)\rho(y)$ is 0, then we take $R^{[\zeta]}(x, dy) = 0$.) It is easy to check that $R^{[\zeta]}$ is reversible with respect to ζ. The kernel R is often called the "proposal kernel". The idea is that $R^{[\zeta]}$ works by proposing a move and then computing a ratio that determines the probability with which the proposed move is "accepted". It is this acceptance scheme that ensures that ζ is the equilibrium distribution. We remark that the Metropolis-Hastings chain is usually defined in the more general case that does not even require R to be reversible (see Section 3), but the reversible case reduces to Equation (9).

For our second theorem, suppose that the chain of interest, P, is a Metropolis-Hastings chain for a proposal chain R with respect to a desired equilibrium density ζ (i.e. $P = R^{[\zeta]}$). Also suppose that ζ can be expressed as a convex combination of a small number of densities $\phi_0, \ldots, \phi_D$ (i.e., ζ is a "mixture density"). Think of running a Metropolis-Hastings chain for each ϕ_j, using the same proposal kernel R as in the original P. (These chains $R^{[\phi_j]}$ are the "pieces" of the original chain.) If the ϕ_j's have some "overlap" in a sense that we describe below, then we can bound the gap of the original chain in terms of the gaps of the Metropolis-Hastings chains for the ϕ_j's.

Roughly speaking, a large overlap in the following theorem corresponds to the rapid mixing of P_H in the preceding theorem. For example, suppose that ϕ_i has substantial overlap with ϕ_{i+1} and with ϕ_{i-1}. If each "piece" is rapidly mixing, then a chain which starts in the i^{th} piece will soon move into a region where ϕ_i and ϕ_{i+1} (or ϕ_{i-1}) overlap. In this way the process can move from one piece to another reasonably efficiently.

Theorem 1.2 (Density Decomposition Theorem) *Let $\phi_0, \ldots, \phi_D$ be probability densities on Ω (with respect to a common reference measure λ), and let $a_0, \ldots, a_D$ be positive numbers that add up to 1. Define the mixture density*

$$\phi_{mix} := \sum_{j=0}^{D} a_j \phi_j \,. \tag{10}$$

Let $R(x, dy)$ be a Markov chain that is reversible with respect to a probability density $\rho(x)$ on Ω. Let Gap_j (respectively, Gap_{mix}) be the spectral gap of the Metropolis-Hastings chain $R^{[\phi_j]}$ (respectively, $R^{[\phi_{mix}]}$). Finally, assume that neighboring ϕ_j's have some "overlap": that is, assume

$$\int \min\{\phi_j(x), \phi_{j+1}(x)\}\, \lambda(dx) \;\geq\; \delta \qquad (j = 0, \ldots, D-1) \tag{11}$$

for some $\delta > 0$. Then

$$Gap_{mix} \geq \frac{\delta}{2D} \min_{j=0,\ldots,D} a_j\, Gap_j. \tag{12}$$

The paper is organized as follows. The rest of Section 1 describes applications of our main results. Theorem 1.1 is closely related to an unpublished result due to Caracciolo, Pelissetto and Sokal. In their framework, the decomposition of the state space arises in the context of simulated tempering. We give a brief introduction to this sampling method and state their result (Theorem 2.1) in section 2. In section 3 we introduce the method of umbrella sampling and state a result due to Madras and Piccioni (1999). In section 4 we prove Theorem 1.1 (the State Decomposition Theorem) using the results from the previous two sections. Theorem 1.2 is proven in section 5 and is independent of the rest of the paper. Finally, Appendices A and B prove Theorem 2.1 and Proposition 3.2 respectively.

1.1 Sampling independent sets

As a simple application of Theorem 1.1, we will consider a Markov chain for sampling independent sets. Let G be a graph with vertex set V and edge set E. An *independent set* is a subset I of V with the property that no two vertices of I are joined by an edge of G. Let Ω be the collection of all independent sets, and let Ω_i be the collection of all independent sets of cardinality i (for $i = 0, 1, \ldots, |V|$). Finally, let γ be a positive real number.

The *hard core model with parameter* γ is the probability distribution h_γ on the collection of all independent sets defined by

$$h_\gamma(I) \;=\; \frac{\gamma^{|I|}}{Z_\gamma} \qquad (I \in \Omega)$$

where Z_γ is the normalizing constant $\sum_{J\in\Omega} \gamma^{|J|}$. This is a model of identical particles with short-range mutual repulsion: The particles can be located at the vertices of G, but a particle at a given vertex forbids any other particle at any adjacent vertex. The parameter γ controls the number of particles; for example, it is not hard to see that the expected size of I is an increasing function of γ.

Consider the following Markov chain on Ω, which produces a sequence $I_0, I_1, \ldots$ of independent sets by randomly inserting or deleting one vertex at a time, or

exchanging two vertices by inserting one and deleting the other in a single step. To formalize the transitions P of this new chain, we let $V^* = V \cup \{v^*\}$ be the original vertex set augmented with one auxiliary vertex; this vertex will enable us to encode which type of move (i.e., insertion, deletion or exchange) we are attempting. Let $I_t \subset V$ be the independent set at time t. Pick two vertices (u_t, v_t) uniformly at random from $V^* \times V^*$. If $v_t = v^*$, we attempt to delete u_t: i.e., if $u_t \in I_t$, then set I_{t+1} equal to $I_t \setminus \{u_t\}$ with probability $\min\{1, \gamma^{-1}\}$. If $u_t = v^*$, we attempt to insert v_t: i.e., if $v_t \in V \setminus I_t$, and if v_t is not adjacent to any vertex of I_t, then set I_{t+1} equal to $I_t \cup \{v_t\}$ with probability $\min\{1, \gamma\}$. Finally, if $u_t, v_t \neq v^*$, we attempt to exchange u_t and v_t: i.e., if $u_t \in I_t$ and v_t is not adjacent to any vertex in $I \setminus \{u_t\}$, then set I_{t+1} equal to $(I \setminus \{u_t\}) \cup \{v_t\}$ with probability 1. With all remaining probability, set I_{t+1} equal to I_t. It is not hard to see that this Markov chain is irreducible, aperiodic, and reversible with respect to h_γ.

The work of Luby and Vigoda (1997, 1999) implies that this chain is rapidly mixing if $\gamma \leq 2/(\Delta - 2)$, where Δ is the maximum number of neighbors of any vertex in G. It has been shown by Borgs et. al. (1999) that this chain is slowly mixing on some graphs if γ is sufficiently large. The problem is that large values of γ cause the particles to get too crowded, which makes it hard for the configurations to change much. We shall see that if we limit the total number of particles to a moderate value n^* (defined below), then this crowding does not occur even if γ is very large, and the modified chain mixes rapidly.

Let $n^* = \lfloor |V|/2(\Delta+1) \rfloor$, and let Ω^* be the collection of all independent sets with at most n^* vertices:

$$\Omega^* = \bigcup_{i=0}^{n^*} \Omega_i \,.$$

Let P^* be the restriction of the above Markov chain P to Ω^* (i.e. if $|I_t| = n^*$ and $u_t = v^*$, then $I_{t+1} = I_t$ with probability one). Then P^* is irreducible on Ω^*, aperiodic, and reversible with respect to the restriction of h_γ to Ω^*. We shall show that P^* is rapidly mixing for every $\gamma \geq 1/(\Delta + 1)$. The results of Luby and Vigoda (1999) can be extended to show that P^* is also rapidly mixing for $\gamma \leq 2/(\Delta - 2)$, so we can conclude that this Markov chain converges quickly for all values of $\gamma > 0$.

Our strategy for bounding the convergence rate of the Markov chain P^* on independent sets can can be described as follows. Let $A_i = \Omega_i \cup \Omega_{i+1}$, whereby $\Omega^* = \cup\, A_i$ is a decomposition of the state space into overlapping pieces, as required for theorem 1.1. We consider in turn the restrictions $P^*_{[A_i]}$ to A_i, for all i, as well as the projection P^*_H. A lower bound on the spectral gap for each of these Markov chains will establish a bound for the original Markov chain P^*, appealing to theorem 1.1.

