A GIBBS SAMPLER ON THE *n*-SIMPLEX

BY AARON SMITH[1]

ICERM, Brown University

We determine the mixing time of a simple Gibbs sampler on the unit simplex, confirming a conjecture of Aldous. The upper bound is based on a two-step coupling, where the first step is a simple contraction argument and the second step is a non-Markovian coupling. We also present a MCMC-based perfect sampling algorithm based on our proof which can be applied with Gibbs samplers that are harder to analyze.

1. Introduction. Given a measure μ on a convex body $K \subset \mathbb{R}^n$, how can we efficiently obtain independent samples from the distribution of μ? This problem arises in the computational sciences, and a frequently-used tool is Markov chain Monte Carlo (MCMC) [5]. Because MCMC methods produce nearly-independent samples only after a lengthy mixing period, a long-standing mathematical question is to analyze the mixing times of the MCMC algorithms in common use.

The analysis of discrete MCMC algorithms is very advanced, with precise bounds for many difficult problems as well as some general theory that has received recent exposition in [1, 11]. For samplers on continuous state spaces, there has been some general theory based on geometric or coupling arguments (see [12, 13, 24] and [20]), but many of the techniques built for discrete chains seem to run into technical difficulties. There are also very few well-understood simple chains, in stark contrast to the discrete theory, which has been built on many detailed analyses of specific chains; though, see [16, 17] for some very nice analyses of two slower walks on the simplex; [18, 19] for group walks; and [10] for some applications. This paper is an attempt to carefully analyze a simple continuous chain, namely a Gibbs sampler on the *n*-simplex. In addition, it illustrates the use of two powerful techniques from the discrete theory: non-Markovian coupling [2, 4, 7] and coupling from the past [15].

The ideas in this paper can be used for a number of other problems. The analysis was initially motivated by a simpler version of Kac's random walk on $S(n)$ or $\mathrm{SO}(n)$ (see [8, 14, 23] and especially [9] for recent progress). It is also a stepping stone toward analysis of Gibbs samplers on more complicated convex sets, such as contingency tables. In the author's thesis and a forthcoming note, we use the

technique in this paper to improve existing analyses of these samplers and some others [22, 23]; there is still substantial room for improvement.

In this paper, we will discuss mixing in terms of the popular total variation distance. For a Markov chain with transition kernel K on a measurable space (Ω, Σ) and unique stationary distribution π, the total variation distance to stationarity after t steps of a Markov chain started at $\omega \in \Omega$ is given by

$$\sup_{A \in \Sigma} |K^t(\omega, A) - \pi(A)|.$$

Most of this paper will be concerned with a specific Gibbs sampler X_t on the n-simplex $\Delta_n = \{X \in \mathbb{R}^n \mid \sum_{i=1}^n X[i] = 1; X[i] \geq 0\}$ whose stationary distribution is the uniform distribution on Δ_n. To take a move in this Markov chain, begin by choosing $1 \leq i < j \leq n$ and $\lambda \in [0, 1]$ independently and uniformly. Then set

$$
\begin{aligned}
(1.1) \qquad & X_{t+1}[i] = \lambda(X_t[i] + X_t[j]), \\
& X_{t+1}[j] = (1 - \lambda)(X_t[i] + X_t[j]), \\
& X_{t+1}[k] = X_t[k] \qquad (k \neq i, j).
\end{aligned}
$$

This sampler was first mentioned in [1], where the mixing time was shown to be $O(n^2 \log n)$. Aldous suggested in his list of open problems that the correct mixing time was $O(n \log n)$, and we confirm this, also demonstrating a pre-cutoff window of moderate size:

THEOREM 1.1 (Simplex mixing time). *Fix $C > 3$ and n satisfying $n > \max(4096, 2C + \frac{7}{2})$ and $\frac{n}{\log n} > \frac{3(1/2 + C)C}{C/2 - 1/4}$. If K_n^t is the t-step transition kernel for the Gibbs sampler described above, and U_n is the uniform distribution on Δ_n, then for all $t > 10Cn \log n$, $x \in \Delta_n$ and $A \subset \Delta_n$ measurable,*

$$|K_n^t(x, A) - U_n(A)| < n^{3-C} + 2n^{-C/2-1/4} + 4n^{11/4-C}.$$

On the other hand, for $0 < C < 1$ and $t < (1 - C)n \log n$,

$$\liminf_{n \to \infty} \sup_{x \in \Delta_n} \sup_{A \subset \Delta_n} |K_n^t(x, A) - U_n(A)| = 1.$$

The conditions on the constant C are not onerous. Choosing $C = 4$ gives a mixing time of at most $40n \log n$ that is effective for $n > 4096$.

