MARKOV CHAIN DECOMPOSITION FOR CONVERGENCE RATE ANALYSIS

BY NEAL MADRAS$^1$ AND DANA RANDALL$^2$

York University and Georgia Institute of Technology

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 gaps 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 $\Omega$, and you want to estimate its spectral gap (loosely, the rate at 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 equilibrates rapidly, then one would expect the entire chain to equilibrate rapidly. This is the spirit of our main results.

Let $\Omega$ 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 $\Omega$” when we really mean “$B$ is a measurable subset of $\Omega$.”

To discuss probability densities on our state space, we need a reference measure $\lambda$ on $\Omega$. (For example, if $\Omega$ is discrete, then $\lambda$ can be counting measure, while if $\Omega = \mathbb{R}^n$, then $\lambda$ can be Lebesgue measure.) If $\rho$ is a probability density function (with respect to $\lambda$) on $\Omega$, then the associated probability of a set $B \subset \Omega$ is denoted

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

[If $\lambda$ 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 $\Omega$ that is reversible with respect to a probability density $\pi$. (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 $\Omega$ such that $\bigcup A_i = \Omega$. (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

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

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 $\pi$ to $A_i$.

Next define

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

and define the “maximum overlap” $\Theta$ of the covering $\{A_1, \ldots, A_m\}$ by

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

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

$$ 1 \leq Z \leq \Theta \leq m. $$

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:

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

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 $\rho$. Let $E_\rho$ denote the expectation with respect to $\rho$:

$$ E_\rho f = \int f(y)\rho(dy). $$

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

$$ \text{Gap}(R) := \inf_{f} \frac{\iint |f(x) - f(y)|^2 \rho(dx)R(x, dy)}{\iint |f(x) - f(y)|^2 \rho(dx)\rho(dy)} $$

where the inf is over all non-constant functions $f$ (i.e., functions that are not constant almost surely with respect to $\rho$) 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 $\rho$. 

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 $1 / \text{Gap}(R)$ iterations.
Alternatively, once a chain is in equilibrium, $k / \text{Gap}(R)$ consecutive observations
of the chain will be “statistically equivalent” to $k$ independent samples from
the equilibrium distribution of the chain. [See, e.g., 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 $\mu_1$ and $\mu_2$ on the state space $\Omega$, 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 $\mu_i$ with respect to $\rho$ (i.e., $f_i$ is the Radon–Nikodym
derivative $d\mu_i / d\rho$). For a probability measure $\mu$ and nonnegative integer $n$, let
$\mu R^n$ 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 $\mu$:

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

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

$$
\| \mu R^n - \rho \|_2 \leq \Gamma^n \| \mu - \rho \|_2.
$$

This says that the sequence of measures $\mu R^n$ converges exponentially rapidly
to the equilibrium measure $\rho$. Moreover, $\Gamma^n \leq \exp(-n \text{Gap}(R))$, with approximate
equality in the usual case that $\text{Gap}(R)$ 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

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

where $X'_0, X'_1, \ldots$ is the Markov chain given by $R$ and started in equilibrium (i.e., $X'_0$ has distribution $\rho$). 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 $\Omega$ 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 $\rho$. A Markov chain would be constructed whose equilibrium distribution is $\rho$, knowing that running the chain for “long enough” would result in a state that was “almost” distributed as $\rho$. 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

$$\text{Gap}(P) \geq \frac{1}{\Omega^2} \text{Gap}(P_H) \left( \min_{i=1,\ldots,m} \text{Gap}(P|_{A_i}) \right).$$

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 $\Omega$ that is reversible with respect to a probability density $\rho$. Let $\zeta$ be another probability density whose support is contained in the support of $\rho$. Then the “Metropolis–Hastings chain for $R$ with respect to $\zeta$” is the new Markov chain whose transition kernel $R^{[\zeta]}$ is defined by

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

$$R^{[\zeta]}(x, \{x\}) = 1 - \int_{\Omega \setminus \{x\}} R^{[\zeta]}(x, dy).$$
[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 $\zeta$. 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 $\zeta$ 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 $\zeta$ (i.e., $P = R^{[\zeta]}$). Also suppose that $\zeta$ can be expressed as a convex combination of a small number of densities $\phi_0, \ldots, \phi_D$ (i.e., $\zeta$ is a “mixture density”). Think of running a Metropolis–Hastings chain for each $\phi_j$, 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 $\phi_j$’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 $\phi_j$’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 $\phi_i$ has substantial overlap with $\phi_{i+1}$ and with $\phi_{i-1}$. If each “piece” is rapidly mixing, then a chain which starts in the $i$th piece will soon move into a region where $\phi_i$ and $\phi_{i+1}$ (or $\phi_{i-1}$) 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 $\Omega$ (with respect to a common reference measure $\lambda$), and let $a_0, \ldots, a_D$ be positive numbers that add up to 1. Define the mixture density