We first establish a bound on the mixing time of each of the restricted Markov chains $P^*_{[A_i]}$. This Markov chain performs exchanges, additions, and deletions, but always stays in the set of independent sets of size i or $i+1$ on the

input graph $G = (V, E)$. Consider a new graph $G' = (V', E)$ which augments the vertex set with an isolated vertex x. The independent sets of size $i+1$ in G' correspond bijectively to the set of independent sets of size i or $i+1$ in G; if x is in the independent set, then its removal defines an independent set of size i in G, and if x is not in the independent set, removing x from the graph leaves an independent set of size $i+1$. Furthermore, taking this vertex x to be the auxiliary vertex in the description of the Markov chain, the transitions of $P^*_{[A_i]}$ can all be described as *exchanges* in G' which keep the number of vertices in the independent set fixed at $i+1$. A variant of this new Markov chain based on exchanges was analyzed by Bubley and Dyer (1997), who show that it is rapidly mixing when $i+1 \leq n^*$. More precisely, letting $n = |V|$, we derive the following bound on the spectral gap.

Theorem 1.3 *Let $A_i = \Omega_i \cup \Omega_{i+1}$, for $0 \leq i \leq n^* - 1$, and let $P^*_{[A_i]}$ be the restriction of the Markov chain P^* to this set. Then*

$$1/\mathit{Gap}(P^*_{[A_i]}) \leq cn^2 \lceil \ln(n) \rceil \max(\gamma^2, \gamma^{-2}),$$

for some constant c.

Next, we consider the chain P^*_H on $\{a_0, \ldots, a_{n^*-1}\}$. Clearly $\Theta = 2$. Observe that $P^*_H(a_i, a_j) = 0$ whenever $|i - j| > 1$. We also have

$$P^*_H(a_i, a_{i+1}) = \frac{h_\gamma(\Omega_{i+1})}{2(h_\gamma(\Omega_i) + h_\gamma(\Omega_{i+1}))} \qquad (0 \leq i < n^* - 1),$$

$$P^*_H(a_i, a_{i-1}) = \frac{h_\gamma(\Omega_i)}{2(h_\gamma(\Omega_i) + h_\gamma(\Omega_{i+1}))} \qquad (0 < i \leq n^* - 1).$$

Notice in particular that $P^*_H(a_i, a_i) = 1/2$ for $i = 1, \ldots, n^* - 2$. We shall show below that $P^*_H(a_i, a_{i+1}) \geq P^*_H(a_i, a_{i-1})$ for each $i = 1, \ldots, n^* - 2$ and $P^*_H(a_0, a_1) \geq 1/4$ (when $\gamma \geq 1/(\Delta + 1)$). Thus P^*_H is essentially a nearest-neighbor random walk on $\{0, 1, \ldots, n^* - 1\}$ with nonnegative drift; hence using the Optional Stopping Theorem for submartingales (see, e.g., Luby, Randall and Sinclair (1995)), it is rapidly mixing.

Theorem 1.4 *Let $\gamma \geq 1/(\Delta+1)$. Then the Markov chain P^*_H is rapidly mixing with*

$$1/\mathit{Gap}(P^*_H) \leq c' n^2,$$

for some constant c'.

Theorems 1.3 and 1.4 together with theorem 1.1 (the State Decomposition Theorem) allow us to conclude that the original chain P^* on Ω^* is rapidly mixing, as claimed.

Theorem 1.5 *The Markov chain P^* on Ω^*, the set of independent sets of size at most $n/2(\Delta+1)$, is rapidly mixing for all values of $\gamma \geq 1/(\Delta+1)$, with*

$$1/Gap(P^*) \leq c'' n^4 \lceil \ln(n) \rceil \max(\gamma^2, \gamma^{-2}),$$

for some constant c''.

It only remains to show that $P^*_H(a_i, a_{i+1}) \geq P^*_H(a_i, a_{i-1})$ for each $i = 1, \ldots, n^* - 2$ and that $P^*_H(a_0, a_1) \geq 1/4$ when $\gamma \geq 1/(\Delta+1)$. Fix such a γ. Since $h_\gamma(\Omega_j) = \gamma^j |\Omega_j|$, it suffices to show that $\gamma|\Omega_{i+1}| \geq |\Omega_i|$ for $i = 0, \ldots, n^* - 1$. Fix such an i and let $\mathcal{N}(i)$ be the number of pairs of independent sets (I, J) such that $I \in \Omega_i$, $J \in \Omega_{i+1}$, and $I \subset J$. For each $J \in \Omega_{i+1}$, there are exactly $i+1$ vertices of J that can be deleted to give a suitable I; therefore

$$\mathcal{N}(i) = |\Omega_{i+1}|(i+1).$$

Conversely, for each $I \in \Omega_i$, there are at least $|V| - i(\Delta+1)$ vertices that are not adjacent to (or equal to) a vertex of I; adding any such vertex to I gives a suitable J. Therefore

$$\mathcal{N}(i) \geq |\Omega_i| \left(|V| - i(\Delta+1) \right).$$

Combining the above two inequalities gives

$$\begin{aligned}
|\Omega_i| &\leq |\Omega_{i+1}| \frac{(i+1)}{|V| - i(\Delta+1)} \\
&\leq |\Omega_{i+1}| \frac{n^*}{|V| - n^*(\Delta+1)} \\
&\leq |\Omega_{i+1}| \frac{|V|/2(\Delta+1)}{|V| - |V|/2} \\
&= |\Omega_{i+1}| \frac{1}{\Delta+1} \\
&\leq |\Omega_{i+1}| \gamma,
\end{aligned}$$

which was what we wanted to prove.

A simpler Markov chain which has also been studied in the context of independent sets is based on just insertions and deletions (without allowing exchanges). The analysis of P^* above can be used to infer that this simpler Markov chain is also rapidly mixing on Ω^* by using the comparison method of Diaconis and Saloff-Coste (1993). We refer the reader to Randall and Tetali (1998) for a description of how the comparison method can be applied in the context of independent sets.

1.2 Other applications

For another example, imagine a Markov chain defined on a state space $\Omega = \cup \Omega_i$, where the pieces Ω_i form a partition. Moreover, assume the sizes $|\Omega_i|$ are unimodal as a function of i. (For example, let Ω be the set of matchings of some

underlying graph G, and Ω_i is the set of matchings of size i. A simple, ergodic Markov chain on the state space of matchings can be defined by adding, removing or exchanging edges in a single transition; see Broder (1986) and Jerrum and Sinclair (1989) for details. In this example $|\Omega_i|$ is always a logconcave, and therefore unimodal, function of i.) Defining $A_i = \Omega_i \cup \Omega_{i+1}$, the Markov chain $P_H(a_i, a_j)$ is a one dimensional random walk with bias towards the mode. Therefore Theorem 1.1 provides a bound on the spectral gap of the Markov chain in terms of the restricted Markov chains $P_{[A_i]}$. (In the case of matchings, Jerrum and Sinclair's method directly bounds the the spectral gap of the Markov chain, and hence this decomposition is not necessary for that application.)

More complicated applications have been worked out, and we shall give only brief descriptions of them here. Madras and Randall (1996) gave a different proof of a bound similar to Theorem 1.1, but that paper worked with conductances instead of spectral gaps, and the multiplicative factor in their inequality was not as good as the Θ^{-2} that appears in the present Theorem 1.1. (That paper then applied the result to a Markov chain for three-colorings on the square lattice, but the application contained an error; to our knowledge designing an efficient sampling algorithm for three-colorings remains open.) Madras and Piccioni (1999) used Theorem 1.2 to analyze an implementation of the method of "simulated tempering" to a special "witch's hat" distribution, as originally studied empirically in Geyer and Thompson (1995). Zheng (1999) uses both of our main results to study the Metropolis-coupled Markov chain method of Geyer (1991) (see Orlandini (1998) for interesting recent applications of this method). Cooper et. al. (2000) have applied our results to show that the Wolff chain for the Potts model is rapidly mixing on an $n \times \mathrm{O}(1) \times ... \times \mathrm{O}(1)$ grid.

2 Simulated Tempering

The method of Simulated Tempering was proposed independently by Marinari and Parisi (1992) in the physics literature and Geyer and Thompson (1995) in the statistics literature (see Madras (1998) for a review). To explain the idea, consider the following motivating example from statistical physics. Let G be a graph, and let Ω be the set of all functions from the vertex set of G into $\{-1, +1\}$. The *Ising model on* G is a certain family of probability distributions on Ω parametrized by a real number β which determines the strength of interactions between neighboring vertices. (There is often a second parameter, the "external field", but we shall fix it equal to 0.) When $\beta = 0$, the distribution is simply the uniform distribution on Ω. When β is large, then the distribution is "bimodal": with high probability, we see either lots of $+1$'s and few -1's, or vice versa. How does one sample from this model by Markov chains? The simplest way is by the single-spin Metropolis algorithm (Section 3): pick a vertex at random and try to change the sign at that vertex. Accept the change with a certain probability

(which is easy to compute). When β equals 0, the change is always accepted, and the Markov chain is rapidly mixing. When β is close to 0, the chain is also rapidly mixing. However, when β is large, the chain takes exponentially long to get from one "mode" to the other, and it mixes exponentially slowly (see for example Madras and Piccioni (1999)). To set the stage for simulated tempering, let ϕ_1 be the distribution with $\beta = 0$, ϕ_m the distribution for a given large β, and ϕ_i $(i = 2, \ldots, m-1)$ be distributions at equally spaced intermediate values of β. For each i, the $\mathcal{T}_i$ that we shall introduce below will be the Metropolis chain for the corresponding ϕ_i. The description for this example will continue below after the general framework is developed.