Sections 2–4 are devoted to proving the upper bound of Theorem 1.1, and Section 5 proves the lower bound. In Section 6, we briefly discuss applications of our method to closely related Markov chains. In Section 7, we use the ideas of the proof to develop a perfect sampling algorithm with wider applicability.

2. Notation, basic lemmas and strategy. 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. The proof relies on the following standard lemma (see [11], Theorem 5.2—they work in discrete space, but their proof does not rely on this assumption):

LEMMA 2.1 (Fundamental coupling lemma). *Assume* (X_t, Y_t) *is a coupling of Markov chains such that if* $X_s = Y_s$, *then* $X_t = Y_t$ *for all* $t > s$. *Assume also that* $X_0 = x$ *and* Y_0 *is distributed according to the stationary distribution of* K. *Define the random time* τ *to be the first time at which* $X_t = Y_t$. *Then* $\sup_{A \in \Sigma} |K^t(x, A) - \pi(A)| \leq P[\tau > t]$.

Throughout this note, we are interested in a coupling of Markov chains (X_t, Y_t), where X_0 starts according to some distribution of our choosing, Y_0 starts out uniformly over the simplex and both marginally evolve as the Gibbs sampler being studied. We will describe a joint evolution of our two chains X_t and Y_t, such that at a specific time, the probability of having coupled is very high. The method for proving this is slightly unusual. In most coupling proofs, including the non-Markovian coupling in [7], there is an attempt to make the two chains get closer throughout the process. In our method, we attempt to couple only at a specific final time, and include many moves that are likely to increase the distance between the chains by a large amount. In fact, our joint distribution will generally assign 0 probability to coupling at any prior time.

In order to develop our global joint coupling, we describe two possible one-step couplings of X_t and Y_t. These are the "proportional" coupling and the "subset" coupling. Throughout, we will always choose to update entries at the same coordinates i, j in both X_t and Y_t at every step; only the uniform variable λ used in representation (1.1) sometimes differs. Because of this, we often describe the couplings by describing only how the update variables λ are coupled.

In the proportional coupling, we choose an i, j and λ for Y_t, and then use the same choices for X_t in representation (1.1), so that, for example, entry i in Y_t is updated to $\lambda(Y_t[i] + Y_t[j])$ while entry i in X_t is updated to $\lambda(X_t[i] + X_t[j])$. The subset coupling is slightly more complicated. As before, we choose two coordinates i, j to be updated in both chains. Next, define the weight $w(S, X)$ that a vector X gives to a subset $S \subset [n]$ to be $w(S, X) = \sum_{s \in S} X[s]$. A subset coupling of X_t and Y_t will always be with respect to some specific subset $S \subset [n] = \{1, 2, \ldots, n\}$. If either $i, j \in S$ or $i, j \notin S$, perform a proportional coupling. Otherwise, assume without loss of generality that $i \in S$ and $j \notin S$ and also that $X_t[i] + X_t[j] \geq Y_t[i] + Y_t[j]$. In this case, call a coupling of X_{t+1} and Y_{t+1} conditioned on X_t, Y_t, i and j a subset coupling if

$$P\big[w(X_{t+1}, S) = w(Y_{t+1}, S)\big] \geq \frac{Y_t[i] + Y_t[j] - |\sum_{k \in S/\{i\}}(Y_t[k] - X_t[k])|}{X_t[i] + X_t[j]}.$$

We will say a subset coupling has succeeded if $w(X_{t+1}, S) = w(Y_{t+1}, S)$, and that it has failed otherwise. We will generally not be concerned with what happens when a subset coupling has failed. We will check now that such a coupling exists. Note that, conditioned on X_t and the coordinates i, j updated at time t, the weight $w(S, X_{t+1})$ is uniformly distributed on $[\sum_{k \in S/\{i\}} X_t[k], \sum_{k \in S/\{i\}} X_t[k] + X_t[i] + X_t[j]]$. Similarly, conditioned on Y_t, i and j, $w(S, Y_{t+1})$ is uniformly distributed on $[\sum_{k \in S/\{i\}} X_t[k], \sum_{k \in S/\{i\}} X_t[k] + X_t[i] + X_t[j]]$.