\begin{equation}
\phi_{\text{mix}} := \sum_{j=0}^{D} a_j \phi_j.
\end{equation}

Let $R(x, dy)$ be a Markov chain that is reversible with respect to a probability density $\rho(x)$ on $\Omega$. Let $\text{Gap}_j$ (respectively, $\text{Gap}_{\text{mix}}$) be the spectral gap of the Metropolis–Hastings chain $R^{[\phi_j]}$ (respectively, $R^{[\phi_{\text{mix}}]}$). Finally, assume that neighboring $\phi_j$’s have some “overlap” that is, assume

\begin{equation}
\int \min\{\phi_j(x), \phi_{j+1}(x)\}\lambda(dx) \geq \delta \quad (j = 0, \ldots, D - 1)
\end{equation}

for some $\delta > 0$. Then

\begin{equation}
\text{Gap}_{\text{mix}} \geq \frac{\delta}{2D} \min_{j=0,\ldots,D} a_j \text{Gap}_j.
\end{equation}
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 $\Omega$ be the collection of all independent sets, and let $\Omega_i$ be the collection of all independent sets of cardinality $i$ ($i = 0, 1, \ldots, |V|$). Finally, let $\gamma$ be a positive real number.

The hard core model with parameter $\gamma$ is the probability distribution $h_\gamma$ on the collection of all independent sets defined by

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

where $Z_\gamma$ 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 $\gamma$ controls the number of particles; for example, it is not hard to see that the expected size of $I$ is an increasing function of $\gamma$.

Consider the following Markov chain on $\Omega$, 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 $u_t$; that is, 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 $v_t$; that is, 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 $v_t$; that is, 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 $h_\gamma$. 
The work of Luby and Vigoda (1997, 1999) implies that this chain is rapidly mixing if $\gamma \leq 2/(\Delta - 2)$, where $\Delta$ 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 $\gamma$ is sufficiently large. The problem is that large values of $\gamma$ 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 $\gamma$ is very large, and the modified chain mixes rapidly.

Let $n^* = \lceil |V|/2(\Delta + 1) \rceil$, and let $\Omega^*$ 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 $\Omega^*$ (i.e., if $|I_i| = n^*$ and $u_i = v^*$, then $I_{i+1} = I_i$ with probability one). Then $P^*$ is irreducible on $\Omega^*$, aperiodic, and reversible with respect to the restriction of $h_\gamma$ to $\Omega^*$. 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 be described as follows. Let $A_i = \Omega_i \cup \Omega_{i+1}$, whereby $\Omega^* = \bigcup A_i$ is a decomposition of the state space into overlapping pieces, as required for Theorem 1.1. We consider in turn the restrictions $P^*_i$ to $A_i$, for all $i$, as well as the projection $P^*_{H_i}$. 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 $P^*_i$. 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 $P^*_i$ 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

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

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

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

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

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, 2001)], it is rapidly mixing.

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

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

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 $\Omega^*$ is rapidly mixing, as claimed.

THEOREM 1.5. The Markov chain $P^*$ on $\Omega^*$, 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

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

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 $\gamma$. Since $h_y(\Omega_j) = \gamma^j |\Omega_j|$, it suffices to show that $\gamma |\Omega_{i+1}| \geq |\Omega_i|$ for $i = 0, \ldots, n^*-1$. Fix such an $i$ and let $N(i)$ 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
\[
\mathcal{N}(i) \geq |\Omega_i|(|V| - i(\Delta + 1)).
\]
Combining the above two inequalities gives
\[
|\Omega_i| \leq |\Omega_{i+1}|(i + 1) \leq |\Omega_{i+1}| |V| - n^*(\Delta + 1) \leq |\Omega_{i+1}| |V| - n^*(\Delta + 1) \leq |\Omega_{i+1}| \gamma,
\]
which is what we want 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 \( \Omega^* \) 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 \( \Omega = \bigcup \Omega_i, \) where the pieces \( \Omega_i \) form a partition. Moreover, assume the sizes \( |\Omega_i| \) are unimodal as a function of \( i. \) [For example, let \( \Omega \) be the set of matchings of some underlying graph \( G, \) and \( \Omega_i \) 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 \( \Theta^{-2} \) 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 \( n \times O(1) \times \cdots \times O(1) \) 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 \( \Omega \) 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 \( \Omega \) parametrized by a real number \( \beta \) 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 \( \Omega \). When \( \beta \) 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 \( \beta \) equals 0, the change is always accepted, and the Markov chain is rapidly mixing. When \( \beta \) is close to 0, the chain is also rapidly mixing. However, when \( \beta \) 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 \( \phi_1 \) be the distribution with \( \beta = 0 \), \( \phi_m \) the distribution for a given large \( \beta \), and \( \phi_i \) (\( i = 2, \ldots, m-1 \)) be distributions at equally spaced intermediate values of \( \beta \). For each \( i \), the \( \mathcal{T}_i \) that we shall introduce below will be the Metropolis chain for the corresponding \( \phi_i \). 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 \( \Omega \) 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 \( \phi_i \).