In the general Simulated Tempering framework, we have a state space Ω with m different probability densities $\phi_1(x), \ldots, \phi_m(x)$ (with respect to a common reference measure $\lambda(dx)$). For each i, let $\mathcal{T}_i(x, dy)$ be a transition kernel that is reversible with respect to ϕ_i.

Next we define the "augmented" state space $\mathcal{S}$ by including the "labels" 1 through m:

$$\mathcal{S} \;=\; \Omega \times \{1, \ldots, m\}. \tag{13}$$

For each $i = 1, \ldots, m$, define

$$\mathcal{S}_i \;=\; \Omega \times \{i\} \;=\; \{(x, i) : \, x \in \Omega\}\,. \tag{14}$$

Thus $\mathcal{S}_1, \ldots, \mathcal{S}_m$ forms a partition of $\mathcal{S}$. Define the transition probability kernel $\mathcal{P}$ on $\mathcal{S}$ as follows:

$$\mathcal{P}((x, i), (dy, j)) \;=\; \begin{cases} 0 & \text{if } j \neq i \\ \mathcal{T}_i(x, dy) & \text{if } j = i \end{cases} \tag{15}$$

Notice that this kernel does not permit transitions from one $\mathcal{S}_i$ to another.

Next, suppose that we associate a positive number c_i with each ϕ_i, such that $\sum_{i=1}^m c_i = 1$. These "weights" permit us to define the transition kernel $\mathcal{Q}$ on $\mathcal{S}$:

$$\mathcal{Q}((x, i), (dy, j)) \;=\; \delta_x(dy) \frac{c_j \phi_j(x)}{\sum_{l=1}^m c_l \phi_l(x)} \tag{16}$$

Thus, $\mathcal{Q}$ keeps x the same, but chooses the label according to the weighted probabilities of x under the different densities. Observe that $\mathcal{Q}^2 = \mathcal{Q}$.

Define the probability measure ψ on $\mathcal{S}$ by

$$\psi(dx, i) \;=\; c_i \phi_i(x) \lambda(dx) \qquad ((x, i) \in \mathcal{S}). \tag{17}$$

One can check that both $\mathcal{Q}$ and $\mathcal{P}$ are reversible with respect to ψ. Notice that the marginal probability of the "label" i is c_i, and the marginal distribution of the "configuration" x is

$$\bar{\phi}(x) \;:=\; \sum_{l=1}^{m} c_l \phi_l(x) \tag{18}$$

Therefore we can view $\mathcal{Q}$ as replacing the current label by sampling from the conditional distribution of the label given the configuration x.

The simulated tempering method is the repeated alternation of $\mathcal{Q}$ with $\mathcal{P}$. So we could describe it as the Markov chain corresponding to $\mathcal{QP}$ or to $\mathcal{PQ}$, or to $\mathcal{QPQ}$ (recall that $\mathcal{Q}^2 = \mathcal{Q}$). We shall use the version $\mathcal{QPQ}$, since it is reversible with respect to ψ (it inherits this property from $\mathcal{Q}$ and $\mathcal{P}$).

(To understand what is happening, we shall refer again to the Ising model. Applying $\mathcal{P}$ causes us to update the configuration from Ω by the Metropolis chain corresponding to the current value of the label i. Then we apply $\mathcal{Q}$, which permits the current value of the label to change. Then we apply $\mathcal{P}$ again, updating the configuration according to the Metropolis chain corresponding to the new value of the label. And so on. The intuition is the following: The label will do a "random walk" on $\{1, \ldots, m\}$. When the label equals m, we observe configurations that have been sampled from ϕ_m, which is what we want. When the label is 1, or close to 1, the chain is mixing rapidly, so that the next time the label gets back up to m we can expect to see a configuration that is pretty different from the last time that the label equaled m. We can make sure that the chain spends enough time with the label taking both extreme values; indeed, in the long run, the fraction of time spent with the label equal to i is c_i. If this intuition is correct, then the overall chain should be rapidly mixing. This seems to work well in practice, although there are some substantial issues of implementation that arise. The rapid mixing of this procedure has been proven for the Ising model on the complete graph, but not on the more interesting graphs corresponding to Euclidean lattices. See Madras and Piccioni (1999) for more details.)

Finally, we define the "aggregated transition matrix" $\overline{\mathcal{Q}}$ (an analogue of P_H defined in the Introduction):

$$\begin{aligned} \overline{\mathcal{Q}}(i,j) &= \frac{1}{c_i} \int_\Omega \frac{c_i \phi_i(x) c_j \phi_j(x)}{\sum_{l=1}^m c_l \phi_l(x)} \lambda(dx) \\ &= c_j \int_\Omega \frac{\phi_i(x)\phi_j(x)}{\bar{\phi}(x)} \lambda(dx) \qquad (i,j = 1, \ldots, m). \end{aligned} \tag{19}$$

The next theorem says roughly that simulated tempering cannot be any slower than the combination of the random walk on labels and the individual chains $\mathcal{T}_i$ within each piece. (At first sight, this is not quite what we want to know: In the Ising model example, $\mathcal{T}_m$ is very slow, but we hope that simulated tempering is efficient. So in such cases, one would want to find different ways to decompose the state space into pieces that are rapidly mixing. But this is not easy to do.)

Theorem 2.1 (Caracciolo–Pelissetto–Sokal) *In the framework of simulated tempering described above, we have*

$$Gap(\mathcal{QPQ}) \geq Gap(\overline{\mathcal{Q}}) \min_{i=1,\ldots,m} Gap(\mathcal{T}_i). \tag{20}$$

This theorem is from a 1992 unpublished manuscript by S. Caracciolo, A. Pelissetto, and A.D. Sokal. Since these three authors do not intend to publish their manuscript in the foreseeable future, they have given us permission to present their proof here. It appears in Appendix A.

3 Metropolis Algorithm and Umbrella Sampling

In this section we discuss two more Markov chains that are useful in Monte Carlo simulations, and we describe some needed results from Madras and Piccioni (1999).

First we mention the *Metropolis-Hastings method*. Let $R(x, dy)$ be the transition kernel of a Markov chain on Ω. Let ζ be a probability measure on Ω. Then the "Metropolis-Hastings chain for R with respect to ζ" is the new Markov chain whose transition kernel $R^{[\zeta]}$ is formally defined by

$$\begin{aligned} R^{[\zeta]}(x, dy) &= R(x, dy) \min\left\{1, \frac{\zeta(dy)R(y, dx)}{\zeta(dx)R(x, dy)}\right\} \qquad \text{if } y \neq x \\ R^{[\zeta]}(x, \{x\}) &= 1 - \int_{\Omega \setminus \{x\}} R^{[\zeta]}(x, dy)\,. \end{aligned} \tag{21}$$

This is a formal definition, and we refer the reader to Tierney (1998) for a discussion of the general situation. Fortunately, in many common situations it is easy to interpret Equation (21). For example suppose that the measure ζ has a density, which we shall also call ζ, with respect to some reference measure λ (i.e. $\zeta(dx) = \zeta(x)\lambda(dx)$). Then:

- If R has a transition density r with respect to λ, so that $R(x, dy) = r(x, y)\lambda(dy)$, then we have

$$R^{[\zeta]}(x, dy) = R(x, dy) \min\left\{1, \frac{\zeta(y)r(y, x)}{\zeta(x)r(x, y)}\right\} \qquad \text{if } y \neq x;$$

- If R is reversible with respect to the density $\rho(x)\lambda(dx)$, then we obtain Equation (9).

In both of these cases it is easy to check that $R^{[\zeta]}$ is reversible with respect to ζ. This paper only considers reversible chains, so we just need the second case, Equation (9). The terminology "Metropolis chain" is frequently used in the second case when ρ is constant.

The following lemma is a consequence of the definition of the spectral gap (7). The proof is essentially the same as was given in Madras and Piccioni (1999) for the case of Metropolis chains.