LEMMA 2.2 (Total variation distance of two uniform distributions). *Let U be distributed uniformly on $[a, a + b]$ and let U' be distributed uniformly on $[a', a' + b']$. Assume $b \le b'$. Then $\|\mathcal{L}(U) - \mathcal{L}(U')\|_{\mathrm{TV}} \le 1 - \frac{b - |a - a'|}{b'}$.*

PROOF. Note that U has density $f(x) = \frac{1}{b}\mathbf{1}_{x \in [a, a+b]}$, and U' has density $g(x) = \frac{1}{b'}\mathbf{1}_{x \in [a', a'+b']}$. Thus, the total variation distance between them is given by

$$
\begin{aligned}
\|\mathcal{L}(U) - \mathcal{L}(U')\|_{\mathrm{TV}} &= 1 - \int_x \min(f(x), g(x))\, dx \\
&= 1 - \frac{1}{b'} \int_{x \in [a, a+b] \cap [a', a'+b']} 1\, dx \\
&= 1 - \frac{1}{b'}[\min(a + b, a' + b') - \max(a, a')] \\
&\le 1 - \frac{1}{b'}[b + \min(a, a') - \max(a, a')] \\
&= 1 - \frac{b - |a - a'|}{b'}.
\end{aligned}
$$
□

Since it is always possible to couple two random variables W, Z so that $P[Z = W] = 1 - \|\mathcal{L}(Z) - \mathcal{L}(W)\|_{\mathrm{TV}}$, Lemma 2.2 implies that subset couplings exist.

We now give a rough and nonrigorous description of the proof strategy, which proceeds by describing a two-step coupling of X_t and Y_t. For the first T_1 steps, X_t and Y_t evolve always under the proportional coupling. This coupling is Markovian, and we prove that under this coupling, the two chains are very close in sup-norm with high probability after about $n \log n$ steps. In the next phase, we record the updated coordinates $(i(t), j(t))$ from time T_1 until a specified time $T = T_1 + T_2$. This information is used to construct a nested sequence P_t of partitions of the set of coordinates $[n]$. With high probability, the sequence will satisfy $P_{T_1} = \{[n]\}$ and $P_{T_1 + T_2} = \{\{1\}, \{2\}, \ldots, \{n\}\}$. We will then couple X_t to Y_t step by step, using only information about the future that is contained in P_t, using a proportional coupling for some steps and a subset coupling for others. We then show that it is possible to keep $w(S, X_t) = w(S, Y_t)$ for all $S \in P_t$ with high probability. If all of these

high-probability events occur, then the final partition consists of only singletons, and this implies that $X_T[i] = Y_T[i]$ for $1 \leq i \leq n$. The two main difficulties are constructing the partition and showing that X_t and Y_t remain close throughout the second phase.

It is worth pointing out that the dependence of the coupling on the future is in fact necessary to get the correct mixing time, or indeed any bound that is $o(n^2)$. This is analogous to the well-known fact that no Markovian coupling of the random transposition walk on S_n can give a coupling time that is $o(n^2)$. See Lemma 8 of [2] for a short proof of this fact for the walk on S_n which applies essentially without modification to this Gibbs sampler.

Here is a list of some commonly used variables that have been reserved, for reference while reading:

X_t, the Markov chain of interest.
Y_t, another instance of the Markov chain, started at stationarity.
P_t, a set partition of $[n]$.
S, a piece of a partition.
i, j, coordinates we update.
$\lambda, \lambda_x, \lambda_y$, uniform random variable used to update a chain, or chains X_t and Y_t.
$w(S, X)$, the weight assigned by vector X to a subset $S \subset [n]$.

3. First coupling stage. Define $Z_t = \|X_t - Y_t\|_2^2$. The following provides an upper bound for $E[Z_t]$ under the proportional coupling described above:

LEMMA 3.1 (Burn-in). *Let X_t and Y_t be two copies of the Markov chain coupled by the proportional coupling, and Z_t defined as above. After $s \geq \frac{3}{2}dn\log n$ steps of the proportional coupling, $E[Z_s] \leq 2n^{-d}$.*

PROOF. The proof is by a one-step contraction estimate. Assume X_t and Y_t are coupled by the proportional coupling from time 0 onwards. Let $\mathcal{F}_t$ be the σ-algebra generated by the random variables X_t and Y_t; note that Z_t is $\mathcal{F}_t$-measurable. Then, under the proportional coupling, the following equality comes from conditioning on the coordinates $(i, j) = (i(t), j(t))$ updated at time t,

$$
E[Z_{t+1}|\mathcal{F}_t]
$$
$$
= \frac{1}{n(n-1)} \sum_{i \neq j} E\Big[\lambda^2 \big(X_t[i] + X_t[j] - Y_t[i] - Y_t[j]\big)^2
$$
$$
+ (1-\lambda)^2 \big(X_t[i] + X_t[j] - Y_t[i] - Y_t[j]\big)^2
$$
$$
+ \sum_{k \neq i, j} \big(X_t[k] - Y_t[k]\big)^2\Big].
$$

