ANALYSIS OF CONVERGENCE RATES OF SOME GIBBS
SAMPLERS ON CONTINUOUS STATE SPACES

AARON SMITH

1. Abstract

We use a non-Markovian coupling and small modifications of techniques from the
theory of finite Markov chains to analyze some Markov chains on continuous state
spaces. The first is a generalization of a sampler introduced by Randall and Winkler,
the second a Gibbs sampler on narrow contingency tables.

2. Introduction

The problem of sampling from a given distribution on high-dimensional continuous
spaces arises in the computational sciences and Bayesian statistics, and a frequently-
used solution is Markov chain Monte Carlo (MCMC); see [14] for many examples.
Because MCMC methods produce good samples only after a lengthy mixing period,
a long-standing mathematical question is to analyze the mixing times of the MCMC
algorithms which are in common use. Although there are many mixing conditions,
the most commonly used is called the mixing time, and is based on the total variation
distance:

For measures \( \nu, \mu \) with common measurable \( \sigma \)-algebra \( A \), the total variation dis-
tance between \( \mu \) and \( \nu \) is

\[
||\mu - \nu||_{TV} = \sup_{A \in A} (\mu(A) - \nu(A))
\]

For an ergodic discrete-time Markov chain \( X_t \) with unique stationary distribution
\( \pi \), the mixing time is

\[
\tau(\epsilon) = \inf \{ t : ||L(X_t) - \pi||_{TV} < \epsilon \}
\]

Although most scientific and statistical uses of MCMC methods occur in continuous
state spaces, much of the mathematical mixing analysis has been in the discrete
setting. The methods that have been developed for discrete chains often break down
when used to analyze continuous chains, though there are some efforts, such as [26],
[22], [16], to create general techniques. This paper extends the author’s previous work
in [25] and work of Randall and Winkler [20], and attempts to provide some more

Date: January 17, 2013.
examples of relatively sharp analyses of continuous chains similar to those used to develop the discrete theory.

The first process that we analyze is a Gibbs sampler on the simplex with a very restricted set of allowed moves. Fix a finite group $G$ of size $n$ with symmetric generating set $R$ of size $m$, with $id \notin R$. For unity of notation, label the group elements with the integers from 1 to $n$. We consider the process $X_t[g]$ on the simplex $\Delta_G = \{X \in \mathbb{R}^n | \sum_{g \in G} X[g] = 1; X[g] \geq 0\}$. At each step, choose $g \in G$, $r \in R$ and $\lambda \in [0, 1]$ uniformly, and set

$$X_{t+1}[g] = \lambda(X_t[g] + X_t[gr])$$

(1)

$$X_{t+1}[gr] = (1 - \lambda)(X_t[g] + X_t[gr])$$

For all other $h \in G$ set $X_{t+1}[h] = X_t[h]$. Let $U_G$ be the uniform distribution on $\Delta_G$; this is also the stationary distribution of $X_t$. Also consider a random walk $Z_t$ on $G$, where in each stage we choose $g \in G$ and $r \in R$ uniformly at random and set $Z_{t+1} = gr$ if $Z_t = g$, set $Z_{t+1} = g$ if $Z_t = gr$, and $Z_{t+1} = Z_t$ otherwise. This is the standard simple random walk on the Cayley graph, slowed down by a factor of about $n$. Let $\gamma$ be the spectral gap of the walk $Z_t$, and follow the notation that $L(X)$ denotes the distribution of a random variable $X$.

**Theorem 1** (Convergence Rate for Gibbs Sampler with Geometry). For $T > \frac{8C}{\gamma} \log(n)$, $C > \frac{103}{4}$, and $n$ satisfying $n > \max\left(4096, \frac{C}{3} + \frac{10}{3}, \frac{C}{3} - \frac{13}{12}\right) \log(n)$

$$||L(X_T) - U_G||_{TV} \leq 7n^{1.5 - \frac{C}{\pi}}$$

and conversely for $T < \frac{k}{\gamma}$,

$$||L(X_T) - U_G||_{TV} \geq \frac{1}{2} e^{-k} - 4n^{-\frac{1}{4}}$$

This substantially generalizes [20] and [25], from samplers corresponding to $G = \mathbb{Z}_n$, and $R = \{1, -1\}$ or $R = \mathbb{Z}_n \setminus \{0\}$ respectively, to general Cayley graphs. In addition to being of mathematical interest, this process is an example of a gossip process with some geometry, studied by electrical engineers and sociologists interested in how information propagates through networks; see [23] for a survey.

The proof of the upper bound will use an auxiliary chain similar to that found in [20], a coupling argument improved from [25], and an unusual use of comparison theory from [7]. The proof of the lower bound is elementary.

The next example consists of narrow contingency tables. Beginning with the work of Diaconis and Efron [4] on independence tests, there has been interest in finding efficient ways to sample uniformly from the collection of integer-valued matrices with given row and column sums. A great deal of this effort has been based on Markov chain Monte Carlo methods. While some of the efforts have dealt directly with Markov chains on these integer-valued matrices, much recent success, including [6] [19], has
involved using knowledge of Gibbs samplers on convex sets in $\mathbb{R}^n$ and clever ways to project from the continuous chain to the desired matrices [18].

Unfortunately, while the general bounds are polynomial in the number of entries in the desired matrix, they often have a large degree and leading coefficient; see [15]. In this paper, we find some better bounds for very specific cases. Like the paper [25], this is part of an attempt to make further use of non-Markovian coupling techniques [11] [1] [3] [17] and also to expand the small set of carefully analyzed Gibbs samplers [20] [21] [5] [6].

We consider the following Gibbs sampler $X_{t}[i,j]$ on the space $M_{n} = \{X \in \mathbb{R}^{2n} : \sum_{i=1}^{n}X[i,j] = n \ \forall \ 1 \leq j \leq 2, \sum_{j=1}^{2}X[i,j] = 2 \ \forall \ 1 \leq i \leq n, X[i,j] \geq 0\}$ of nonnegative $n$ by 2 matrices with column sums fixed to be $n$ and row sums fixed to be 2. To make a step of the Gibbs sampler, choose two distinct integers $1 \leq i < j \leq n$ and update the four entries $X_{t+1}[i,1], X_{t+1}[i,2], X_{t+1}[j,1]$ and $X_{t+1}[j,2]$ to be uniform conditional on all other entries of $X_{t}$. Let $U_{n}$ be the uniform distribution on $M_{n}$ inherited from Lebesgue measure. Then we find the following reasonable bound on the mixing time of this sampler:

**Theorem 2** (Convergence Rate for Narrow Matrices). Fix $C > 23$ and set $a = \frac{2}{11}(C + 18.25)$. Then, for $n > \max(4096, \frac{2}{11}C + \frac{75}{11})$ satisfying $\frac{n}{\log(n)} > \frac{6(2a-13)(a-7)}{2a-15}$ and $T > Cn\log(n)$,

$$||\mathcal{L}(X_{T}) - U_{n}||_{TV} \leq 11n^{-\frac{2}{11} + \frac{12}{44}}$$

and conversely for fixed $0 < C < 1$ with $n$ sufficiently large and $T < (1-C)n\log(n)$,

$$||\mathcal{L}(X_{T}) - U_{n}||_{TV} \geq 1 - 2n^{-k}$$

3. **General Strategy and the Partition Process**

Both of our bounds will be obtained using a similar strategy, ultimately built on the classical coupling lemma. We recall that a coupling of Markov chains with transition kernel $K$ is a process $(X_t, Y_t)$ so that marginally both $X_t$ and $Y_t$ are Markov chains with transition kernel $K$. Although we always couple entire paths $\{X_t\}_{t=0}^T$ and $\{Y_t\}_{t=0}^T$, we often use the shorthand notation of saying that we are coupling $X_t$ and $Y_t$. In order to describe a coupling, note that for both walks being studied, the evolution of the Markov chain $X_t$ can be represented by $X_{t+1} = f(X_t, i(t), j(t), \lambda(t))$, where $f$ is a deterministic function, $i(t), j(t)$ are random coordinates (either elements of $[n]$ or of a group $G$), and $\lambda(t)$ is drawn from Lebesgue measure on $[0,1]$. These representations are given in equations (10), (11) and (1) respectively. To couple $X_t$ and $Y_t$, it is thus enough to couple the update variables $i(t)^\alpha, j(t)^\alpha, \lambda(t)^\alpha$, with $\alpha \in \{x, y\}$, used to construct $X_t$ and $Y_t$ respectively.

Our couplings will provide bounds on mixing times through the following lemma (see [13], Theorem 5.2 - they work in discrete space, but their proof doesn’t rely on this assumption):
Lemma 3 (Fundamental Coupling Lemma). If \((X_t, Y_t)\) is a coupling of Markov chains, \(Y_0\) is distributed according to the stationary distribution of \(K\), and \(\tau\) is a random time with the property that \(X_t = Y_t\) for \(t \geq \tau\), then

\[ ||\mathcal{L}(X_t) - \mathcal{L}(Y_t)||_{TV} \leq P[\tau > t] \]

In each chain, then, we begin with \(X_t\) started at a distribution of our choice, and \(Y_t\) started at stationarity. For any fixed (large) \(T\), we will then couple \(X_t\) and \(Y_t\) so that they will have coupled by time \(T\) with high probability. Each coupling will have two phases: an initial phase from time 0 to time \(T_1\) in which \(X_t\) and \(Y_t\) get close with high probability, and a non-Markovian coupling phase from time \(T_1\) to time \(T = T_1 + T_2\) in which they are forced to collide. Unlike many coupling proofs, the time of interest \(T\) must be specified before constructing the coupling.

While the initial contraction phases are quite different for the two chains, the final coupling phase can be described in a unified way. The unifying device is the partition process introduced in \([25]\) for a special case of the first sampler treated here. This partition process contains some information about the coordinates \(i(t), j(t)\) used by \(Y_t\) throughout the entire process, and is the only source of information from the future that is used to construct the non-Markovian coupling. Critically, we don’t use any information about the random variables \(\lambda(t)\) used at each step, which makes it trivial to check that the couplings constructed in this paper have the correct marginal distributions.

The process \(\{P_t\}_{t=T_1}^{T} \) will consist of a set of nested partitions of \([n]\), \(P_{T_1} \leq P_{T_1+1} \leq \ldots \leq P_T = \{\{1\}, \{2\}, \ldots, \{n\}\}\), where we say partition \(A\) is less than partition \(B\) if every element of partition \(B\) is a subset of an element of partition \(A\). To construct \(P_t\), we first look at the sequence of graphs \(G_t\) with vertex set \([n]\) and edge set \(\{(i(s), j(s)) : s \geq t\}\). Then let \(P_t\) consist of the connected components of \(G_t\). While constructing \(P_t\), we will also record a series of ‘marked times’ \(T - 1 = t_1 > t_2 > \ldots > t_m\) and associated special subsets \(S(t_j, 1)\) and \(S(t_j, 2)\) of \([n]\). We will set \(t_1 = T - 1\), and then inductively set \(t_j = \sup\{t : t < t_{j-1}, P_t \neq P_{t+1}\}\). Finally, note that if \(P_{t-1} \neq P_t\), the only difference between them is that two elements of \(P_t\) have been merged into a single element in \(P_{t-1}\). Label the set merged at time \(t_j\) with fewer elements \(S(t_j, 1)\), and label the other one \(S(t_j, 2)\). If both sets have the same number of elements, set \(S(t_j, 1)\) to be the one containing the smallest element (this is, of course, quite arbitrary).

We will be interested in the smallest time \(\tau\) such that \(P_{T-\tau} = [n]\), a single block (set \(\tau = \infty\) if \(P_t\) is never a single block). Lemma 4.2 of \([25]\), a small adaptation of classical arguments (see e.g. chapter 7 of \([2]\)), says:
Lemma 4 (Connectedness). Let \( \epsilon > 0 \) and assume \( n > 4 \) satisfies \( \frac{n}{\log(n)} > \frac{3(1+2\epsilon)(1+2\epsilon)}{\epsilon^2} \). For the Gibbs sampler on narrow matrices,

\[
P \left[ \tau < \left( \frac{1}{2} + 2\epsilon \right) n \log(n) \right] \leq 2n^{-\epsilon}
\]

The analogous lemma for the other example will be proved in Section 5, Lemma 7.

For both of our walks, we will use two types of coupling, the ‘proportional’ coupling and the ‘subset’ coupling. In both cases, we will set \( i(t)^x = i(t)^y \) and \( j(t)^x = j(t)^y \) at each step. In the proportional coupling, we will also set \( \lambda(t)^x = \lambda(t)^y \).

To discuss the subset coupling, we must define the weight of \( X_t \) on a subset \( S \subset [n] \), which we call \( w(X_t, S) \). For the simplex walk, we define \( w(X_t, S) = \sum_{s \in S} X_t[s] \). For narrow matrices, we define \( w(X_t, S) = \sum_{s \in S} X_t[s, 1] \). The subset couplings associated with subset \( S \subset [n] \) is defined immediately prior to Lemma 5 in terms of the walk on the simplex (the coupling for the walk on narrow matrices is identical, but uses the representation of that walk in equations (10) and (11) rather than the representation for the simplex walk given in equation (1)). Roughly, the subset coupling at time \( t \) will often set \( w(X_{t+1}, S) = w(Y_{t+1}, S) \). We say that a subset coupling of subset \( S \) at time \( t \) succeeds if that equality holds; otherwise, we say it fails.

In each case, the coupling of \( X_t \) and \( Y_t \) during the non-Markovian coupling phase will be as follows. At marked times \( t_j \), we will perform a subset coupling of \( X_{t_j}, Y_{t_j} \) with respect to \( S(t_j, 1) \). At all other times, we will perform a proportional coupling. This leads to:

Lemma 5 (Final Coupling). Assume the non-Markovian coupling phase lasts from time \( T_1 \) to \( T \), that \( P_{T_1} = \{[n]\} \), and that all subset couplings succeed. Then \( X_T = Y_T \).

Proof. Let \( \mathcal{F}_t \) denote the collection of equations \( w(X_t, S) = w(Y_t, S) \) for all \( S \in P_t \). We will show by induction on \( t \) that the equations \( \mathcal{F}_t \) hold for all \( T_1 \leq t \leq T \). At time \( T_1 \), we have \( w(X_{T_1}, [n]) = w(Y_{T_1}, [n]) = 1 \). By definition of the partition process, if \( t \) is not a marked time and all equations \( \mathcal{F}_t \) hold, then all equations \( \mathcal{F}_{t+1} \) also hold. In fact, this is true for any coupling of \( \lambda(t)^x, \lambda(t)^y \) at that step, not just the proportional coupling.