Lemma 3.1 *Let r_1 and r_2 be two densities with respect to λ on Ω. Suppose that $R(x, dy)$ is reversible with respect to the density ρ. Also suppose that there are constants a and b such that*

$$a \leq \frac{r_1(x)}{r_2(x)} \leq b \tag{22}$$

for all $x \in \Omega$ such that r_1 and r_2 do not vanish simultaneously. Then the spectral gaps of the associated Metropolis-Hastings chains satisfy

$$\frac{a}{b}\, Gap(R^{[r_2]}) \;\leq\; Gap(R^{[r_1]}) \;\leq\; \frac{b}{a}\, Gap(R^{[r_2]})\,. \tag{23}$$

Next we discuss the method known as *Umbrella Sampling*. Consider a probability density κ on Ω which is written as a convex combination of m other densities: that is, there are m probability densities $\phi_1, \ldots, \phi_m$ and m positive constants $c_1, \ldots, c_m$ such that $\sum_{i=1}^m c_i = 1$ and

$$\kappa(x) \;=\; \sum_{i=1}^{m} c_i \phi_i(x)\,. \tag{24}$$

Now consider a transition kernel $R(x, dy)$ of a Markov chain on Ω that is reversible with respect to a probability density ρ. For each $i = 1, \ldots, m$, define the transition kernel $\mathcal{T}_i(x, dy) = R^{[\phi_i]}$, the Metropolis-Hastings chain for R with respect to ϕ_i.

In physical applications, the ϕ_i's are often natural distributions from which we want to sample, and the mixture κ is an artificial "umbrella" distribution. Using the Metropolis-Hastings chain $R^{[\kappa]}$ to sample from κ (together with the classical Monte Carlo technique of *importance sampling*) allows one to sample from all ϕ_i's in a single simulation run. Torrie and Valleau (1977) first realized the power of this approach for physical systems, and it was they who introduced the term "umbrella sampling". See Madras and Piccioni (1999) for further discussion. Moreover, this umbrella sampling can be far more efficient than running the m chains $R^{[\phi_i]}$ separately. It also turns out that umbrella sampling is at least as good as simulated tempering, in the following sense.

Proposition 3.2 (Madras–Piccioni) *Suppose that the above Ω, ϕ_i's, c_i's, and $\mathcal{T}_i$'s are used to define the simulated tempering chain $\mathcal{QPQ}$ of Section 2 on the augmented state space $\Omega \times \{1, \ldots, m\}$. Then*

$$Gap(\mathcal{QPQ}) \;\leq\; Gap(R^{[\kappa]})\,.$$

Note that the invariant measure of $\mathcal{QPQ}$, ψ, is described in Equation (17). Proposition 3.2 is formulated in Madras and Piccioni (1999) in a slightly different way. It is easier to present the (short) proof than to explain how to modify it, so we shall do this in Appendix B.

4 State Decomposition

To prove the State Decomposition Theorem 1.1, we shall interpret it in terms of the constructions of Section 3. We are given $P(x, dy)$, a Markov transition kernel on Ω that is reversible with respect to the density $\pi(x)$ (all densities are with respect to $\lambda(dx)$). We are also given $A_1, \ldots, A_m$, subsets of Ω such that $\cup A_i = \Omega$. For each $i = 1, \ldots, m$, let ϕ_i be the normalized restriction of π to A_i:

$$\phi_i(x) \;=\; \frac{\pi(x)\, 1_{A_i}(x)}{\pi[A_i]}\,. \tag{25}$$

Recall

$$Z \;=\; \sum_{i=1}^{m} \pi[A_i] \quad \text{and} \quad \Theta \;:=\; \max_{x\in\Omega} |\{i : x \in A_i\}|. \tag{26}$$

Also let

$$c_i \;=\; \frac{\pi[A_i]}{Z} \qquad (i = 1, \ldots, m), \tag{27}$$

and define

$$\kappa(x) \;=\; \sum_{i=1}^{m} c_i \phi_i(x)\,. \tag{28}$$

Then κ is a probability density, and

$$\frac{1}{Z}\pi(x) \;\le\; \kappa(x) \;\le\; \frac{\Theta}{Z}\pi(x)\,. \tag{29}$$

Let $P^{[\kappa]}$ be the Metropolis-Hastings chain for P with respect to κ. Since P is reversible with respect to π, we have

$$P^{[\pi]} \;=\; P\,. \tag{30}$$

Therefore Lemma 3.1 and Equation (29) imply that

$$\frac{1}{\Theta}\mathrm{Gap}(P) \;\le\; \mathrm{Gap}(P^{[\kappa]}) \;\le\; \Theta\mathrm{Gap}(P)\,. \tag{31}$$

The restriction $P_{[A_i]}$ of P to A_i is the same as the Metropolis-Hastings chain $P^{[\phi_i]}$. Let

$$R(x, dy) = P^{[\kappa]}(x, dy) \quad \text{and} \quad \mathcal{T}_i(x, dy) = R^{[\phi_i]}(x, dy) \quad (i = 1, \ldots, m). \tag{32}$$

Observe that we can write

$$\mathcal{T}_i(x, dy) \;=\; P(x, dy) \min\left\{1, \frac{\kappa(y)\pi(x)}{\kappa(x)\pi(y)}\right\} \min\left\{1, \frac{\pi(y)1_{A_i}(y)\kappa(x)}{\pi(x)1_{A_i}(x)\kappa(y)}\right\} \quad \text{whenever } x \neq y \text{ and } x, y \in A_i. \tag{33}$$

It follows from this and Equation (29) that

$$\frac{1}{\Theta}P^{[\phi_i]}(x,dy) \;\le\; \mathcal{T}_i(x,dy) \;\le\; P^{[\phi_i]}(x,dy) \qquad \text{whenever } x \neq y. \tag{34}$$

Since $P^{[\phi_i]}$ and $\mathcal{T}_i$ are both reversible with respect to ϕ_i, it follows from the above bounds and Equation (7) that

$$\frac{1}{\Theta}\mathrm{Gap}(P^{[\phi_i]}) \;\le\; \mathrm{Gap}(\mathcal{T}_i) \;\le\; \mathrm{Gap}(P^{[\phi_i]})\,. \tag{35}$$

The aggregated transition matrix of Section 2 is

$$\begin{aligned}
\overline{\mathcal{Q}}(i,j) &= c_j \int_\Omega \frac{\phi_i(x)\phi_j(x)}{\kappa(x)}\,\lambda(dx) \\
&= c_j \int_{A_i\cap A_j} \frac{\pi(x)^2}{\kappa(x)\pi[A_i]\pi[A_j]}\,\lambda(dx) \\
&= \frac{1}{\pi[A_i]} \int_{A_i\cap A_j} \frac{\pi(x)^2}{Z\kappa(x)}\,\lambda(dx)\,.
\end{aligned} \tag{36}$$

Recalling the definition of P_H (Equation (5)), and using Equations (29) and (36), we see that

$$P_H(a_i,a_j) \;\le\; \overline{\mathcal{Q}}(i,j) \;\le\; \Theta P_H(a_i,a_j) \qquad \text{for } i \neq j. \tag{37}$$

Since both P_H and $\overline{\mathcal{Q}}$ are reversible with respect to the same probability distribution (namely, the one whose weights are the c_i's), Equations (37) and (7) imply that

$$\mathrm{Gap}(P_H) \;\le\; \mathrm{Gap}(\overline{\mathcal{Q}}) \;\le\; \Theta\mathrm{Gap}(P_H)\,. \tag{38}$$

Finally we put the pieces together:

$$\begin{aligned}
\mathrm{Gap}(P) &\ge \frac{1}{\Theta}\mathrm{Gap}(P^{[\kappa]}) \quad \text{(by Equation (31))} \\
&\ge \frac{1}{\Theta}\mathrm{Gap}(\mathcal{Q}\mathcal{P}\mathcal{Q}) \quad \text{(by Proposition 3.2 with } R=P \text{ and } \rho=\pi) \\
&\ge \frac{1}{\Theta}\mathrm{Gap}(\overline{\mathcal{Q}}) \min_{i=1,\ldots,m} \mathrm{Gap}(\mathcal{T}_i) \quad \text{(by Theorem 2.1)} \\
&\ge \frac{1}{\Theta^2}\mathrm{Gap}(P_H) \min_{i=1,\ldots,m} \mathrm{Gap}(P^{[\phi_i]}) \quad \text{(by Equations (38) and (35))}
\end{aligned} \tag{39}$$

Since $P^{[\phi_i]} = P_{[A_i]}$, this completes the proof of Theorem 1.1.