Next we define the “augmented” state space \( \mathcal{S} \) by including the “labels” 1 through \( m \):

\[
\mathcal{S} = \Omega \times \{1, \ldots, m\}.
\]
For each $i = 1, \ldots, m$, define

$$
\delta_i = \Omega \times \{i\} = \{(x, i) : x \in \Omega\}.
$$

Thus $\delta_1, \ldots, \delta_m$ forms a partition of $\delta$. Define the transition probability kernel $\mathcal{P}$ on $\delta$ as follows:

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

Notice that this kernel does not permit transitions from one $\delta_i$ to another.

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

$$
\mathcal{Q}((x, i), (dy, j)) = \delta_x(dy) \frac{c_j \phi_j(x)}{\sum_{i=1}^m c_i \phi_i(x)}.
$$

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 $\psi$ on $\delta$ by

$$
\psi(dx, i) = c_i \phi_i(x) \lambda(dx) \quad ((x, i) \in \delta).
$$

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

$$
\tilde{\phi}(x) := \sum_{i=1}^m c_i \phi_i(x).
$$

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 $\mathcal{Q} \mathcal{P}$ or to $\mathcal{P} \mathcal{Q}$, or to $\mathcal{Q} \mathcal{P} \mathcal{Q}$ (recall that $\mathcal{Q}^2 = \mathcal{Q}$). We shall use the version $\mathcal{Q} \mathcal{P} \mathcal{Q}$, since it is reversible with respect to $\psi$ (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 $\Omega$ 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 $\phi_m$, 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” $\mathcal{Q}$ (an analogue of $P_H$
defined in the Introduction):

$$\mathcal{Q}(i, j) = \frac{1}{c_i} \int_{\Omega} \frac{c_i \phi_i(x) c_j \phi_j(x)}{\sum_{i=1}^{m} c_i \phi_i(x)} \lambda(dx)$$

(19)

$$= c_j \int_{\Omega} \frac{\phi_i(x) \phi_j(x)}{\phi_i(x)} \lambda(dx) \quad (i, j = 1, \ldots, m).$$

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

$$\text{Gap}(\mathcal{Q}, \mathcal{P}) \geq \text{Gap}(\mathcal{Q}) \min_{i=1,\ldots,m} \text{Gap}(\mathcal{T}_i).$$

(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 $\Omega$. Let $\xi$ be a probability measure on $\Omega$.
Then the “Metropolis–Hastings chain for $R$ with respect to $\xi$” is the new Markov
chain whose transition kernel $R^{[\xi]}$ is formally defined by

$$R^{[\xi]}(x, dy) = R(x, dy) \min\left\{1, \frac{\xi(dy) R(y, dx)}{\xi(dx) R(x, dy)}\right\} \quad \text{if } y \neq x,$$

$$R^{[\xi]}(x, \{x\}) = 1 - \int_{\Omega \setminus \{x\}} R^{[\xi]}(x, dy).$$

(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 \( \zeta \) has a density, which we shall also call \( \zeta \), with respect to some reference measure \( \lambda \) (i.e., \( \zeta(dx) = \zeta(x)\lambda(dx) \)). Then:

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

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

- 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 \( \zeta \). 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 \( \rho \) 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 \( \lambda \) on \( \Omega \). Suppose that \( R(x, dy) \) is reversible with respect to the density \( \rho \). Also suppose that there are constants \( a \) and \( b \) such that

\[
a \leq \frac{r_1(x)}{r_2(x)} \leq b
\]

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

\[
\frac{a}{b} \text{ Gap}(R^{[r_2]}) \leq \text{ Gap}(R^{[r_1]}) \leq \frac{b}{a} \text{ Gap}(R^{[r_2]}).
\]

Next we discuss the method known as *Umbrella Sampling*. Consider a probability density \( \kappa \) on \( \Omega \) 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

\[
\kappa(x) = \sum_{i=1}^{m} c_i \phi_i(x).
\]