Assume \( t = t_j \) is a marked time, and that the equations \( \mathcal{F}_{t_j} \) hold. Then if the equations \( \mathcal{F}_{t_j+1} \) don’t all hold, we must have that either \( w(X_{t_j+1}, S(t_j, 1)) \neq w(Y_{t_j+1}, S(t_j, 1)) \) or \( w(X_{t_j+1}, S(t_j, 2)) \neq w(Y_{t_j+1}, S(t_j, 2)) \), since none of the terms in the other equations change. By assumption, all subset couplings have succeeded, so \( w(X_{t_j+1}, S(t_j, 1)) = w(Y_{t_j+1}, S(t_j, 1)) \). By construction, \( w(X_{t_j+1}, S(t_j, 2)) = w(X_{t_j+1}, S(t_j, 1) \cup S(t_j, 2)) - w(Y_{t_j+1}, S(t_j, 1)) \) and similarly for \( Y_{t_j+1} \), so \( w(X_{t_j+1}, S(t_j, 2)) = w(Y_{t_j+1}, S(t_j, 2)) \). Thus, the inductive claim has been proved.
Finally, we note that if \( w(X_t, \{i\}) = w(Y_t, \{i\}) \) for any singleton \( \{i\} \), then \( X_t[i] = Y_t[i] \) for the sampler on the simplex (respectively \( X_t[i, j] = Y_t[i, j] \) for \( j \in \{1, 2\} \) for the other sampler). Since \( P_T = \{\{1\}, \{2\}, \ldots, \{n\}\} \), this proves the lemma.

So, in both cases, to show that coupling has succeeded, it is sufficient to show that all subset couplings succeed with high probability.

4. Contraction for Gibbs Samplers on the Simplex with Geometry

In this section, we prove a contraction lemma for Gibbs samplers on the simplex associated with a group \( G \) and symmetric generating set \( R \) of \( G \) (that is, \( R^{-1} = R \)), where \( |G| = n \), \( |R| = m \), and \( id \) is the identity element of \( G \). We recall briefly some definitions. We write \( \Delta G = \{X \in \mathbb{R}^G | X[g] \geq 0, \sum_{g \in G} X[g] = 1\} \). If \( X_t \in \Delta G \) is a copy of the Markov chain, we take a step by choosing \( g \in G, r \in R \) and \( \lambda \in [0, 1] \) uniformly and setting \( X_{t+1}[g] = \lambda(X_t[g] + X_t[gr]), X_{t+1}[gr] = (1 - \lambda)X_t[g] + X_t[gr] \), and for all other entries \( X_{t+1}[h] = X_t[h] \). This walk is closely related to a slow simple random walk on the group. In particular, we let \( Z_t \in G \) be the random walk that evolves by choosing at each time step a group element \( g \in G \) and generator \( r \in R \) uniformly at random, and setting \( Z_{t+1} = Z_tr \) if \( Z_t = g \), and \( Z_{t+1} = Z_t \) otherwise.

Let \( \tilde{K} \) be the transition kernel associated with the random walk \( Z_t \). Since \( R \) is symmetric, the random walk is reversible, so \( \tilde{K} \) can be written in a basis of orthogonal eigenvectors with real eigenvalues \( 1 = \hat{\lambda}_1 > \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_n \geq -1 \). Since it is \( \frac{1}{2} \)-lazy, all eigenvalues are in fact nonnegative. Let \( \hat{\gamma} = 1 - \hat{\lambda}_2 \) be the spectral gap of \( \tilde{K} \). In this section we will show that

**Lemma 6** (Contraction Estimate for Gibbs Sampler on Cayley Graphs). Let \( X_t, Y_t \) be two copies of the Gibbs sampler on the simplex associated with \( G \) and \( R \), with joint distribution given by a proportional coupling at each step. Then

\[
E[\|X_t - Y_t\|_2^2] \leq 4ne^{-\frac{\hat{\gamma}t}{2}}
\]

**Proof.** We will construct an auxilliary Markov chain on \( G \) associated with \( X_t \), and compare it to the standard random walk \( Z_t \). Let \( X_t, Y_t \) be two copies of the walk, and couple them at each step with the proportional coupling. For \( h \in G \), let \( S_t^h = \sum_{g \in G} (X_t[g] - Y_t[g])(X_t[hg] - Y_t[hg]) \). We will analyze the evolution of the vector \( S_t = (S_t^i, \ldots) \).
There are three cases to analyze: $h \notin R$ and $h \neq id$, $h \in R$ and $h \neq id$, and $h = id$. Let $\mathcal{F}_t$ be the $\sigma$-algebra generated by $X_s$ and $Y_s$, $0 \leq s \leq t$. For case 1, we have

$$E[S^h_{t+1}|\mathcal{F}_t] = \left(1 - \frac{4}{n} + \frac{2}{mn}\right) S^h_t + \frac{1}{2mn} \sum_{iG} \sum_{r \in R, r \neq h, h^{-1}} [(X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri])(X_t[hi] - Y_t[hi])]$$

$$+ (X_t[ri] + X_t[i] - Y_t[ri] - Y_t[i])(X_t[hrri] - Y_t[hrri])$$

$$+ (X_t[h^{-1}i] - Y_t[h^{-1}i])(X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri])$$

$$+ (X_t[h^{-1}ri] - Y_t[h^{-1}ri])(X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri])$$

$$+ \frac{2}{mn} \sum_{iG} \left[\frac{1}{6}(X_t[i] + X_t[hi] - Y_t[i] - Y_t[hi])^2 + (X_t[i] - Y_t[i])(X_t[hi] - Y_t[hi]) + (X_t[i] - Y_t[i])(X_t[h^2i] - Y_t[h^2i])\right]$$

$$= \left(1 - \frac{2}{n} + \frac{2}{3mn}\right) S^h_t + \frac{2}{3mn} S^{id}_t + \frac{2}{mn} S^{h^2}_t$$

$$+ \frac{1}{2mn} \sum_{r \in R, r \neq h, h^{-1}} (S^{hr^{-1}}_t + S^{hr}_t + S^{rh}_t + S^{rh^{-1}}_t)$$

and we note that the sum of the coefficients is $1 - \frac{2}{3mn}$. For case 2, we have

$$E[S^{id}_{t+1}|\mathcal{F}_t] = \left(1 - \frac{4}{n}\right) S^{id}_t + \frac{2m}{mn} S^{id}_t$$

$$+ \frac{1}{2mn} \sum_{r \in R} (S^{hr^{-1}}_t + S^{hr}_t + S^{rh}_t + S^{rh^{-1}}_t)$$

$$= \left(1 - \frac{2}{n}\right) S^{id}_t + \frac{1}{2mn} \sum_{r \in R} (S^{hr^{-1}}_t + S^{hr}_t + S^{rh}_t + S^{rh^{-1}}_t)$$