5 Density Decomposition

This section consists of the proof of the Density Decomposition Theorem 1.2, which is essentially independent of the rest of the paper. As usual, all densities are with respect to a reference measure λ on Ω.

For a given probability density h, we let E_h and V_h respectively denote expectation and variance with respect to h. We write E_j and E_{mix} instead of E_{ϕ_j} and $E_{\phi_{mix}}$, and similarly for V_j and V_{mix}.

For an arbitrary function f on the state space, and for $j = 0, \ldots, D$, define

$$B_j(f) \;=\; \int\int \Big(f(x) - f(y)\Big)^2 R(x, dy) \min\left\{\frac{\phi_j(x)}{\rho(x)}, \frac{\phi_j(y)}{\rho(y)}\right\} \rho(x)\lambda(dx)$$

(and define $B_{mix}(f)$ analogously). Then the spectral gap of the Metropolis-Hastings chain for R with respect to ϕ_j is given by

$$\mathrm{Gap}_j \;=\; \inf_f \frac{B_j(f)}{2V_j(f)} \tag{40}$$

Since

$$\min\left\{\frac{\phi_{mix}(x)}{\rho(x)}, \frac{\phi_{mix}(y)}{\rho(y)}\right\} \;\geq \sum_{j=0}^{D} a_j \min\left\{\frac{\phi_j(x)}{\rho(x)}, \frac{\phi_j(y)}{\rho(y)}\right\},$$

it follows that for every f

$$\begin{aligned}
B_{mix}(f) \;&=\; \int\int \Big(f(x) - f(y)\Big)^2 R(x, dy) \min\left\{\frac{\phi_{mix}(x)}{\rho(x)}, \frac{\phi_{mix}(y)}{\rho(y)}\right\} \rho(x)\lambda(dx) \\
&\geq\; \sum_j a_j B_j(f) \\
&\geq\; \sum_j a_j \mathrm{Gap}_j 2V_j(f) \\
&\geq\; 2 \min_i\{a_i \mathrm{Gap}_i\} \sum_{j=0}^{D} V_j(f).
\end{aligned} \tag{41}$$

Thus, to prove the theorem it suffices to show that for every f such that $E_{mix}(f^2) < \infty$,

$$2V_{mix}(f) \;\leq\; \frac{4D}{\delta} \sum_{j=0}^{D} V_j(f). \tag{42}$$

Let f be an arbitrary function such that $E_{mix}(f^2) < \infty$. Then we also have $E_i(f^2) < \infty$ for every $i = 1, \ldots, m$. For $i, j = 0, \ldots, D$, define

$$C_{ij}(f) \;=\; \int\int \Big(f(x) - f(y)\Big)^2 \phi_i(x)\phi_j(y)\, \lambda(dx)\, \lambda(dy)\,.$$

Then

$$C_{jj}(f) \;=\; 2V_j(f)$$

and

$$\begin{aligned} 2V_{mix}(f) &= \int\int \Big(f(x)-f(y)\Big)^2 \left(\sum_i a_i\phi_i(x)\right)\left(\sum_j a_j\phi_j(y)\right)\,\lambda(dx)\,\lambda(dy) \\ &= \sum_{i,j} a_i a_j C_{ij}(f). \end{aligned}$$

In particular, we have

$$2V_{mix}(f) \;\le\; \max_{i,j} C_{ij}(f). \tag{43}$$

Fix j. By the overlap condition (11), there exist probability densities η, τ, and ψ such that

$$\phi_j \;=\; \delta\eta + (1-\delta)\tau \quad \text{ and } \quad \phi_{j+1} \;=\; \delta\eta + (1-\delta)\psi.$$

Then

$$\begin{aligned} 2V_j(f) &= \int\int \Big(f(x)-f(y)\Big)^2 \\ &\qquad \Big(\delta\eta(x)+(1-\delta)\tau(x)\Big)\Big(\delta\eta(y)+(1-\delta)\tau(y)\Big)\,\lambda(dx)\,\lambda(dy) \\ &= 2\delta^2 V_\eta(f) + 2(1-\delta)^2 V_\tau(f) \\ &\qquad + 2\delta(1-\delta)\int\int \Big(f(x)-f(y)\Big)^2 \eta(x)\tau(y)\,\lambda(dx)\,\lambda(dy) \end{aligned} \tag{44}$$

and

$$\begin{aligned} &C_{j,j+1}(f) \\ &= \int\int \Big(f(x)-f(y)\Big)^2 \\ &\qquad \Big(\delta\eta(x)+(1-\delta)\tau(x)\Big)\Big(\delta\eta(y)+(1-\delta)\psi(y)\Big)\,\lambda(dx)\,\lambda(dy) \\ &= 2\delta^2 V_\eta(f) + (1-\delta)^2 \int\int \Big(f(x)-f(y)\Big)^2 \tau(x)\psi(y)\,\lambda(dx)\,\lambda(dy) \\ &\qquad + \delta(1-\delta)\int\int \Big(f(x)-f(y)\Big)^2 \Big(\eta(x)\psi(y)+\eta(y)\tau(x)\Big)\,\lambda(dx)\,\lambda(dy)\,. \end{aligned} \tag{45}$$

¿From (44) we find

$$\int\int \Big(f(x)-f(y)\Big)^2 \eta(x)\tau(y)\,\lambda(dx)\,\lambda(dy) \;\le\; \Big(V_j(f)-\delta^2 V_\eta(f)\Big)/(\delta(1-\delta))\,. \tag{46}$$

Using $(u+v)^2 \le 2u^2 + 2v^2$, we obtain

$$\begin{aligned}
&\int\int \Big(f(x)-f(y)\Big)^2 \tau(x)\psi(y)\,\lambda(dx)\,\lambda(dy) \\
&= \int\int\int \Big(f(x)-f(z)+f(z)-f(y)\Big)^2 \tau(x)\psi(y)\eta(z)\,\lambda(dx)\,\lambda(dy)\,\lambda(dz) \\
&\le 2\int\int\int \Big(\Big(f(x)-f(z)\Big)^2 + \Big(f(z)-f(y)\Big)^2\Big)\tau(x)\psi(y)\eta(z)\,\lambda(dx)\,\lambda(dy)\,\lambda(dz) \\
&= 2\int\int \Big(f(x)-f(z)\Big)^2 \tau(x)\eta(z)\,\lambda(dx)\,\lambda(dz) \\
&\qquad\qquad + 2\int\int \Big(f(z)-f(y)\Big)^2 \psi(y)\eta(z)\,\lambda(dy)\,\lambda(dz)\,.
\end{aligned}$$

Inserting this into (45), and then applying (46) (and the analogue of (46) for $j+1$), we see

$$\begin{aligned}
&C_{j,j+1}(f) \\
&\le 2\delta^2 V_\eta(f) + \Big(2(1-\delta)^2 + \delta(1-\delta)\Big)\Big(\int\int \Big(f(x)-f(z)\Big)^2 \tau(x)\eta(z)\,\lambda(dx)\,\lambda(dz) \\
&\qquad\qquad + \int\int \Big(f(z)-f(y)\Big)^2 \psi(y)\eta(z)\,\lambda(dy)\,\lambda(dz)\Big)\Big) \\
&= 2\delta^2 V_\eta(f) + (2-\delta)(1-\delta)\Big(\int\int \Big(f(x)-f(y)\Big)^2 \eta(x)\tau(y)\,\lambda(dx)\,\lambda(dy) \\
&\qquad\qquad + \int\int \Big(f(x)-f(y)\Big)^2 \eta(x)\psi(y)\,\lambda(dx)\,\lambda(dy)\Big) \\
&\le 2\delta^2 V_\eta(f) + (2-\delta)\Big(V_j(f) - \delta^2 V_\eta(f) + V_{j+1}(f) - \delta^2 V_\eta(f)\Big)/\delta \\
&\le \frac{2-\delta}{\delta}\Big(V_j(f) + V_{j+1}(f)\Big)\,. \qquad (47)
\end{aligned}$$

Let $X^{(0)}, \ldots, X^{(D)}$ be independent random variables, with $X^{(i)}$ having density ϕ_i. Then for $i < j$,

$$\begin{aligned}
C_{ij}(f) &= E\Big(\Big(f(X^{(i)}) - f(X^{(j)})\Big)^2\Big) \\
&= E\Big(\Big(\sum_{k=i}^{j-1} f(X^{(k)}) - f(X^{(k+1)})\Big)^2\Big) \\
&\le E\Big((j-i)\sum_{k=i}^{j-1}\Big(f(X^{(k)}) - f(X^{(k+1)})\Big)^2\Big) \\
&= (j-i)\sum_{k=i}^{j-1} C_{k,k+1}(f)
\end{aligned}$$

(where we used the Schwarz inequality in the third line). Therefore, applying (47), we see that for all $i \neq j$,

$$\begin{aligned} C_{ij}(f) &\leq D \sum_{k=0}^{D-1} C_{k,k+1}(f) \\ &\leq \frac{2(2-\delta)D}{\delta} \sum_{l=0}^{D} V_l(f)\,. \end{aligned} \tag{48}$$

Notice that the last expression in (48) is also a bound for the case $i = j$, because $C_{jj}(f) = 2V_j(f)$. Therefore (48) and (43) imply (42), and the theorem is proven.