Now consider a transition kernel \( R(x, dy) \) of a Markov chain on \( \Omega \) that is reversible with respect to a probability density \( \rho \). For each \( i = 1, \ldots, m \), define the transition kernel \( T_i(x, dy) = R^{[\phi_i]} \), the Metropolis–Hastings chain for \( R \) with respect to \( \phi_i \).
In physical applications, the $\phi_i$’s are often natural distributions from which we want to sample, and the mixture $\kappa$ is an artificial “umbrella” distribution. Using the Metropolis–Hastings chain $R^{[\kappa]}$ to sample from $\kappa$ (together with the classical Monte Carlo technique of importance sampling) allows one to sample from all $\phi_i$’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 $\Omega$, $\phi_i$’s, $c_i$’s, and $T_i$’s are used to define the simulated tempering chain $Q\mathcal{P}Q$ of Section 2 on the augmented state space $\Omega \times \{1, \ldots, m\}$. Then

$$\text{Gap}(Q\mathcal{P}Q) \leq \text{Gap}(R^{[\kappa]}).$$

Note that the invariant measure of $Q\mathcal{P}Q$, $\psi$, 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 $\Omega$ 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 $\Omega$ such that $\bigcup A_i = \Omega$. For each $i = 1, \ldots, m$, let $\phi_i$ be the normalized restriction of $\pi$ to $A_i$:

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

Recall

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

Also let

$$c_i = \frac{\pi[A_i]}{Z} \quad (i = 1, \ldots, m),$$

and define

$$\kappa(x) = \sum_{i=1}^{m} c_i \phi_i(x).$$
Then $\kappa$ is a probability density, and

\begin{equation}
\frac{1}{Z} \pi(x) \leq \kappa(x) \leq \Theta \frac{1}{Z} \pi(x).
\end{equation}

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

\begin{equation}
P^{[\pi]} = P.
\end{equation}

Therefore Lemma 3.1 and equation (29) imply that

\begin{equation}
\frac{1}{\Theta} \text{Gap}(P) \leq \text{Gap}(P^{[\kappa]}) \leq \Theta \text{Gap}(P).
\end{equation}

The restriction $P^{[A_i]}$ of $P$ to $A_i$ is the same as the Metropolis–Hastings chain $P^{[\phi_i]}$. Let $R(x, dy) = P^{[\kappa]}(x, dy)$ and $T_i(x, dy) = R^{[\phi_i]}(x, dy)$ ($i = 1, \ldots, m$).

Observe that we can write

\begin{equation}
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\}
\end{equation}

whenever $x \neq y$ and $x, y \in A_i$.

It follows from this and equation (29) that

\begin{equation}
\frac{1}{\Theta} P^{[\phi_i]}(x, dy) \leq T_i(x, dy) \leq P^{[\phi_i]}(x, dy)
\end{equation}

whenever $x \neq y$.

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

\begin{equation}
\frac{1}{\Theta} \text{Gap}(P^{[\phi_i]}) \leq \text{Gap}(T_i) \leq \text{Gap}(P^{[\phi_i]}).
\end{equation}

The aggregated transition matrix of Section 2 is

\begin{equation}
\overline{Q}(i, j) = c_j \int_{\Omega} \frac{\phi_i(x) \phi_j(x)}{\kappa(x)} \lambda(dx)
\end{equation}

\begin{equation}
= c_j \int_{A_i \cap A_j} \frac{\pi(x)^2}{\kappa(x) \pi[A_i] \pi[A_j]} \lambda(dx)
\end{equation}

\begin{equation}
= \frac{1}{\pi[A_i]} \int_{A_i \cap A_j} \frac{\pi(x)^2}{Z \kappa(x)} \lambda(dx).
\end{equation}

Recalling the definition of $P_H$ [equation (5)], and using equations (29) and (36), we see that

\begin{equation}
P_H(a_i, a_j) \leq \overline{Q}(i, j) \leq \Theta P_H(a_i, a_j)
\end{equation}

for $i \neq j$. 

Since both $P_H$ and $\overline{Q}$ 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

$$\text{Gap}(P_H) \leq \text{Gap}(\overline{Q}) \leq \Theta \text{Gap}(P_H).$$

Finally we put the pieces together:

$$\text{Gap}(P) \geq \frac{1}{\Theta} \text{Gap}(P^{[c]}) \quad \text{[by equation (31)]}$$

$$\geq \frac{1}{\Theta} \text{Gap}(Q \cdot P \cdot Q) \quad \text{(by Proposition 3.2 with } R = P \text{ and } \rho = \pi)$$

$$\geq \frac{1}{\Theta} \text{Gap}(\overline{Q}) \min_{i=1,\ldots,m} \text{Gap}(T_i) \quad \text{(by Theorem 2.1)}$$

$$\geq \frac{1}{\Theta^2} \text{Gap}(P_H) \min_{i=1,\ldots,m} \text{Gap}(P^{[\phi_i]}) \quad \text{[by equations (38) and (35)]}$$

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 $\lambda$ on $\Omega$.