Note $E[\lambda^2] = E[(1-\lambda)^2] = \frac{1}{3}$. Expanding the above, we obtain

$$E[Z_{t+1}|\mathcal{F}_t]$$
$$= \frac{1}{n(n-1)} \sum_{i \neq j} \left[\frac{2}{3}(X_t[i] - Y_t[i])^2 + \frac{2}{3}(X_t[j] - Y_t[j])^2 \right.$$

$$+ \frac{4}{3}(X_t[i] - Y_t[i])(X_t[j] - Y_t[j])$$

$$\left. + \sum_{k \neq i,j} (X_t[k] - Y_t[k])^2 \right].$$

Collecting coefficients of Z_t and using the fact that $Z_t = \sum_k (X_t[k] - Y_t[k])^2$, this equals

$$\left(1 - \frac{2}{3n}\right) Z_t + \frac{4}{3n(n-1)} \sum_{i \neq j} (X_t[i] - Y_t[i])(X_t[j] - Y_t[j]).$$

Noting that $\sum_{i=1}^n (X_t[i] - Y_t[i]) = 0$, the last term can be rewritten as

$$\sum_{i \neq j} (X_t[i] - Y_t[i])(X_t[j] - Y_t[j]) = \left(\sum_{i=1}^n (X_t[i] - Y_t[i]) \right)^2 - \sum_{i=1}^n (X_t[i] - Y_t[i])^2$$

$$= -Z_t.$$

Putting this together, we find that

$$E[Z_{t+1}|\mathcal{F}_t] = \left(1 - \frac{2}{3n} - \frac{4}{3n(n-1)} \right) Z_t.$$

And so in particular,

$$E[Z_t|\mathcal{F}_0] = E\left[E[Z_t|\mathcal{F}_{t-1}] | \mathcal{F}_0 \right]$$

$$\leq \left(1 - \frac{2}{3n} \right) E[Z_{t-1}|\mathcal{F}_0].$$

By induction on t, it is then easy to see that

$$E[Z_t|\mathcal{F}_0] \leq \left(1 - \frac{2}{3n} \right)^t Z_0.$$

Bound Z_0 by

$$Z_0 \leq \sum_k (X_0[k]^2 + Y_0[k]^2)$$

$$\leq \sum_k X_0[k] + Y_0[k]$$

$$= 2.$$

We conclude that at times $t \geq \frac{3}{2} dn \log n$, $E[Z_t] \leq 2n^{-d}$. □

Using the obvious inequality $|X_t[i] - Y_t[i]| \leq \sqrt{Z_t}$ and Markov's inequality, $P[|X_t[i] - Y_t[i]| > \delta] \leq \delta^{-1} n^{-d/2}$ for all $\delta > 0$ and $i \in [n]$.

4. Second coupling stage. Let $T = (\frac{1}{2} + \varepsilon) n \log n$ be fixed, for some $\varepsilon > 0$ to be decided later. Let Y_0 be chosen from the uniform distribution on the simplex, and let X_0 satisfy $\|X_0 - Y_0\|_1 \leq n^{-d}$. We describe a coupling (X_t, Y_t) from time 0 to time T with the property that $X_T = Y_T$ with high probability as n goes to infinity, for any fixed $\varepsilon > 0$ and d sufficiently large. First, we choose a sequence of pairs of distinct elements $1 \leq i(t) \neq j(t) \leq n$ independently and uniformly for times $0 \leq t \leq T$. These pairs $(i(t), j(t))$ will be the coordinates updated at time t in both X_t and Y_t. Then define a sequence of graphs G_t for $0 \leq t \leq T - 1$ to have vertex set $[n]$ and edge set $E_t = \{(i(t), j(t)), (i(t+1), j(t+1)), \ldots, (i(T-1), j(T-1))\}$, throwing out repeated edges, if any. We also define G_T to be the graph on $[n]$ with no edges. From this sequence, construct a sequence of partitions of $[n]$, $P(0), P(1), \ldots, P(T)$ by letting the sets in P_t be exactly the connected components of G_t.