Remark: An inspection of the final paragraph of the proof shows that Theorem 1.2 can be generalized to the case that the overlapping ϕ_j's are not linearly arranged. More precisely, we can replace the last two sentences in the statement of the theorem with the following: *Fix $\delta > 0$, and consider a graph whose vertices are $0, 1, \ldots, D$, with an edge joining i to j if and only if*

$$\int \min\{\phi_i(x), \phi_j(x)\}\, \lambda(dx) \;\geq\; \delta\,.$$

Let M be the diameter of this graph, i.e.

$$M = \max_{i,j} \;\{\text{minimum number of edges in a path from } i \text{ to } j\;\}$$

Then

$$\mathrm{Gap}_{mix} \geq \frac{\delta}{2M} \min_{j=0,\ldots,D} a_j \mathrm{Gap}_j.$$

6 Acknowledgments

We are grateful to Sergio Caracciolo, Andrea Pelissetto, and Alan Sokal for discussing their unpublished work with us, and for allowing us to present their proof as Appendix A of this paper. We also thank the referee for comments that have made the paper more readable (we hope), and for additional references.

A The Caracciolo–Pelissetto–Sokal Result

This Appendix contains the proof of Theorem 2.1 as a consequence of a more general result (Theorem A.1 below), due to Caracciolo, Pelissetto, and Sokal (1992). The proof given here is their proof, with only editorial changes.

The proof is based on the the theory of operators in a Hilbert space. To start, we shall describe the spaces in which we work, and give some equivalent descriptions of the spectral gap of a reversible (self-adjoint) operator.

Let ρ be a probability measure on a state space $\mathcal{S}$. For functions f and g on $\mathcal{S}$, we define

$$(f,g)_\rho \;=\; \int_{\mathcal{S}} f(x)g(x)\,\rho(dx) \tag{49}$$

the inner product on $L^2(\rho)$, the Hilbert space all functions that are square-integrable with respect to ρ. (If $\mathcal{S}$ is discrete, then of course all integrals become sums.)

Let $R(x,dy)$ be the transition kernel of a Markov chain that is reversible with respect to ρ. That is,

$$\rho(dx)R(x,dy) \;=\; \rho(dy)R(y,dx), \tag{50}$$

or, more formally,

$$(f,Rg)_\rho \;=\; (Rf,g)_\rho \qquad \text{for all } f,g \in L^2(\rho), \tag{51}$$

where we write

$$Rf(x) \;=\; \int_{\mathcal{S}} R(x,dy)f(y)\,.$$

Let Π_ρ be the projection operator that sends each function to the constant function identical to its mean:

$$(\Pi_\rho f)(x) \;:=\; (f,1)_\rho \;=\; \int_{\mathcal{S}} f(y)\,\rho(dy) \qquad \text{for all } x \in \mathcal{S}. \tag{52}$$

The spectral gap of R, $\mathrm{Gap}(R)$, is defined by

$$\mathrm{Gap}(R) \;=\; \inf \frac{(f,(I-R)f)_\rho}{(f,(I-\Pi_\rho)f)_\rho} \tag{53}$$

$$\;=\; \inf \frac{\int\int |f(x)-f(y)|^2\,\rho(dx)R(x,dy)}{2\int |f(x)-(f,1)_\rho|^2\,\rho(dx)} \tag{54}$$

where the inf is over all non-constant functions f in $L^2(\rho)$. Notice that the denominator in (53) equals the variance of $f(X)$ if X is a random variable with distribution ρ.

Let $\mathbf{1}^\perp$ be the orthogonal complement of the constant functions in $L^2(\rho)$:

$$\mathbf{1}^\perp \;:=\; \{f \in L^2(\rho) :\; (f,1)_\rho = 0\} \;=\; \{f \in L^2(\rho) :\; \int_{\mathcal{S}} f(x)\,\rho(dx) = 0\}\,.$$

Observe that $(I-R)f \in \mathbf{1}^\perp$ whenever $f \in \mathbf{1}^\perp$; therefore we can view $I-R$ as an operator on the Hilbert space $\mathbf{1}^\perp$. We shall write $\mathrm{Spec}_{\mathbf{1}^\perp}(T)$ to denote

the spectrum of the operator T on $\mathbf{1}^\perp$. For the reversible probability transition operator R, it is well known that $\mathrm{Spec}_{\mathbf{1}^\perp}(R)$ is a subset of the real interval $[-1, 1]$.

Observe that the equations (53) and (54) still hold if we take the inf over $f \in \mathbf{1}^\perp$. Thus we obtain

$$\begin{aligned} \mathrm{Gap}(R) &= \inf_{f \in \mathbf{1}^\perp} \frac{(f, (I-R)f)_\rho}{(f,f)_\rho} \\ &= \inf \mathrm{Spec}_{\mathbf{1}^\perp}(I-R) \qquad \text{(by p. 320 of Yosida (1980))} \\ &= 1 - \sup \mathrm{Spec}_{\mathbf{1}^\perp}(R) \end{aligned} \tag{55}$$

In the case that $\mathcal{S}$ is finite, this simply says that $\mathrm{Gap}(R)$ is one minus the second-largest eigenvalue of R.

The preceding paragraphs are very general. For the theorem presently under consideration, consider a probability measure ψ on the state space $\mathcal{S}$, and let $\mathcal{P}(x, dy)$ be the transition kernel of a Markov chain that is reversible with respect to ψ. Suppose further that the state space is partitioned into m disjoint pieces:

$$\mathcal{S} \,=\, \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_m \,. \tag{56}$$

For each $i = 1, \ldots, m$, define $\mathcal{P}_i$, the restriction of $\mathcal{P}$ to $\mathcal{S}_i$, by rejecting jumps that leave $\mathcal{S}_i$:

$$\mathcal{P}_i(x, B) \,=\, \mathcal{P}(x, B) + 1_{\{x \in B\}} \mathcal{P}(x, \mathcal{S} \setminus \mathcal{S}_i) \quad \text{ for } x \in \mathcal{S}_i, B \subset \mathcal{S}_i. \tag{57}$$

Also define the transition kernel $\overline{\mathcal{P}}$ on $\mathcal{S}$ which suppresses all jumps between different pieces:

$$\overline{\mathcal{P}}(x, A) \,=\, \mathcal{P}_i(x, A \cap \mathcal{S}_i) \quad \text{ if } x \in \mathcal{S}_i \text{ and } A \subset \mathcal{S}. \tag{58}$$

Let ψ_i be the normalized restriction of ψ to $\mathcal{S}_i$:

$$\psi_i(A) \,=\, \frac{\psi(A \cap \mathcal{S}_i)}{b_i}, \quad \text{ where } \quad b_i \,=\, \psi(\mathcal{S}_i). \tag{59}$$

This defines ψ_i to be a measure on $\mathcal{S}$ whose support is in $\mathcal{S}_i$. However, when discussing $\mathcal{P}_i$, we shall want to interpret ψ_i as a measure that is defined on $\mathcal{S}_i$ only. We shall not bother to introduce a different notation for this. Similarly, when discussing $\mathcal{P}_i$ and a function f that is defined on all of $\mathcal{S}$ (e.g. in (ii) below), we shall really mean the new function obtained from f by restricting its domain to $\mathcal{S}_i$.

The following observations are easy to check:
(i) $\psi \,=\, \sum_{i=1}^m b_i \psi_i$;
(ii) $(f, \overline{\mathcal{P}} g)_\psi = \sum_{i=1}^m b_i (f, \mathcal{P}_i g)_{\psi_i}$;
(iii) $\mathcal{P}_i$ is reversible with respect to ψ_i (on the state space $\mathcal{S}_i$); and
(iv) $\mathcal{P}$ is reversible with respect to ψ.