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 $E_{\text{mix}}$ instead of $E_{\phi_j}$ and $E_{\text{mix}}$, and similarly for $V_j$ and $V_{\text{mix}}$.

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

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

[and define $B_{\text{mix}}(f)$ analogously]. Then the spectral gap of the Metropolis–Hastings chain for $R$ with respect to $\phi_j$ is given by

$$\text{Gap}_j = \inf_f \frac{B_j(f)}{2V_j(f)}.$$

Since

$$\min \left\{ \frac{\phi_{\text{mix}}(x)}{\rho(x)}, \frac{\phi_{\text{mix}}(y)}{\rho(y)} \right\} \geq \sum_{j=0}^D a_j \min \left\{ \frac{\phi_j(x)}{\rho(x)}, \frac{\phi_j(y)}{\rho(y)} \right\},$$
it follows that for every \( f \),

\[
B_{\text{mix}}(f) = \int \left( (f(x) - f(y))^2 R(x, dy) \right) \min \left\{ \frac{\phi_{\text{mix}}(x)}{\rho(x)}, \frac{\phi_{\text{mix}}(y)}{\rho(y)} \right\} \times \rho(x) \lambda(dx)
\]

\[
\geq \sum_j a_j B_j(f) \geq \sum_j a_j \text{Gap}_j 2V_j(f)
\]

\[
\geq 2 \min \{ a_i \text{Gap}_i \} \sum_{j=0}^D V_j(f).
\]

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

\[
2V_{\text{mix}}(f) \leq \frac{4D}{\delta} \sum_{j=0}^D V_j(f).
\]

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

\[
C_{ij}(f) = \int \left( (f(x) - f(y))^2 \phi_i(x) \phi_j(y) \lambda(dx) \lambda(dy) \right).
\]

Then

\[
C_{jj}(f) = 2V_j(f)
\]

and

\[
2V_{\text{mix}}(f) = \int \left( (f(x) - f(y))^2 \left( \sum_i a_i \phi_i(x) \right) \left( \sum_j a_j \phi_j(y) \right) \lambda(dx) \lambda(dy) \right)
\]

\[
= \sum_{i,j} a_i a_j C_{ij}(f).
\]

In particular, we have

\[
2V_{\text{mix}}(f) \leq \max_{i,j} C_{ij}(f).
\]

Fix \( j \). By the overlap condition (11), there exist probability densities \( \eta, \tau \) and \( \psi \) such that

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

Then

\[
2V_j(f) = \int \left( (f(x) - f(y))^2 \right)
\]

\[
\times \left( \delta \eta(x) + (1 - \delta) \tau(x) \right) \left( \delta \eta(y) + (1 - \delta) \tau(y) \right) \lambda(dx) \lambda(dy)
\]

\[
= 2\delta^2 V_\eta(f) + 2(1 - \delta)^2 V_\tau(f)
\]

\[
+ 2\delta(1 - \delta) \int \left( (f(x) - f(y))^2 \eta(x) \tau(y) \lambda(dx) \lambda(dy) \right)
\]

\[
= 2\delta^2 V_{\eta}(f) + 2(1 - \delta)^2 V_{\tau}(f) + 2\delta(1 - \delta) \int \left( (f(x) - f(y))^2 \eta(x) \tau(y) \lambda(dx) \lambda(dy) \right)
\]
and

\[
C_{j,j+1}(f) = \iint (f(x) - f(y))^2 (\delta \eta(x) + (1 - \delta) \tau(x)) \times (\delta \eta(y) + (1 - \delta) \psi(y)) \lambda(dx) \lambda(dy)
\]

\[
= 2 \delta^2 V_\eta(f) + (1 - \delta)^2 \iint (f(x) - f(y))^2 \tau(x) \psi(y) \lambda(dx) \lambda(dy)
\]

\[
+ \delta (1 - \delta) \iint (f(x) - f(y))^2 (\eta(x) \psi(y) + \eta(y) \tau(x)) \times \lambda(dx) \lambda(dy).
\]

From (44) we find

\[
\iint (f(x) - f(y))^2 \eta(x) \tau(y) \lambda(dx) \lambda(dy) \leq (V_j(f) - \delta^2 V_\eta(f))/(\delta(1 - \delta)).
\]

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

\[
\iint (f(x) - f(y))^2 \tau(x) \psi(y) \lambda(dx) \lambda(dy)
\]

\[
= \iiint (f(x) - f(z) + f(z) - f(y))^2 \tau(x) \psi(y) \eta(z) \lambda(dx) \lambda(dy) \lambda(dz)
\]

\[
\leq 2 \iiint ((f(x) - f(z))^2 + (f(z) - f(y))^2) \tau(x) \psi(y) \eta(z) \lambda(dx) \lambda(dy) \lambda(dz)
\]

\[
= 2 \iint (f(x) - f(z))^2 \tau(x) \eta(z) \lambda(dx) \lambda(dy)
\]

\[
+ 2 \iint (f(z) - f(y))^2 \psi(y) \eta(z) \lambda(dy) \lambda(dz).
\]