where the sum of the coefficients is 1. Finally, in case 3, we have

$$E[S^{id}_{t+1}|\mathcal{F}_t] = \left(1 - \frac{2}{n}\right) S^{id}_t + \frac{2}{3mn} \sum_{r \in R} \sum_{iG} (X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri])^2$$

$$= \left(1 - \frac{2}{3n}\right) S^{id}_t + \frac{4}{3mn} \sum_{r \in R} S^r_t$$

and here the sum of the coefficients is $1 + \frac{2}{3n}$. If we rewrite $U^{id}_t = \frac{1}{2} S^{id}_t$, and $U^q_t = S^q_t$ for $g \neq id$, and $U_t = (U^{id}_t, \ldots)$, then we find the following transformations. For case
1, we have

\[
E[U_{t+1}^h | F_t] = \left(1 - \frac{2}{n} + \frac{2}{3mn}\right) U_t^h + \frac{4}{3mn} U_t^{id} + \frac{2}{mn} U_t^{h^2} + \frac{1}{2mn} \sum_{r \in R, r \neq h, h^{-1}} (U_t^{hr^{-1}} + U_t^{hr} + U_t^{rh} + U_t^{r^{-1}h})
\]

For case 2, we have

\[
E[U_{t+1}^h | F_t] = \left(1 - \frac{2}{n}\right) U_t^h + \frac{1}{2mn} \sum_{r \in R} (U_t^{hr^{-1}} + U_t^{hr} + U_t^{rh} + U_t^{r^{-1}h})
\]

Finally, in case 3, we have

\[
E[U_{t+1}^{id} | F_t] = \left(1 - \frac{2}{3n}\right) U_t^{id} + \frac{2}{3mn} \sum_{r \in R} U_t^r
\]

where the sum of the coefficients is now 1 in all three cases. Since the coefficients are now all nonnegative numbers summing to 1, the equations (2) to (4) define the transition kernel of a Markov chain on $G$. From equation (4), this chain sends the identity to itself with probability $1 - \frac{2}{3n}$, and to a uniformly chosen element of $R$ with the remaining probability; Equations (2) and (3) describe transitions from $h \in R$ and $h \neq id, h \notin R$ respectively. Call the transition kernel $K$.

Before analyzing the chain, we note that $\sum_{i \in G} (X_t[i] - Y_t[i]) = 0$, and so

\[
0 = \left(\sum_{i \in G} (X_t[i] - Y_t[i])\right)^2 = \sum_{i \in G} (X_t[i] - Y_t[i])^2 + \sum_{i \neq j} (X_t[i] - Y_t[i])(X_t[j] - Y_t[j]) = S_t^{id} + \sum_{h \neq id} S_t^h
\]

From this calculation, if $\langle v, (2, 1, 1, \ldots, 1) \rangle = 0$, then $\langle Kv, (2, 1, 1, \ldots, 1) \rangle = 0$ as well. By direct computation, $\pi = \frac{1}{n+1}(2, 1, 1, \ldots, 1)$ is a reversible measure for $K$. It is also clear that the distribution $\hat{\pi} = \frac{1}{n}(1, 1, \ldots, 1)$ is the reversible measure for $\hat{K}$.

We are now ready to compare the chains. Recall from [8] that the Dirichlet form associated to a Markov chain with transition kernel $Q$ and stationary distribution $\nu$ is given by

\[
\mathcal{E}(\phi) = \frac{1}{2} \sum_{h, g \in G} \nu(g) Q(g, h)(\phi(g) - \phi(h))^2
\]
Let $\mathcal{E}$ and $\mathcal{\hat{E}}$ be the Dirichlet forms associated with $K$ and $\hat{K}$ respectively. Then by comparing terms, it is clear that $\mathcal{E}(\phi) \geq \frac{1}{4}\mathcal{\hat{E}}(\phi)$ for any $\phi$ and $\frac{\gamma}{4}, \frac{\hat{\gamma}}{8} \leq 2$. By e.g. Lemma 13.12 of [13], this implies $\gamma \geq \frac{1}{8}\hat{\gamma}$.

Recall that if $\langle v, \pi \rangle = 0$, then $\langle K^m v, \pi \rangle = 0$ as well. In particular $K$ applied to the subspace orthogonal to $\pi$ has $L^2 \to L^2$ operator norm at most $1 - \gamma$. Thus, we have for any $v$ in that subspace

$$||K^m v||_2 \leq e^{-\frac{\gamma m}{8}}||v||_2$$

going back to our original situation, we are interested in the vector $(S^0_t)$. At time 0, $S^0_0 \leq 2$, and by Cauchy-Schwarz $|S^0_t| \leq |S^0_0|$. Thus, $||U_0||_1 \leq 2n$, and of course $|S^d_t| \leq ||S_t||_1 \leq 2||U_t||_1$. So we find that

$$E[|S^d_t|] \leq 4ne^{-\frac{t\gamma}{8}}$$

which is the contraction estimate in Lemma 6. ■

5. Coupling for Gibbs Samplers on the Simplex with Geometry

Having shown contraction, we must now show convergence in total variation distance. First, the analogue to Lemma 4.

**Lemma 7** (Connectedness for Gibbs Sampler on Cayley Graphs). Let $\tau$ be as defined immediately before Lemma 4 and let $\hat{\gamma}$ be as defined immediately before Theorem 7. Then for $t > \frac{8(C+3)\log(n)}{\hat{\gamma}}$, we have

$$P[\tau > t] \leq 2n^{-C}$$

**Proof.** We consider a graph-valued process $G_t$, where $G_0$ is a graph with no edges, and vertex set equal to the group $G$. To construct $G_{t+1}$ from $G_t$, choose elements $g \in G$ and $r \in R$ uniformly at random, and add the edge $(g, gr)$ if it isn’t already in $G_t$. We note that $\tau > t$ if and only if $G_t$ is not connected, so we would like to estimate the time at which $G_t$ becomes connected.

First, fix two elements $x, y \in G$. We’d like to see if $x, y$ are in the same component of $G_t$. To do so, let $X_t, Y_t$ be two copies of the Gibbs sampler described in the last section, with $X_0 = x, Y_0 = y$. Couple $X_t, Y_t$ and $G_t$ by the proportional coupling. Then assume $x, y$ are in different components $C_x, C_y$ at time $t$. We would have

$$\sum_g |X_t[g] - Y_t[g]|^2 \geq \sum_{g \in C_x} \frac{1}{|C_x|^2} + \sum_{g \in C_y} \frac{1}{|C_y|^2} \geq \frac{4}{n}$$

By Markov’s inequality, then,

$$P[C_x \neq C_y] \leq \frac{n}{4}E[\sum_g |X_t - Y_t|^2]$$
and so, by standard union bound for fixed $x$ over all $y$, if $A_t$ is the event that $G_t$ is disconnected,

$$P[A_t] \leq \frac{n^2}{4} \sup_{\mu, \nu} E[\sum_g |X_t - Y_t|^2 |X_0 = \mu, Y_0 = \nu]$$

$$\leq n^3 e^{-\frac{|c|^2}{8}}$$

where the last inequality is due to Lemma 6.