We shall write Π for the projection operator Π_ψ, defined in (52):

$$(\Pi f)(x) \equiv (\Pi f)(x) = \int_{\mathcal{S}} f(y)\,\psi(dy) \quad \text{for all } x \in \mathcal{S}. \tag{60}$$

Similarly, define the operator which projects onto functions that are constant in each piece:

$$(\overline{\Pi} f)(x) = \int_{\mathcal{S}_i} f(y)\,\psi_i(dy) \quad \text{if } x \in \mathcal{S}_i. \tag{61}$$

Let $V_{\mathcal{S}}$ be the vector space of functions on $\mathcal{S}$ that are constant within each $\mathcal{S}_i$. Then $\overline{\Pi}$ is the orthogonal projection onto $V_{\mathcal{S}}$ in $L^2(\psi)$.

Let $\mathcal{Q}(x, dy)$ be another transition kernel that is reversible with respect to ψ. Then let $\overline{\mathcal{Q}}$ be the following "aggregated transition matrix":

$$\overline{\mathcal{Q}}(i,j) = \frac{1}{b_i} \int_{y \in \mathcal{S}_j} \int_{x \in \mathcal{S}_i} \psi(dx)\mathcal{Q}(x, dy) \quad (i,j = 1, \ldots, m). \tag{62}$$

Observe that

$$b_i \overline{\mathcal{Q}}(i,j) = b_j \overline{\mathcal{Q}}(j,i); \tag{63}$$

that is, if we view the vector $b = (b_1, \ldots, b_m)$ as a probability measure on $\{1, \ldots, m\}$, then $\overline{\mathcal{Q}}$ is reversible with respect to b.

Theorem A.1 (Caracciolo–Pelissetto–Sokal) *Assume that $\mathcal{Q}$ is nonnegative definite. Let $\mathcal{Q}^{1/2}$ be the nonnegative square root of $\mathcal{Q}$. Then*

$$Gap(\mathcal{Q}^{1/2}\mathcal{P}\mathcal{Q}^{1/2}) \geq Gap(\overline{\mathcal{Q}}) \min_{i=1,\ldots,m} Gap(\mathcal{P}_i). \tag{64}$$

Given this theorem, we deduce Theorem 2.1 directly, as follows.

Proof of Theorem 2.1: Let $\mathcal{S}$, $\mathcal{S}_i$, $\mathcal{P}$, ψ, and $\mathcal{Q}$ in this appendix be the objects of the same names of Section 2. With this choice, we observe that $\mathcal{P}_i$ and b_i of this appendix respectively correspond to $\mathcal{T}_i$ and c_i of Section 2, and the measure $\psi_i(dx)$ on $\mathcal{S}_i$ in this appendix corresponds to the measure $\psi_i(dx, j) = \delta_i(j)\phi_i(x)\lambda(dx)$ on $\mathcal{S}_i = \Omega \times \{i\}$ in Section 2. With these definitions, the operators $\overline{\mathcal{Q}}$ defined by equations (19) and (62) are the same. Finally, we observed in Section 2 that $\mathcal{Q}$ is reversible and $\mathcal{Q}^2 = \mathcal{Q}$, so we conclude that $\mathcal{Q}$ is positive definite and $\mathcal{Q}^{1/2} = \mathcal{Q}$. This completes the translation between Theorems 2.1 and A.1. □

Before we undertake the proof of Theorem A.1, we record two lemmas.

Lemma A.2 *Let A and B be operators on a Hilbert space, and let c be a nonzero complex number. Then c is in the spectrum of AB if and only if c is in the spectrum of BA.*

Proof: This is Problem 76 of Halmos (1982) (solved on page 224). □

Corollary A.3 *Let A and B be transition kernels on the state space $\mathcal{S}$, such that AB and BA are both nonnegative definite and reversible with respect to the probability measure ρ. Then $Gap(AB) = Gap(BA)$.*

Proof: Lemma A.2 shows that $(0,\infty) \cap \mathrm{Spec}_{\mathbf{1}^\perp}(AB) = (0,\infty) \cap \mathrm{Spec}_{\mathbf{1}^\perp}(BA)$. Therefore $\sup \mathrm{Spec}_{\mathbf{1}^\perp}(AB)$ can differ from $\sup \mathrm{Spec}_{\mathbf{1}^\perp}(BA)$ only if both of these numbers are nonpositive. But the spectrum of a reversible nonnegative definite operator is a nonempty subset of $[0,\infty)$; so if $\sup \mathrm{Spec}_{\mathbf{1}^\perp}(AB)$ and $\sup \mathrm{Spec}_{\mathbf{1}^\perp}(BA)$ are nonpositive, then they must both equal 0. Hence these two sups must be equal. The Corollary now follows from (55). □

Remark: If $\mathcal{S}$ is finite, then we can omit the assumption about nonnegative definiteness in Corollary A.3. This is because in finite dimensions it is well-known that AB and BA have the same spectrum, including multiplicities of all eigenvalues.

Proof of Theorem A.1: Let $G_* = \min_{i=1,\ldots,m} \mathrm{Gap}(\mathcal{P}_i)$. Then, for every i,

$$(f, (I - \mathcal{P}_i)f)_{\psi_i} \;\geq\; G_*(f, (I - \overline{\Pi})f)_{\psi_i} \quad \text{for every } f \in L^2(\psi_i). \tag{65}$$

Multiplying this inequality by b_i and summing over i gives

$$(f, (I - \overline{\mathcal{P}})f)_{\psi} \;\geq\; G_*(f, (I - \overline{\Pi})f)_{\psi} \quad \text{for every } f \in L^2(\psi). \tag{66}$$

Since $\mathcal{P}(x, dy) \geq \overline{\mathcal{P}}(x, dy)$ whenever $x \neq y$, we have

$$\begin{aligned}
(f, (I - \mathcal{P})f)_{\psi} &= \frac{1}{2} \int\int |f(x) - f(y)|^2 \psi(dx) \mathcal{P}(x, dy) \\
&\geq \frac{1}{2} \int\int |f(x) - f(y)|^2 \psi(dx) \overline{\mathcal{P}}(x, dy) \\
&= (f, (I - \overline{\mathcal{P}})f)_{\psi} \quad \text{for every } f \in L^2(\psi).
\end{aligned} \tag{67}$$

By Corollary A.3 with $A = \overline{\Pi}\mathcal{Q}^{1/2}$ and $B = \mathcal{Q}^{1/2}\overline{\Pi}$, we find that

$$\mathrm{Gap}(\overline{\Pi}\mathcal{Q}\overline{\Pi}) = \mathrm{Gap}(\mathcal{Q}^{1/2}\overline{\Pi}^2\mathcal{Q}^{1/2}). \tag{68}$$

It is straightforward to see that the restriction of the operator $\overline{\Pi}\mathcal{Q}\overline{\Pi}$ to the m-dimensional vector space $V_{\mathcal{S}}$ (defined above) corresponds to the matrix $\overline{\mathcal{Q}}$. Hence the eigenvalues of $\overline{\Pi}\mathcal{Q}\overline{\Pi}$ on $V_{\mathcal{S}}$ are exactly the same as those of $\overline{\mathcal{Q}}$. In particular,

$$\mathrm{Gap}(\overline{\Pi}\mathcal{Q}\overline{\Pi}) = \mathrm{Gap}(\overline{\mathcal{Q}}). \tag{69}$$

Combining equations (68) and (69) and using $\overline{\Pi}^2 = \overline{\Pi}$, we conclude that

$$\mathrm{Gap}(\mathcal{Q}^{1/2}\overline{\Pi}\mathcal{Q}^{1/2}) \;=\; \mathrm{Gap}(\overline{\mathcal{Q}})\,. \tag{70}$$

Putting the pieces together, we find that for every f in $L^2(\psi)$,

$$\begin{aligned}
&(f,(I-\mathcal{Q}^{1/2}\mathcal{P}\mathcal{Q}^{1/2})f)_\psi \\
&\quad = \quad (f,(I-\mathcal{Q})f)_\psi + (\mathcal{Q}^{1/2}f,(I-\mathcal{P})\mathcal{Q}^{1/2}f)_\psi \\
&\quad \geq \quad (f,(I-\mathcal{Q})f)_\psi + G_*(\mathcal{Q}^{1/2}f,(I-\overline{\Pi})\mathcal{Q}^{1/2}f)_\psi \qquad \text{(by (67) and (66))} \\
&\quad = \quad (1-G_*)(f,(I-\mathcal{Q})f)_\psi + G_*(f,(I-\mathcal{Q}^{1/2}\overline{\Pi}\mathcal{Q}^{1/2})f)_\psi \\
&\quad \geq \quad 0 + G_*\mathrm{Gap}(\overline{\mathcal{Q}})(f,(I-\Pi)f)_\psi \qquad \text{(by (70))}
\end{aligned}$$

The Theorem follows. □

B The Madras-Piccioni Result

This appendix contains the proof of Proposition 3.2.