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

\[
C_{j,j+1}(f) \leq 2 \delta^2 V_\eta(f) + (2(1 - \delta)^2 + \delta(1 - \delta))
\]

\[
\times \left( \iint (f(x) - f(z))^2 \tau(x) \eta(z) \lambda(dx) \lambda(dz)
\right.
\]

\[
+ \iint (f(z) - f(y))^2 \psi(y) \eta(z) \lambda(dy) \lambda(dz) \right)
\]

\[
= 2 \delta^2 V_\eta(f) + (2 - \delta)(1 - \delta)
\]

\[
\times \left( \iint (f(x) - f(y))^2 \eta(x) \tau(y) \lambda(dx) \lambda(dy)
\right.
\]

\[
+ \iint (f(x) - f(y))^2 \eta(x) \psi(y) \lambda(dx) \lambda(dy) \right)
\]

\[
\leq 2 \delta^2 V_\eta(f) + (2 - \delta)(V_j(f) - \delta^2 V_\eta(f) + V_{j+1}(f) - \delta^2 V_\eta(f))/\delta
\]

\[
\leq \frac{2 - \delta}{\delta} (V_j(f) + V_{j+1}(f)).
\]
Let \( X^{(0)}, \ldots, X^{(D)} \) be independent random variables, with \( X^{(i)} \) having density \( \phi_i \). Then for \( i < j \),
\[
C_{ij}(f) = E \left( (f(X^{(i)}) - f(X^{(j)}))^2 \right)
\]
\[
= E \left( \left( \sum_{k=i}^{j-1} f(X^{(k)}) - f(X^{(k+1)}) \right)^2 \right)
\]
\[
\leq E \left( (j - i) \sum_{k=i}^{j-1} (f(X^{(k)}) - f(X^{(k+1)}))^2 \right)
\]
\[
= (j - i) \sum_{k=i}^{j-1} C_{k,k+1}(f)
\]
(\text{where we used the Schwarz inequality in the third line}). Therefore, applying (47),
we see that for all \( i \neq j \),
\[
C_{ij}(f) \leq D \sum_{k=0}^{D-1} C_{k,k+1}(f) \leq \frac{2(2 - \delta)D}{\delta} \sum_{l=0}^{D} V_l(f).
\]
(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.

\textbf{Remark.} An inspection of the final paragraph of the proof shows that 
Theorem 1.2 can be generalized to the case that the overlapping \( \phi_j \)'s are not 
linearly arranged. More precisely, we can replace the last two sentences in 
the statement of the theorem with the following: \textit{Fix} \( \delta > 0 \), \textit{and consider a graph 
whose vertices are} \( 0, 1, \ldots, D \), \textit{with an edge joining} \( i \) \textit{to} \( j \) \textit{if and only if}
\[
\int \min[\phi_i(x), \phi_j(x)] \lambda(dx) \geq \delta.
\]

\textit{Let} \( M \) \textit{be the diameter of this graph, that is,}
\[
M = \max_{i,j} \{ \text{minimum number of edges in a path from} \ i \ \text{to} \ j \}.
\]

\textit{Then}
\[
\text{Gap}_{\text{mixture}} \geq \frac{\delta}{2M} \min_{j=0, \ldots, D} a_j \text{Gap}_j.
\]

\textbf{Appendix}

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

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

the inner product on $L^2(\rho)$, the Hilbert space all functions that are square-integrable with respect to $\rho$. (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 $\rho$. That is,

$$
\rho(dx)R(x, dy) = \rho(dy)R(y, dx),
$$

or, more formally,

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

where we write

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

Let $\Pi_\rho$ be the projection operator that sends each function to the constant function identical to its mean:

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

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

$$
\text{Gap}(R) = \inf \frac{(f, (I - R)f)_\rho}{(f, (I - \Pi_\rho f)_\rho}
$$

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

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 $\rho$.

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

$$
1^\perp := \{ f \in L^2(\rho): (f, 1)_\rho = 0 \} := \{ f \in L^2(\rho): \int_{\mathcal{S}} f(x)\rho(dx) = 0 \}.
$$
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 \(\text{Spec}_{\mathbf{1}^\perp}(T)\) to denote the spectrum of the operator \(T\) on \(\mathbf{1}^\perp\). For the reversible probability transition operator \(R\), it is well known that \(\text{Spec}_{\mathbf{1}^\perp}(R)\) is a subset of the real interval \([-1, 1]\).