Next, we define subset couplings and discuss success probabilities for this walk. Fix points $X_t, Y_t$, subset $S \subset [n]$ and updated coordinates $i = i(t) \in S, j = j(t) \not\in S$. The next step is to construct a pair of uniform random variables $\lambda_x = \lambda(t)^x$ and $\lambda_y = \lambda(t)^y$ with which to update the chains $X_t$ and $Y_t$ respectively. Assume first that $\frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} < 1$, and choose $\lambda_y$ uniformly in $[0, 1]$. Then set

$$\lambda_x = \lambda_y \frac{Y_t[i] + Y_t[j]}{X_t[i] + X_t[j]} + \frac{1}{X_t[i] + X_t[j]} \sum_{s \in S \setminus \{i\}} (Y_t[s] - X_t[s])$$

(5)

if that results in a value between 0 and 1. Otherwise, choose $\lambda_x$ independently of $\lambda_y$, according to the density:

$$f(\lambda) = C \left(1 - \frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} \frac{1}{X_t[i] + X_t[j]} \sum_{s \in S \setminus \{i\}} (Y_t[s] - X_t[s]) \right)^{\frac{1}{\lambda}}$$

(6)

where $C$ is a normalizing constant, and

$$g(\lambda) = \lambda \frac{Y_t[i] + Y_t[j]}{X_t[i] + X_t[j]} + \frac{1}{X_t[i] + X_t[j]} \sum_{s \in S \setminus \{i\}} (Y_t[s] - X_t[s])$$

From the assumption that $\frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} < 1$, it is easy to see that $f$ really is a density on $[0, 1]$. From its construction as a remainder density, it is easy to check that under this coupling, $\lambda_x$ is uniformly distributed on $[0, 1]$. If $\frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} > 1$, an analogous construction will work. More precisely, in this case choose $\lambda_x$ first, and then choose $\lambda_y$ to satisfy equation 5 if the result is in $[0, 1]$, rather than choosing $\lambda_y$ first. If the result is not in $[0, 1]$, then choose $\lambda_y$ according to its remainder measure, given by equation 6 with $X_t$ and $Y_t$ switched and $g$ replaced by $g^{-1}$. Note that if equation 5 is satisfied, then $w(S, X_{t+1}) = w(S, Y_{t+1})$.

For a pair of points $(x, y)$ in the simplex, a pair of update entries $(i, j)$, and a subset $S \subset [n]$ of interest such that $i \in S$ and $j$ not in $S$, we define $p(x, y, i, j, S)$ to be the probability that the associated subset coupling succeeds. Then Lemma 4.4 from [25] gives a lower bound on this probability:
Lemma 8 (Subset Coupling). Assume \( n \geq 6 \), and fix \( 0 \leq b \leq f - 1 \). For a pair of vectors \((x, y)\) satisfying \( \sup_i |x_i - y_i| \leq n^{-f} \) and \( \inf_i x_i, \inf_i y_i \geq n^{-b} \), we have \( p(x, y, i, j, S) \geq 1 - 3n^{b+1-f} \) uniformly in \( S \) and possible \( i, j \).

In general, it is possible to choose \( x, y, i, j, S \) so that the probability of success is 0 under any coupling, and the lemma is quite restrictive. Having bounded the probability of failure when \( X_t, Y_t \) are close, we must show that they remain close with high probability. Define for \( v \in \mathbb{R}^n \) and \( S \subset [n] \) the quantity \( ||v||_S = \sum_{s \in S} |v[s]| \). We will need Lemma 4.5 from \([25]\).

Lemma 9 (Closeness). Let \( X_t, Y_t \) be either the of the chains described in this paper, coupled as described in Section 3, and assume that \( P_{T_1} = \{[n]\} \), that all subset couplings up to time \( s - 1 \) have succeeded, and that \( ||X_{T_1} - Y_{T_1}||_1 < \epsilon \). Then \( ||X_s - Y_s||_S < \epsilon \) for every \( S \) in \( P_t \).

Related to this, Lemma 4.6 from \([25]\) shows that \( X_t, Y_t \) rarely have entries close to 0:

Lemma 10 (Largeness). \( P[\inf_{1 \leq i \leq n} \inf_{0 \leq t \leq T-1} Y_t[i] \leq Tn^{-2.5-k}] \leq 2Tn^{-k} \) for \( n > \max(2k, 4096) \).

This lets us complete the calculation. We split the run of \( \frac{8C\log(n)}{\gamma} \) steps into an initial contractive phase of length \( T_1 = 8(\frac{5}{8}C + \frac{7}{3}\frac{\log(n)}{\gamma}) \) and a second, coupling, phase of length \( T_2 = 8(\frac{1}{6}C - \frac{7}{2}\frac{\log(n)}{\gamma}) \). By Lemma 5, \( X_T = Y_T \) unless one of the subset couplings fails or \( \tau > T_2 \). By Lemma 7, \( P[\tau > T_2] \leq 2n^{-\frac{5}{6}C + 3.5} \). Thus, it remains only to bound the probability that a subset coupling fails, assuming \( \tau \leq T_2 \).

By Lemma 6, Markov’s inequality and the bound \( ||v||_1 \leq \sqrt{2n}||v||_2 \), \( P[||X_{T_1} - Y_{T_1}||_1 \geq n^{-\frac{C}{6} + \frac{1}{2}}] \leq 4\sqrt{2n}n^{-\frac{C}{6} - 1} \). It is easy to show that \( \frac{\gamma}{4} \geq \frac{1}{2n^4} \) for simple random walk on any Cayley graph (e.g. by naive bounds with Theorem 13.14 of \([13]\)). Combining this with Lemma 10, we have \( P[\inf_{g \in G, T_1 \leq t \leq T_3} Y_t[g] \leq n^{-\frac{C}{6} - 3}] \leq 2n^{1.5 - \frac{C}{6}} \) for \( n \) satisfying \( n > \max(4096, 2b, 2(\frac{C}{6} - 3.5) \log(n)) \). By Lemmas 8 and 9, as long as \( \inf_{g \in G, T_1 \leq t \leq T_3} Y_t[g] > n^{-b} \) and \( ||X_{T_1} - Y_{T_1}||_1 < n^{-\frac{C}{6} + \frac{1}{2}} \) and \( C > \frac{103}{4} \), we conclude that the all subset couplings succeed with probability at least \( 1 - 3n^{4.5 - \frac{C}{6}} \). Combining the bounds in this paragraph, as long as \( \tau \leq T_2 \), all subset couplings succeed with probability at least \( 1 - 6n^{4.5 - \frac{C}{6}} \).

Putting together the bounds in the last line of each of the proceeding two paragraphs, \( P[Y_T \neq X_T] \leq 7n^{4.5 - \frac{C}{6}} \), which proves the upper bound in the theorem.