Since the edges satisfy $E_s \subset E_t$ for every $s > t$, it is clear that for any $A \in P_s$, there must be some $B \in P_t$ with $A \subset B$. In this sense, the sequence of partitions is nested. Also note from the construction that either P_t and P_{t+1} are the same, or they differ by having a single set in P_t split into two sets in P_{t+1}. Define the sequence of marked time $0 \leq t_1 < \cdots < t_k = T - 1$ as the times at which $P_{t_\ell} \neq P_{t_\ell+1}$. Then, for marked time t_ℓ, define $S(t_\ell, 1)$ and $S(t_\ell, 2)$ to be the two sets that were split apart at time t_ℓ, labeled so that $|S(t_\ell, 1)| \leq |S(t_\ell, 2)|$. Note that there are at most $n - 1$ marked times, and that there are $n - 1$ if and only if $P_0 = [n]$.

Note also that $P_0 = [n]$ if and only if G_0 is connected. The question of whether or not the random graph G_0 is connected is a classical question in random graph theory. The following result, found in [3] among other places, is good enough for our purposes:

LEMMA 4.1 (Connectedness for Erdos–Renyi graphs). *Let $\varepsilon > 0$ be fixed, and let $T = T_\varepsilon$ be the first time that $(\frac{1}{2} + \varepsilon) n \log n$ distinct edges have been chosen. Then the probability that G_0 is connected is at least $1 - n^{-\varepsilon}$.*

This has the immediate corollary:

LEMMA 4.2 (Connectedness for G_0). *Let $\varepsilon > 0$, and assume $n > 4$ satisfies $\frac{n}{\log n} > \frac{3(1+2\varepsilon)(1/2+2\varepsilon)}{\varepsilon}$. Then let $T > (\frac{1}{2} + 2\varepsilon) n \log n$. Then the probability that G_0 is connected is at least $1 - 2n^{-\varepsilon}$.*

PROOF. Ignoring the ordering of vertices in edges, define
$$A_t = \mathbf{1}_{(i(t), j(t)) \notin \{(i(0), j(0)), \ldots, (i(t-1), j(t-1))\}} \mathbf{1}_{t-1 < T_\varepsilon}.$$

Note that $P[A_t = 1 | A_1, \dots, A_{t-1}] \leq \frac{(1+2\varepsilon)\log n}{n-1}$. In particular, if B has binomial $((\frac{1}{2} + 2\varepsilon)n \log n, \frac{2(1/2+\varepsilon)\log n}{n-1})$ distribution, then $P[\sum_{s=0}^{(1/2+2\varepsilon)n\log n} A_s > x] \leq P[B > x]$ for all $x > 0$. For n satisfying $\frac{n}{\log n} > \frac{3\varepsilon}{(1+2\varepsilon)(1/2+2\varepsilon)}$, Chernoff's inequality gives the bound

$$P[T_\varepsilon > (\tfrac{1}{2} + 2\varepsilon)n\log n] \leq P[B > \varepsilon n \log n]$$

$$\leq e^{-n\varepsilon/2},$$

which is less than $n^{-\varepsilon}$ for $n \geq 4$. Let $\mathcal{E}_T$ be the event that G_0 is disconnected. Since $P[G_0] \leq P[T_\varepsilon > T] + P[E_T | T_\varepsilon < T]$, the result follows immediately from this bound on T_ε and Lemma 4.2. $\square$

Having constructed this partition, we now couple X_t and Y_t for time $0 \leq t \leq T$. First, we need to choose the coordinates to update; we do this by updating coordinates $i(t)$ and $j(t)$ at time t in both chains. Next, we must describe the coupling of the coordinates. If t is a marked time, then perform a subset coupling for the set $S(t, 1)$. Otherwise, do a proportional coupling. Note that if t is a marked time, then one of $i(t)$ or $j(t)$ is in $S(t, 1)$ and the other is in $S(t, 2)$, so the coupling proceeds according to the description in Section 2 in the case $i \in S$, $j \in S^c$. We claim that this couples the two walks by time T with high probability:

LEMMA 4.3 (Coupling for close chains). *For $\varepsilon > 0$, $d > \frac{11}{2}$, $n > \max(d + \frac{3}{2}, 4096)$, $\frac{n}{\log n} > \frac{3(1+2\varepsilon)(1/2+2\varepsilon)}{\varepsilon}$ and $T > (\frac{1}{2} + 2\varepsilon)n\log n$, the coupling described in this section has the property*

$$P[X_T \neq Y_T] \leq 2n^{-\varepsilon} + 5n^{(15-2d)/4}.$$

We begin by showing that subset couplings succeed with high probability:

LEMMA 4.4 (Subset coupling). *Assume $n \geq 6$, and let (X_t, Y_t) be a pair of elements of Δ_n satisfying $\sup_k |X_t[k] - Y_t[k]| = n^{-f}$ and $\inf_k X_t[k], \inf_k Y_t[k] \geq 2n^{-b}$, with $f \geq b + 1$. Then for all $S \subset [n]$ and update coordinates $i \in S$, $j \notin S$, $P[w(X_{t+1}, S) = w(Y_{t+1}, S)] \geq 1 - 3n^{b+1-f}$ under the subset coupling.*

PROOF. Assume that $X_t[i] + X_t[j] \geq Y_t[i] + Y_t[j]$. Then, from its definition, the subset coupling succeeds with probability at least

$$P[w(X_{t+1}, S) = w(Y_{t+1}, S)]$$

$$\geq \frac{Y_t[i] + Y_t[j] - |\sum_{k \in S/\{i\}}(Y_t[k] - X_t[k])|}{X_t[i] + X_t[j]}$$

$$\geq \frac{Y_t[i] + Y_t[j] - 2|S|n^{-f}}{Y_t[i] + Y_t[j] + 4n^{-f}}$$

$$\geq (1 - 2n^{1-f+b})(1 - 4n^{-f+b} - 8n^{-2f+2b}),$$

which, for $n \geq 6$, is at least $1 - 3n^{b+1-f}$. $\square$

Having bounded the probability of failure when X_t, Y_t are close, we must show that they remain close as long as all subset couplings succeed. For $S \subset [n]$, define $\|X\|_S = \sum_{s \in S} |X[s]|$. Then:

LEMMA 4.5 (Closeness).   *Let X_t, Y_t be coupled as described above, and assume that $P_0 = \{[n]\}$, that all subset couplings up to time t have succeeded and that $\|X_0 - Y_0\|_1 < \varepsilon$. Then $\|X_t - Y_t\|_S < \varepsilon$ for every S in P_t*

PROOF.   There are two types of coupling to take care of. For a proportional coupling with coordinates i and j,

$$
\begin{aligned}
&|X_{t+1}[i] - Y_{t+1}[i]| + |X_{t+1}[j] - Y_{t+1}[j]| \\
&= \lambda_t |X_t[i] + X_t[j] - Y_t[i] - Y_t[j]| \\
&\quad + (1 - \lambda_t)|X_t[i] + X_t[j] - Y_t[i] - Y_t[j]| \\
&\leq |X_t[i] - Y_t[i]| + |X_t[j] - Y_t[j]|.
\end{aligned}
$$

Since i and j always connect elements of the same set in P_t, this shows that proportional couplings never increase $\|X_t - Y_t\|_S$. Otherwise, assume that at time t we had a successful subset coupling for the subset S along edges i and j. Without loss of generality, assume that $i \in S := S(t, 1)$ and $j \in R := S(t, 2)$. Since $w(X_0, [n]) = w(Y_0, [n]) = 1$, and all subset couplings up to time t have succeeded, we have $w(X_t, Q) = w(Y_t, Q)$ for all $Q \in P_t$. In particular, $w(X_t, S \cup R) = w(Y_t, S \cup R)$. Then we note that

$$
\begin{aligned}
X_{t+1}[i] - Y_{t+1}[i] &= \sum_{s \in S \setminus \{i\}} (Y_t[s] - X_t[s]) \\
&= X_t[i] - Y_t[i] + \sum_{s \in R} (X_t[s] - Y_t[s])
\end{aligned}
$$

and so

$$
|X_{t+1}[i] - Y_{t+1}[i]| \leq |X_t[i] - Y_t[i]| + \|X_t - Y_t\|_R,
$$

which immediately implies that

$$
\|X_{t+1} - Y_{t+1}\|_S \leq \|X_t - Y_t\|_{S \cup R}.
$$

An analogous calculation shows that

$$
\|X_{t+1} - Y_{t+1}\|_R \leq \|X_t - Y_t\|_{R \cup S}
$$

as well. By induction on t, this implies that $\|X_{t+1} - Y_{t+1}\|_S \leq \|X_0 - Y_0\|_1$ and $\|X_{t+1} - Y_{t+1}\|_R \leq \|X_0 - Y_0\|_1$. $\square$

LEMMA 4.6 (Largeness). $P[\inf_{1 \leq i \leq n} \inf_{0 \leq t \leq n^2 - 1} Y_t[i] \leq n^{-4.5 - k}] \leq 2n^{-k}$ for $n > \max(2k, 4096)$.