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

\[
\text{Gap}(R) = \inf_{f \in \mathbf{1}^\perp} \frac{(f, (I - R)f)_{\rho}}{(f, f)_{\rho}}
\]

(55)

\[
= \inf \text{Spec}_{\mathbf{1}^\perp}(I - R) \quad [\text{Yosida (1980), page 320}]
\]

\[
= 1 - \sup \text{Spec}_{\mathbf{1}^\perp}(R).
\]

In the case that \(\delta\) is finite, this simply says that \(\text{Gap}(R)\) 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 \(\psi\) on the state space \(\delta\), and let \(\mathcal{P}(x, dy)\) be the transition kernel of a Markov chain that is reversible with respect to \(\psi\). Suppose further that the state space is partitioned into \(m\) disjoint pieces:

(56)

\[
\delta = \delta_1 \cup \cdots \cup \delta_m.
\]

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

(57)

\[
\mathcal{P}_i(x, B) = \mathcal{P}(x, B) + 1_{\{x \in B\}} \mathcal{P}(x, \delta \setminus \delta_i) \quad \text{for } x \in \delta_i, B \subset \delta_i.
\]

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

(58)

\[
\overline{\mathcal{P}}(x, A) = \mathcal{P}_i(x, A \cap \delta_i) \quad \text{if } x \in \delta_i \text{ and } A \subset \delta.
\]

Let \(\psi_i\) be the normalized restriction of \(\psi\) to \(\delta_i\):

(59)

\[
\psi_i(A) = \frac{\psi(A \cap \delta_i)}{b_i} \quad \text{where } b_i = \psi(\delta_i).
\]

This defines \(\psi_i\) to be a measure on \(\delta\) whose support is in \(\delta_i\). However, when discussing \(\mathcal{P}_i\), we shall want to interpret \(\psi_i\) as a measure that is defined on \(\delta_i\) 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 \(\delta\) [e.g., in (ii) below], we shall really mean the new function obtained from \(f\) by restricting its domain to \(\delta_i\).

The following observations are easy to check:

(i) \(\psi = \sum_{i=1}^m b_i \psi_i\);
(ii) \((f, \mathcal{P} g)_{\psi} = \sum_{i=1}^m b_i (f, \mathcal{P}_i g)_{\psi_i}\);
(iii) \(\mathcal{P}_i\) is reversible with respect to \(\psi_i\) (on the state space \(\delta_i\)); and
(iv) \(\mathcal{P}\) is reversible with respect to \(\psi\).
We shall write $\Pi$ for the projection operator $\Pi_\psi$, defined in (52):

\[(60) \quad (\Pi f)(x) \equiv (\Pi f)(x) = \int_\delta f(y) \psi(dy) \quad \text{for all } x \in \delta.\]

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

\[(61) \quad (\overline{\Pi} f)(x) = \int_{\delta_i} f(y) \psi_i(dy) \quad \text{if } x \in \delta_i.\]

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

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

\[(62) \quad \overline{Q}(i, j) = \frac{1}{b_i} \int_{y \in \delta} \int_{x \in \delta_i} \psi(dx) Q(x, dy) \quad (i, j = 1, \ldots, m).\]

Observe that

\[(63) \quad b_i \overline{Q}(i, j) = b_j \overline{Q}(j, i);\]

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

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

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

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

**Proof of Theorem 2.1.** Let $\delta$, $\delta_i$, $\mathcal{P}$, $\psi$, and $Q$ 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 $T_i$ and $c_i$ of Section 2, and the measure $\psi_i(dx)$ on $\delta_i$ in this appendix corresponds to the measure $\psi_i(dx, j) = \delta_i(j) \phi_i(x) \lambda(dx)$ on $\delta_i = \Omega \times \{i\}$ in Section 2. With these definitions, the operators $\overline{Q}$ defined by equations (19) and (62) are the same. Finally, we observed in Section 2 that $Q$ is reversible and $Q^2 = Q$, so we conclude that $Q$ is positive definite and $Q^{1/2} = Q$. This completes the translation between Theorems 2.1 and A.1. \(\square\)

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 non-zero 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 $\rho$. Then $\text{Gap}(AB) = \text{Gap}(BA)$. 