6. LOWER BOUNDS FOR GIBBS SAMPLERS ON THE SIMPLEX WITH GEOMETRY

In this section, we prove lower bounds on the mixing time of the Gibbs sampler on the simplex. The results are similar to those of \([20]\), though the method is different.
and elementary. Begin by calculating

\begin{align}
E[X_{t+1}[g]|X_t] &= \left(1 - \frac{2}{n}\right)X_t[g] + \frac{1}{n} \sum_{r \in R} \frac{1}{m} (X_t[g] + X_t[gr]) \\
&= \left(1 - \frac{1}{n}\right)X_t[g] + \frac{1}{n} \sum_{r \in R} X_t[gr]
\end{align}

In particular, let \( K \) be the transition matrix on \( G \) given by

\[ K[g,g] = 1 - \frac{1}{n}, \quad K[g,gr] = \frac{1}{nm} \quad \text{for } r \in R. \]

This is the standard ‘edge’-based random walk on \( G \) with generating set \( R \) described above. By equation (7), \( E[X_t] = K^tX_0 \). Note that this is not the same \( K \) as was used earlier in the section while proving the upper bound on the mixing time. By the earlier assumptions on \( R \), \( K \) is reversible with respect to the uniform measure on \( G \). Furthermore, it is orthogonally diagonalizable with real eigenvalues \( 1 = \beta_1 > \beta_2 \geq \ldots \geq \beta_n \geq 0. \)

Next, let \( v \) be an eigenvector of \( K \) with eigenvalue \( \beta_2 \), normalized so that \( \|v\|_2 = 1 \) and \( \|v - K v\|_2 = \gamma \), the spectral gap of \( K \). Let \( \Pi \) be the collection of vectors with nonnegative entries summing to 1, and let \( w \in \Pi \) maximize the inner product \( \langle v, w \rangle \) among such vectors; such a vector exists by the compactness of \( \Pi \). Let \( X_t \) be a copy of the Markov chain begun from \( X_0 \overset{D}{=} w \), then \( E[\langle X_t, v \rangle] = (1 - \gamma)^t \langle w, v \rangle \). On the other hand, if \( A_{t,d} \) is the event that \( \langle X_t, v \rangle > d \),

\[ E[\langle X_t, v \rangle] = E[\langle X_t, v \rangle 1_{A_{t,d}}] + E[\langle X_t, v \rangle 1_{A_{t,d}}^c] \leq \langle X_0, v \rangle P[A_{t,d}] + d \]

where the second inequality takes advantage of the maximality of \( X_0 \). Thus,

\[ P[A_{t,d}] \geq \frac{E[\langle X_t, v \rangle] - d}{\langle X_0, v \rangle} \]

putting the two inequalities together,

\begin{align}
P[A_{t,d}] &\geq (1 - \gamma)^t - \frac{d}{\langle X_0, v \rangle} \tag{9}
\end{align}

The next step is to prove that \( \langle X_0, v \rangle \geq \frac{1}{2\sqrt{n}} \). Let \( P \subset [n] \) be the collection of indices so that \( v[p] \geq 0 \) for \( p \in P \). Without loss of generality, assume \( \sum_{p \in P} v_p^2 > \frac{1}{2} \).

Now set \( \lambda^{-1} = \sum_{p \in P} v_p \leq \sqrt{n} \). Then consider the distribution given by \( \mu_p = \lambda v_p \) for \( p \in P \), and \( \mu_p = 0 \) for \( p \notin P \):

\[ \langle \mu, v \rangle = \lambda \sum_{p \in P} v_p^2 \geq \frac{1}{2\sqrt{n}} \]
and so by inequality (9),
\[ P[A_{t,d}] \geq (1 - \gamma)^t - 2d\sqrt{n} \]

Now, let \( Y \in \Delta_G \) be chosen according to the uniform distribution. Then \( E[\langle Y, v \rangle] = 0 \), and
\[
E[\langle Y, v \rangle^2] = E[\left( \sum_{i \in G} Y[i]v_i \right)^2] \\
= \sum_{i \in G} v_i^2 E[Y[i]^2] + \sum_{i \neq j \in G} v_i v_j E[Y[i]Y[j]] \\
\leq \frac{2}{n^2} + 0
\]

So, by Chebyshev’s inequality, \( P[\langle Y, v \rangle > d] \leq \frac{2}{d^2n^2} \). Putting this together with the inequality above, letting \( d = n^{-\frac{5}{6}} \) and defining \( P[A_{\infty,d}] = \lim_{t \to \infty} P[A_{t,d}] \),
\[ P[A_{t,d}] - P[A_{\infty,d}] \geq \gamma^t - 4n^{-\frac{1}{3}} \]
And the lower bound follows immediately.

7. CONTRACTION AND NARROW MATRICES

We begin with some quick observations about the geometry of our space. It is the part of an \((n - 1)\)-dimensional affine subspace of \( \mathbb{R}^{2n} \) that lies in the upper orthant. Our updates are in fact moves along 1-dimensional pieces of this subspace, even though we are updating four entries. While the original motivation for this sampler comes from statistics (see e.g. [4]), it is being treated here primarily as an example of a chain that is somewhere between the standard Gibbs sampler on the simplex and an analogous Gibbs sampler on doubly-stochastic matrices or Kac’s famous walk on the orthogonal group. The former was analyzed by the author in [25]. Matching bounds on Total variation mixing time are not known for either the Gibbs sampler on doubly-stochastic matrices or Kac’s walk. The best such bounds to date can be found in [24] and [12] respectively. Both mixing bounds are polynomials with small but probably incorrect degrees, and both are based on much more complicated non-Markovian coupling arguments.

In this section, we will prove the following contractivity estimate for the Gibbs sampler on narrow matrices. The original proof was a direct translation of the path-coupling argument for \( k \) by \( n \) matrices found in [24]. The following greatly simplified proof was suggested by an extremely helpful reviewer.

**Lemma 11** (Weak Convergence on Narrow Matrices). If \( X_t \) and \( Y_t \) are coupled under the proportional coupling for time \( 0 \leq t \leq T_1 = 3kn \log(n) \), then
\[ P[\|X_{T_1} - Y_{T_1}\|_1 \geq \epsilon] \leq \epsilon^{-1}n^{-k} \]
We make some basic remarks about the chain, beginning with an alternative description of the transition probabilities. Define \( \delta_t[i, j] = 2 - X_t[i, 1] - X_t[j, 1] \), and \( \epsilon_t[i, j] = 2 - Y_t[i, 1] - Y_t[j, 1] \). Then a step of the chain can be defined in the following way. Choose \( i, j \) as before, and choose \( \lambda \overset{\text{i.i.d.}}{\sim} U[0, 1] \). If \( \delta_t[i, j] \geq 0 \), then we update according to:

\[
\begin{align*}
X_{t+1}[i, 1] &= \lambda(2 - \delta_t[i, j]) \\
X_{t+1}[j, 1] &= (1 - \lambda)(2 - \delta_t[i, j]) \\
X_{t+1}[i, 2] &= 2(1 - \lambda) + \lambda\delta_t[i, j] \\
X_{t+1}[j, 2] &= 2\lambda + (1 - \lambda)\delta_t[i, j]
\end{align*}
\]

If \( \delta_t[i, j] < 0 \), we update according to:

\[
\begin{align*}
X_{t+1}[i, 1] &= 2\lambda - (1 - \lambda)\delta_t[i, j] \\
X_{t+1}[j, 1] &= 2(1 - \lambda) - \lambda\delta_t[i, j] \\
X_{t+1}[i, 2] &= (1 - \lambda)(2 + \delta_t[i, j]) \\
X_{t+1}[j, 2] &= \lambda(2 + \delta_t[i, j])
\end{align*}
\]

Note that in both cases, a larger value of \( \lambda \) means a larger value of \( X_{t+1}[i, 1] \). We are now ready to describe the proportional coupling: as in the simplex case, we choose the same value of \( \lambda \) for both chains in the above representation. This leads to the following contraction estimate:

**Lemma 12** \((L^2 \text{ Contractivity})\). If \( X_t, Y_t \) are coupled under the proportional coupling for time \( 0 \leq t \leq T_1 \), then

\[
E[||X_{T_1} - Y_{T_1}||_2^2] \leq (1 - \frac{2}{3\eta})^T_1 ||X_0 - Y_0||_2^2
\]

**Proof.** We begin the proof by calculating the change in the \( L^2 \) norm during a single move. Let \( F_t[i, j] \) be the event that coordinates \( i, j \) are updated at time \( t \). If \( \delta_t[i, j]\epsilon_t[i, j] \geq 0 \), we find:

\[
\Delta_t \equiv E[(X_{t+1}[i, 1] - Y_{t+1}[i, 1])^2 + (X_{t+1}[j, 1] - Y_{t+1}[j, 1])^2 + (X_{t+1}[i, 2] - Y_{t+1}[i, 2])^2 + (X_{t+1}[j, 2] - Y_{t+1}[j, 2])^2 | F_t[i, j]]
\]

\[
= 2E[(\lambda(2 - \delta_t[i, j]) - \lambda(2 - \epsilon_t[i, j]))^2 + (\lambda\delta_t[i, j] - \lambda\epsilon_t[i, j])^2 | F_t[i, j]]
\]

\[
= \frac{4}{3}(\delta_t[i, j] - \epsilon_t[i, j])^2
\]

If \( \delta_t[i, j]\epsilon_t[i, j] < 0 \), we find:
\[
\Delta_t = E[(X_{t+1}[i, 1] - Y_{t+1}[i, 1])^2 + (X_{t+1}[j, 1] - Y_{t+1}[j, 1])^2 \\
+ (X_{t+1}[i, 2] - Y_{t+1}[i, 2])^2 + (X_{t+1}[j, 2] - Y_{t+1}[j, 2])^2|F_t[i, j]] \\
= 2E[(\lambda(2 - \delta[i, j]) - 2\lambda + (1 - \lambda)\epsilon[i, j])^2|F_t[i, j]] \\
+ E[((1 - \lambda)(2 - \delta_t) - 2(1 - \lambda) + \lambda\epsilon[i, j])^2|F_t[i, j]] \\
= \frac{4}{3}(\delta_t[i, j] - \epsilon_t[i, j])^2 + \frac{1}{3}\delta_t[i, j]\epsilon_t[i, j] \\
< \frac{4}{3}(\delta_t[i, j] - \epsilon_t[i, j])^2
\]

As in the simplex case, we can calculate the sums of terms like \((\epsilon_t[i, j] - \delta_t[i, j])^2\) in terms of sums of terms like \((X_t[i, j] - Y_t[i, j])^2\). We first note that

\[
0 = \left(\sum_{i=1}^{n} X_t[i, 1] - Y_t[i, 1]\right)^2 \\
= \sum_{i=1}^{n} (X_t[i, 1] - Y_t[i, 1])^2 + 2 \sum_{i<j} (X_t[i, 1] - Y_t[i, 1])(X_t[j, 1] - Y_t[j, 1])
\]

If \(\delta_t[i, j]\epsilon_t[i, j] \geq 0\) for all \(i, j\), we can write the first line of the following computation:

\[
\sum_{i \neq j} (\delta_t[i, j] - \epsilon_t[i, j])^2 = \sum_{i \neq j} (X_t[i, 1] + X_t[j, 1] - Y_t[i, 1] - Y_t[j, 1])^2 \\
= \sum_{i \neq j} [(X_t[i, 1] - Y_t[i, 1])^2 + (X_t[j, 1] - Y_t[j, 1])^2 \\
+ 2(X_t[i, 1] - Y_t[i, 1])(X_t[j, 1] - Y_t[j, 1])] \\
= (n - 2) \sum_{j=1}^{n} \sum_{i=1}^{n} (X_t[i, j] - Y_t[i, j])^2
\]
Then the final contraction is given by:
\[
\Delta_t = E[||X_{t+1} - Y_{t+1}||^2_2|X_t, Y_t]
\]
\[
= \frac{1}{n(n-1)} \sum_{k=1}^{2} \sum_{m=1}^{n} \sum_{i \neq j} E[(X_{t+1}[m, k] - Y_{t+1}[m, k])^2|F_t[i, j]]
\]
\[
= \frac{1}{n(n-1)} \sum_{k=1}^{2} \sum_{m=1}^{n} \left( \sum_{i \neq j} (X_t[m, k] - Y_t[m, k])^2 + 2 \sum_{j \neq m} E[(X_{t+1}[m, k] - Y_{t+1}[m, k])^2|F_t[i, j]]\right)
\]
\[
= (1 - \frac{2}{n})||X_t - Y_t||_2^2 + \frac{1}{n(n-1)} \sum_{i \neq j} \frac{4}{3} (\delta_t[i, j] - \epsilon_t[i, j])^2
\]
\[
\leq (1 - \frac{2}{3n})||X_t - Y_t||_2^2
\]
where the last line is due to equation (12). □

Lemma 11 follows from this bound, the inequality $|X_t[i] - Y_t[i]| \leq ||X_t - Y_t||_2$, and Markov’s inequality.

8. Coupling for Narrow Matrices

In this section, we show that subset couplings are likely to succeed, and finish the proof of Theorem 2. The main lemma is:

**Lemma 13 (Coupling for Nearby Points).** Fix $a > 7.5$ and $n > \max(4096, a + 3.5)$ satisfying $\frac{n}{\log(n)} > \frac{3(1+2c)(\frac{1}{2} + 2c)}{c}$. Let $Y_t$ be a copy of the chain started at stationarity and assume that $X_t$ is a copy of the chain which satisfies $||X_{T_1} - Y_{T_1}||_1 \leq n^{-a}$. Then for $T_2 > (\frac{1}{2} + 2c)n \log(n)$ and $T = T_1 + T_2$ we have $P[X_T \neq Y_T] \leq 10n^{3.75 - \frac{a}{2} - n^{-c}}$.

Construct a partition process from time $T_1$ to time $T = T_1 + (\frac{1}{2} + 2c)n \log(n)$. Our first step is to define subset couplings and show that if $X_t$ and $Y_t$ are very close to each other and not too close to certain hyperplanes, then any subset couplings are likely to succeed. To define a subset coupling of $X_t$ and $Y_t$, fix the subset $S$ of interest and common update variables $i = i(t) \in S$ and $j = j(t) \notin S$. If $\delta_t[i, j] \epsilon_t[i, j] \geq 0$, then the coupling of $\lambda^x_t$ and $\lambda^y_t$ is exactly as described for the other walk immediately before Lemma 8. Otherwise, assume $\delta_t[i, j] < 0$ and $\epsilon_t[i, j] > 0$, and choose $\lambda^x_t$ from $[0, 1]$ uniformly at random. Then set $\lambda^y_t$ to be the number which satisfies $w(X_{t+1}, S) = w(Y_{t+1}, S)$ if such a number exists and is in the interval $[0, 1]$. Just as with equation (5), the measure (with mass less than 1) on $\lambda^y_t$ that this assignment defines minorizes the uniform distribution, and so leaves a remainder distribution analogous to that given in equation (6). If there is no value of $\lambda^y_t$ in $[0, 1]$ which would allow $w(X_{t+1}, S) =$
Lemma 15 (Subset Coupling). Assume \( n \geq 6 \) and fix \( a - 2 > b > 0 \). For a pair of matrices \((x, y)\) satisfying \( \sup_{m,k} |x[m,k] - y[m,k]| \leq n^{-a} \) and \( \inf_{m,k} (x[m,k], y[m,k], 2 - x[m,k], 2 - y[m,k]) \geq n^{-b} \), we have for all sufficiently large \( n \) that \( p(x, y, i, j, S) \geq 1 - 4n^{b+2-a} \) uniformly in \( S \subseteq [n] \) and pairs \( i \in S, j \notin S \).
Next, as in Lemma 9, note that after a successful subset coupling involving sets $S$ and $R$ at time $t$, we have $||X_{t+1} - Y_{t+1}||_{1,S} \leq ||X_t - Y_t||_{1,S \cup R}$. Thus, if all subset couplings until time $t$ have succeeded,