The spectral gap of the chain $\mathcal{Q}\mathcal{P}\mathcal{Q}$ is given by the following inf over all nonconstant functions f on $\Omega\times\{1,\ldots,m\}$ that are in $L^2(\psi)$:

$$\mathrm{Gap}(\mathcal{Q}\mathcal{P}\mathcal{Q}) \;=\; \inf \frac{\int_\Omega\int_\Omega \sum_i\sum_j \,|f(x,i)-f(y,j)|^2\,\psi(dx,i)(\mathcal{Q}\mathcal{P}\mathcal{Q})((x,i),(dy,j))}{\int_\Omega\int_\Omega \sum_i\sum_j \,|f(x,i)-f(y,j)|^2\,\psi(dx,i)\psi(dy,j)} \tag{71}$$

We obtain an upper bound on $\mathrm{Gap}(\mathcal{Q}\mathcal{P}\mathcal{Q})$ by restricting the inf to functions that do not depend on i, that is functions f of the form $f(x,i)=g(x)$. Then the numerator of (71) equals

$$\begin{aligned}
&\int_\Omega\int_\Omega \sum_i\sum_j \,|g(x)-g(y)|^2\,\psi(dx,i)(\mathcal{Q}\mathcal{P}\mathcal{Q})((x,i),(dy,j)) \\
&\quad = \quad \int_\Omega\int_\Omega \,|g(x)-g(y)|^2 \sum_i c_i\phi_i(x)\lambda(dx) \quad \times \\
&\qquad\qquad\qquad \sum_k \frac{c_k\phi_k(x)}{\kappa(x)} R(x,dy)\min\left\{1,\frac{\phi_k(y)\rho(x)}{\phi_k(x)\rho(y)}\right\} \\
&\quad \leq \quad \int_\Omega\int_\Omega \,|g(x)-g(y)|^2 R(x,dy)\min\left\{\sum_k c_k\phi_k(x),\sum_k c_k\frac{\phi_k(y)\rho(x)}{\rho(y)}\right\}\lambda(dx) \\
&\quad = \quad \int_\Omega\int_\Omega \,|g(x)-g(y)|^2 R^{[\kappa]}(x,dy)\kappa(x)\,\lambda(dx).
\end{aligned}$$

Also, the denominator of (71) equals

$$\int_\Omega\int_\Omega \,|g(x)-g(y)|^2\kappa(x)\,\lambda(dy)\,\kappa(y)\,\lambda(dy),$$

so we have

$$\mathrm{Gap}(\mathcal{Q}\mathcal{P}\mathcal{Q}) \leq \inf \frac{\int_\Omega\int_\Omega \,|g(x)-g(y)|^2 R^{[\kappa]}(x,dy)\kappa(x)\,\lambda(dx)}{\int_\Omega\int_\Omega \,|g(x)-g(y)|^2\kappa(x)\,\lambda(dy)\,\kappa(y)\,\lambda(dy)}$$

where the inf is over all nonconstant g in $L^2(\kappa)$. Since the right hand side of the last inequality is $\mathrm{Gap}(R^{[\kappa]})$, the proposition follows. □

References

[1] Borgs, C., Chayes, J.T., Frieze, A., Kim, J.H., Tetali, P., Vigoda, E. and Vu, V.H. (1999). Torpid mixing of some MCMC algorithms in statistical physics. In *Proc. 40th IEEE Symposium on Foundations of Computer Science*, 218–229.

[2] Broder, A.Z. (1986). How hard is it to marry at random? (On the approximation of the permanent). In *Proc. 18th ACM Symposium on Theory of Computing*, 50–58. Erratum in Broder, A.Z. (1988), *Proc. 20th ACM Symposium on Theory of Computing* 551.

[3] Bubley, R. and Dyer, M. (1997). Path coupling, Dobrushin uniqueness, and approximate counting. University of Leeds, School of Computer Science technical report 97.04.

[4] Caracciolo, S., Pelissetto, A., and Sokal, A.D. (1992). Two remarks on simulated tempering. Unpublished draft manuscript.

[5] Cooper, C., Dyer, M., Frieze, A., and Rue, R. (2000). Mixing Properties of the Swendsen-Wang process on the complete graph and narrow grids. *Journal of Mathematical Physics* **41**, 1499–1527.

[6] Diaconis, P. and Saloff-Coste, L. (1993a). Comparison techniques for random walks on finite groups. *Ann. Probab.* **21**, 2131–2156.

[7] Diaconis, P. and Saloff-Coste, L. (1993b). Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.* **3**, 696–730.

[8] Diaconis, P. and Stroock, D. (1993). Geometric bounds for Markov chains. *Ann. Appl. Probab.* **1**, 36–61.

[9] Geyer, C.J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E.M. Keramidas, ed.), 156–163. Interface Foundation, Fairfax Station.

[10] Geyer, C.J. and Thompson, E.A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90**, 909–920.

[11] Halmos, P.R. (1982). *A Hilbert Space Problem Book* (Second Edition). Springer-Verlag, New York.

[12] Jerrum, M.R. and Sinclair, A.J. (1989). Approximating the permanent. *SIAM J. Computing* **18**, 1149–1178.

[13] Luby, M., Randall, D., and Sinclair, A.J. (1995). Markov chain algorithms for planar lattice structures. In *Proc. 36th IEEE Symposium on Foundations of Computer Science*, 150–159.

[14] Luby, M. and Vigoda, E. (1997). Approximately counting up to four. In *Proc. 29th ACM Symposium on Theory of Computing*, 682–687.

[15] Luby, M. and Vigoda, E. (1999). Fast Convergence of the Glauber dynamics for sampling independent sets. *Random Structures and Algorithms* **15**, 229–241.

[16] Madras, N. (1998). Umbrella sampling and simulated tempering. In *Numerical Methods for Polymeric Systems* (S.G. Whittington, ed.), *IMA Volumes in Mathematics and Its Applications* **102**, 19–32. Springer, New York.

[17] Madras, N. and Piccioni, M. (1999). Importance sampling for families of distributions. *Ann. Appl. Probab.* **9**, 1202–1225.

[18] Madras, N. and Randall, D. (1996), Factoring graphs to bound mixing rates. In *Proc. 37th IEEE Symposium on Foundations of Computer Science*, 194–203.

[19] Madras, N. and Slade, G. (1993). *The Self-Avoiding Walk*. Birkhäuser, Boston.

[20] Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19**, 451–458.

[21] Orlandini, E. (1998). Monte Carlo study of polymer systems by Multiple Markov Chain method. In *Numerical Methods for Polymeric Systems* (S.G. Whittington, ed.), *IMA Volumes in Mathematics and Its Applications* **102**, 33–58. Springer, New York.

[22] Randall, D. and Tetali, P. (1998). Analyzing Glauber dynamics by comparison of Markov chains. In *Proceedings of LATIN '98: Theoretical Informatics* (C. Lucchesi and A. Moura, eds.), *Lecture Notes in Computer Science* **1380**, 292–304. Springer, New York.

[23] Roberts, G.O. and Rosenthal, J.S. (1997). Geometric ergodicity and hybrid Markov chains. *Elect. Comm. in Probab.* **2**, 13–25.

[24] Roberts, G.O. and Tweedie, R.L. (2000). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. Preprint.

[25] Sinclair, A.J. (1993). *Randomized algorithms for counting and generating combinatorial structures*. Birkhäuser, Boston.

[26] Sokal, A.D. and Thomas, L.E. (1989). Exponential convergence to equilibrium for a class of random-walk models. *J. Statist. Phys.* **54**, 797–828.

[27] Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8**, 1–9.

[28] Torrie, G.M. and Valleau, J.P. (1977). Nonphysical sampling distributions in Monte Carlo free energy estimation: Umbrella sampling. *J. Comp. Phys.* **23**, 187–199.

[29] Welsh, D.J.A. (1993). *Complexity: Knots, Colourings, and Counting*. London Math. Soc. Lecture Note Series **186**. Cambridge University Press, Cambridge.

[30] Yosida, K. (1980). *Functional Analysis* (Sixth Edition). Springer-Verlag, Berlin.

[31] Zheng, Z. (1999). Analysis of swapping and tempering Monte Carlo algorithms. Ph.D. thesis, York University.