PROOF. Lemma A.2 shows that $(0, \infty) \cap \text{Spec}_{\perp}(AB) = (0, \infty) \cap \text{Spec}_{\perp}(BA)$. Therefore $\sup \text{Spec}_{\perp}(AB)$ can differ from $\sup \text{Spec}_{\perp}(BA)$ 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 $\sup \text{Spec}_{\perp}(AB)$ and $\sup \text{Spec}_{\perp}(BA)$ are nonpositive, then they must both equal 0. Hence these two supers 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 $G_* = \min_{i=1, \ldots, m} \text{Gap}(\mathcal{P}_i)$. Then, for every $i$,

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

Multiplying this inequality by $b_i$ and summing over $i$ gives

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

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

$$ (f, (I - \mathcal{P})f)_{\psi} = \frac{1}{2} \iint |f(x) - f(y)|^2 \psi(dx, dy) \mathcal{P}(x, dy) $$

$$ \geq \frac{1}{2} \iint |f(x) - f(y)|^2 \psi(dx, dy) \overline{\mathcal{P}}(x, dy) $$

$$ = (f, (I - \overline{\mathcal{P}})f)_{\psi} \quad \text{for every } f \in L^2(\psi). $$

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

$$ \text{Gap}(\overline{\mathcal{P}}\mathcal{Q}\overline{\mathcal{P}}) = \text{Gap}(\mathcal{Q}^{1/2}\overline{\mathcal{P}}\overline{\mathcal{P}}\mathcal{Q}^{1/2}). $$

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

$$ \text{Gap}(\overline{\mathcal{P}}\mathcal{Q}\overline{\mathcal{P}}) = \text{Gap}(\overline{\mathcal{Q}}). $$
Combining equations (68) and (69) and using $\bar{\Pi}^2 = \bar{\Pi}$, we conclude that

$$\text{Gap}(\mathcal{Q}^{1/2}\bar{\Pi}\mathcal{Q}^{1/2}) = \text{Gap}(\mathcal{Q}).$$

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

$$\left(f, (I - \mathcal{Q}^{1/2}\mathcal{P}\mathcal{Q}^{1/2}) f \right)_\psi$$

$$= \left(f, (I - \mathcal{Q}) f \right)_\psi + \left(\mathcal{Q}^{1/2} f, (I - \mathcal{P})\mathcal{Q}^{1/2} f \right)_\psi$$

$$\geq \left(f, (I - \mathcal{Q}) f \right)_\psi + G_*(\mathcal{Q}^{1/2} f, (I - \bar{\Pi})\mathcal{Q}^{1/2} f) \quad \text{[by (68) and (66)]}$$

$$= (1 - G_*)\left(f, (I - \mathcal{Q}) f \right)_\psi + G_*(f, (I - \mathcal{Q}^{1/2}\bar{\Pi}\mathcal{Q}^{1/2}) f)_\psi$$

$$\geq 0 + G_\ast \text{Gap}(\mathcal{Q}) (f, (I - \Pi) f)_{\psi} \quad \text{[by (70)]}.$$

The theorem follows. \qed

**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 $L^2(\psi)$:

$$\text{Gap}(\mathcal{Q}\mathcal{P}\mathcal{Q})$$

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

We obtain an upper bound on $\text{Gap}(\mathcal{Q}\mathcal{P}\mathcal{Q})$ 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

$$\int_\Omega \int_\Omega \sum_i \sum_j |g(x) - g(y)|^2 \psi(dx, i)(\mathcal{Q}\mathcal{P}\mathcal{Q})(\psi, i, (dy, j))$$

$$= \int_\Omega \int_\Omega |g(x) - g(y)|^2 \sum_i c_i \phi_i(x) \lambda(dx)$$

$$\times \sum_k c_k \phi_k(x) \kappa(x) R(x, dy) \min \left\{ 1, \frac{\phi_k(y) \rho(x)}{\phi_k(x) \rho(y)} \right\}$$

$$\leq \int_\Omega \int_\Omega |g(x) - g(y)|^2 R(x, dy) \min \left\{ \sum_k c_k \phi_k(x), \sum_k c_k \frac{\phi_k(y) \rho(x)}{\rho(y)} \right\} \lambda(dx)$$

$$= \int_\Omega \int_\Omega |g(x) - g(y)|^2 R^{[\kappa]}(x, dy) \kappa(x) \lambda(dx).$$
Also, the denominator of (71) equals
\[
\int_{\Omega} \int_{\Omega} |g(x) - g(y)|^2 \kappa(x) \lambda(dy) \kappa(y) \lambda(dx),
\]
so we have
\[
\text{Gap}(Q, P \otimes Q) \leq \inf \frac{\int_{\Omega} \int_{\Omega} |g(x) - g(y)|^2 R^{[\kappa]}(x, dy) \kappa(x) \lambda(dx)}{\int_{\Omega} \int_{\Omega} |g(x) - g(y)|^2 \kappa(x) \lambda(dy) \kappa(y) \lambda(dy)}
\]
where the inf is over all nonconstant g in $L^2(\kappa)$. Since the right hand side of the last inequality is $\text{Gap}(R^{[\kappa]})$, the proposition follows.

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.

REFERENCES