PROOF. Let $q_1, \ldots, q_n$ be independent random variables chosen from the exponential distribution with mean 1, and let $Q = \sum_{i=1}^{n} q_n$. It is well known (see, e.g., Algorithm 2.7.1 of [21]) that $(\frac{q_1}{Q}, \ldots, \frac{q_n}{Q})$ is distributed uniformly on the simplex Δ_n. In particular, $Y_t \overset{D}{=} (\frac{q_1}{Q}, \ldots, \frac{q_n}{Q})$. Taking a union bound over $1 \leq i \leq n$ and $0 \leq t \leq n^2 - 1$, it is thus sufficient to show

$$P\left[\frac{q_1}{Q} \leq n^{-1.5 - k}\right] \leq 2n^{-k}.$$

Let E be the event that $\frac{q_1}{Q} < n^{-1.5 - k}$, E_1 the event that $q_1 < n^{-k - 0.25}$, and E_2 the event that $Q > n^{1.25}$, and observe that $E \subset E_1 \cup E_2$. It is immediate that

$$P[E_1] = 1 - e^{-n^{-k - 0.25}}$$
$$\leq n^{-k - 0.25} + \tfrac{1}{2} n^{-2k - 0.5}.$$

For $n > 4096$, this is certainly less than n^{-k}. To bound the probability that Q is large, note that for all $0 < \theta < 1$,

$$E[e^{\theta Q}] = E[e^{\theta q_1}]^n$$
$$= \frac{1}{(1 - \theta)^n}.$$

Setting $\theta = 1 - n^{-0.25}$, Markov's inequality gives

$$P[E_2] \leq e^{(1/4)n \log n + n - n^{1.25}}.$$

It is straightforward to check that, for $n > \max(2k, 4096)$, this is less than n^{-k}. Since $P[E] \leq P[E_1] + P[E_2]$, this proves the lemma. $\square$

Finally, it is possible to prove that in fact most couplings will succeed:

LEMMA 4.7 (Weight lemma). Fix $d > \frac{11}{2}$ and $n > \max(k + \frac{3}{2}, 4096)$. Assume $P_0 = \{[n]\}$ and that $\|X_0 - Y_0\|_1 \leq n^{-d}$. Let E be the event that the equality

(4.1) $$w(S, X_t) = w(S, Y_t)$$

holds for all $0 \leq t \leq T$ and all $S \subset P(t)$. Then

$$P[E] \geq 1 - 5n^{(15 - 2d)/4}.$$

PROOF. The equality (4.1) clearly holds at time 0. Also note that if it holds at an unmarked time t, it must also hold at time $t + 1$, since at unmarked times the weights of parts S of the partition P_t cannot change in either X_t or Y_t. So,

assume that equality (4.1) holds for all times $t \leq t_k$ for some marked time t_k. If the subset coupling is successful at time t_k, then $w(S(t_k + 1, 1), X_{t_k+1}) = w(S(t_k + 1, 1), Y_{t_k+1})$ by construction. However, by the assumption that equality (4.1) holds until time t_k, $w(S(t_k, 1) \cup S(t_k, 2), X_{t_k}) = w(S(t_k, 1) \cup S(t_k, 2), Y_{t_k})$. Since $w(A \cup B, X) = w(A, X) + w(B, X)$ for any disjoint sets A, B and any vector X, this implies $w(S(t_k + 12), X_{t_k+1}) = w(S(t_k + 1, 2), Y_{t_k+1})$ as well. Since none of the other parts of P_{t_k} change weight, this implies that $w(S, X_{t_k+1}) = w(S, Y_{t_k+1})$ holds for all $S \in P_{t_k+1}$.

It remains to bound only the probability that the first subset coupling to fail occurs at time t_k. By Lemma 4.3 and the assumption of this lemma,

$$\|X_{t_k} - Y_{t_k}\|_1 \leq \sum_{S \in P_{t_k}} \|X_{t_k} - Y_{t_k}\|_S$$

(4.2)
$$\leq \sum_{S \in P_{t_k}} n^{-d}$$

$$\leq n^{1-d}.$$

Set $q = \frac{d}{2} + \frac{3}{4}$. By Lemma 4.6, $\inf_{i,t} Y_t[i] \geq n^{-q}$ with probability at least $1 - 2n^{4.5-q}$. Assuming this holds, Lemma 4.4 along with inequality (4.2) implies that any particular subset coupling succeeds with probability at least $1 - 3n^{2+q-d}$. Taking a union bound over all at most $n - 1$ subset couplings, all subset couplings succeed with probability at least $1 - 3n^{3+q-d} - 2n^{4.5-q} = 1 - 5n^{(15-2d)/4}$. $\square$