\begin{equation}
||X_t - Y_t||_{1,A} \leq ||X_{T_1} - Y_{T_1}||_1
\end{equation}

for all $S \in P_t$.

We are ready to prove Lemma 13. By Lemma 5, $X_T = Y_T$ unless at least one subset coupling has failed or $P_{T_1} \neq \{[n]\}$. Let $E_1$ be the event that $P_{T_1} \neq \{[n]\}$, let $E_2$ be the event that $\inf_{i,j} \inf_{T_i \leq T_j} (Y_t[i,j], 2 - Y_t[i,j]) \geq 3^{-25/75}$, and let $E_3$ be the intersection of $E_2^c$ with the event that at least one subset coupling fails. By Lemma 4, $P[E_1] \leq 2n^{3.75-2\epsilon}$. By Lemma 15 and Lemma 9, $P[E_3] \leq 4n^{3.75-\epsilon}$. By Lemma 14, $P[E_2] \leq 4n^{3.75-\epsilon}$. This completes the proof of Lemma 13.

Finally, we prove Theorem 2. Fix $a = \frac{2}{11}(C + 18.25)$. We divide our run of length $T = Cn \log(n)$ into a contractive phase of length $T_1 = (\frac{2}{11}a - 11.25) n \log(n)$ and a coupling phase of length $T_2 = (a - 7) n \log(n)$. Let $E_1$ be the event that $||X_{T_1} - Y_{T_1}||_1 > n^{-a}$, and let $E_2$ be the intersection of $E_1^c$ with the event that $X_T \neq Y_T$. By Lemma 11, $P[E_1] \leq n^{3.75-\epsilon}$. By Lemma 13, $P[E_2] \leq 10n^{3.75-\epsilon}$. This completes the proof of the upper bound in Theorem 2.

To prove the lower bound, let $\tau$ be the (random) first time at which all $2n$ coordinates have been updated. Then fix the starting position $X_0$ of the Markov chain and let $H_{i,j} = \{X \in \mathbb{R}^{2n} | X[i,j] = X_0[i,j]\}$ and set $H = \cup_{i,j} H_{i,j}$. Then $P[X_t \in H] - U_n(H) \geq P[\tau > t]$. Since only four of $2n$ coordinates are chosen at a time, the classical coupon-collector results in $\cite{10}$ tell us that at time $T = \frac{1}{\epsilon}n(\log(n) - c)$, $|K^T_n(x, H) - \pi(H)| \geq 1 - \exp(-\exp(c)) + o(1)$ as $n$ goes to infinity.

\section*{References}

\begin{enumerate}
\item Olena Blumberg. A coupling proof for random transpositions. \textit{Preprint}, 2011.
\item Bela Bollobas. \textit{Random Graphs}, volume 73 of \textit{Cambridge Studies in Advanced Mathematics}. Press Syndicate of the University of Cambridge, Cambridge, 2001.
\item Robert Burton and Yevgeniy Kovchegov. Mixing times via super-fast coupling. \textit{Technical Report ORST-MATH}, 2011.
\item Persi Diaconis and Bradley Efron. Testing for independence in a two-way table: New interpretations of the chi-square statistic. \textit{Ann. Stat.}, 13(3):845–874, 1985.
\item Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, exponential families and orthogonal polynomials. \textit{Statistical Science}, 23:151–178, 2008.
\item Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, conjugate priors and coupling. \textit{Sankhya}, 72-A(1):136–169, 2009.
\item Persi Diaconis and Laurent Saloff-Coste. Comparison techniques for random walks on finite groups. \textit{Annals of Probability}, 21(4):2131–2156, 1993.
\item Martin Dyer, Leslie Goldberg, Mark Jerrum, and Russel Martin. Markov chain comparison. \textit{Probability Surveys}, 3:89–111, 2006.
\end{enumerate}
[9] Martin Dyer, Ravi Kannan, and John Mount. Sampling contingency tables. Random Structures and Algorithms, 10:487–506, 1997.
[10] Paul Erdos and Alfred Renyi. On a classical problem of probability theory. Magyar Tud. Akad. Mat. Kutato. Int. Kozl, 1961.
[11] Tom Hayes and Eric Vigoda. A non-markovian coupling for randomly sampling colorings. FOCS proceedings, 2003.
[12] John Jiang. Polynomial mixing time of the kac random walk on the orthogonal group. Preprint, 2012.
[13] David Levin, Yuval Peres, and Elizabeth Wilmer. Markov Chains and Mixing Times. American Mathematical Society, Providence, Rhode Island, 2009.
[14] Jun Liu. Monte Carlo Strategies in Scientific Computing. Springer Series in Statistics. Springer, 2001.
[15] Laslo Lovasz. Hit-and-run mixes fast. Math. Prog., 1998.
[16] Laslo Lovasz and Santosh Vempala. Hit and run is fast and fun. Technical Report - Microsoft Research, 2003.
[17] Peter Matthews. Mixing rates for a random walk on the cube. SIAM. J. on Algebraic and Discrete Methods, 8(4):746–752, 1987.
[18] Ben Morris. Random walks in convex sets. PhD Thesis, University of California, Berkeley, 2000.
[19] Ben Morris. Improved bounds for sampling contingency tables. Random Structures and Algorithms, 21:135–146, 2002.
[20] Dana Randall and Peter Winkler. Mixing points on a circle. Lecture Notes in Computer Science, 3624:426–435, 2005.
[21] Dana Randall and Peter Winkler. Mixing points on an interval. Proceedings of ANALCO, 2005.
[22] Jeffrey Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. JASA, 90:558–566, 1995.
[23] Devavrat Shah. Gossip algorithms. Foundations and Trends in Networking, 3(1):1–125, 2009.
[24] Aaron Smith. Some analyses of markov chains by the coupling method. PhD Thesis, Stanford University, 2012.
[25] Aaron Smith. A gibbs sampler on the n-simplex. To Be Published in Annals of Applied Probability, 2013.
[26] Wai Kong Yuen. Applications of geometric bounds to convergence rates of of markov chains and markov processes on rn. PhD Thesis, University of Toronto, 2001.

Department of Mathematics, Stanford University, Stanford, CA 94305
E-mail address: asmith3@math.stanford.edu