It is now time to prove Lemma 4.2. Recall that if $P_0 = [n]$ and all components $Q \in P_t$ satisfy $w(Q, X_t) = w(Q, Y_t)$ for $0 \leq t \leq T$, then at time T the two walks have coupled. There are only two ways for this to fail to happen. The first is the event E_1 that $P_0 \neq [n]$. By Lemma 4.2, $P[E_1] \leq 2n^{-\varepsilon}$. The second is the event E_2 that at least one subset coupling fails. By Lemma 4.7 and our assumption that $\frac{n}{\log n} > \frac{1}{2} + 2\varepsilon$, which implies $T < n^2$, we have the bound $P[E_2] \leq 5n^{(15-2d)/4}$. Combining these two bounds proves the lemma.

Finally, we prove Theorem 1.1. We will run the proportional coupling until time $T_1 = 9Cn \log n$, and then we will run the second phase coupling from time T_1 until time $T = 10Cn \log n$. There are only two ways to have $X_T \neq Y_T$. The first is the event E_1 that $\|X_{T_1} - Y_{T_1}\|_1 > n^{-(2C+2)}$. By the comment immediately after Lemma 3.1, $P[E_1] \leq n^{3-C}$. The second is the event E_2 that the second phase coupling fails. By Lemma 4.3, $P[E_2] \leq 2n^{C/2-1/4} + 4n^{11/4-C}$. Combining these two bounds proves the theorem.

We also note that it is possible to improve the top of the pre-cutoff window from 30 to 12 by being more careful in the above proofs, but there is no hope of actually proving a cutoff without a substantially new argument.

5. Lower bound. Since our walk is over a continuous space, the total variation distance to stationarity of the Markov chain at time t must be at least the probability that not all coordinates have been chosen by time t. Since only two coordinates are chosen at a time, the classical coupon-collector results in [6] tell us that at time $T = \frac{1}{2}n(\log n - c)$, $\sup_{A \in \Sigma} |K_n^T(x, A) - \pi(A)| \geq 1 - \exp(-\exp(c)) + o(1)$ as n goes to infinity.

It is possible to improve the constant a little bit. Let $X_0 = (1, 0, \ldots, 0)$, and let $Q_t \in \{0, 1\}^n$ be a vector keeping track of updates in X_t, started at $Q_0 = (0, 0, \ldots, 0)$. If coordinates i and j are updated in X_t at time t, set $Q_{t+1}[i] = Q_{t+1}[j] = 1$ if at least one of $X_t[i]$, $X_t[j]$ are nonzero, and set $Q_{t+1}[k] = Q_t[k]$ for all $k \neq i, j$. If $X_t[i] = X_t[j] = 0$, then set $Q_{t+1} = Q_t$. Next, let $\tau_j = \inf\{t | Q_{t+1} \neq Q_t, t > \tau_{j-1}\}$ with $\tau_0 = 0$. We note that $E[\tau_1] = \frac{n}{2}$, and for $j > 1$, $E[\tau_j] = \frac{n(n-1)}{2j(n-j)}$. Thus, letting $\tau = \sum_{j=1}^{n-1} \tau_j$,

$$E[\tau] = \frac{n}{2} + \frac{n^2}{2} \sum_{j=2}^{n-1} \frac{1}{j(n-j)}$$

$$= n \log n + O(n).$$

Similarly, since τ_i and τ_j are independent for $i \neq j$, it is easy to calculate that the variance $V[\tau] \leq 6n^2$. By Chebyshev's inequality, for all $\varepsilon > 0$ and n sufficiently large,

$$(5.1) \qquad P[\tau < (1 - \varepsilon)n \log n] = O\left(\frac{1}{\log n^2}\right).$$

Finally, observe that for $t < \tau$, at least one entry of X_t is 0, and so taking $H_j = \{X \in \Delta_n | X[j] = 0\}$ and $A \in \Sigma$ to be $\bigcup_j H_j$, we find $|K_n^T((1, 0, \ldots, 0), A) - U_n(A)| \geq P[T < \tau]$. Combined with inequality (5.1), this proves the lower bound on the mixing time.

6. Closely related walks. It is worth pointing out a small number of cases where the above argument goes through with very few changes. The first allows us to go from sampling from the uniform distribution to sampling from a large class of distributions on the simplex, including symmetric Dirichlet distributions. At each step of the random walk, instead of choosing λ according to the uniform distribution on $[0, 1]$, choose it according to some other distribution with twice differentiable cdf F satisfying $F[x] = 1 - F[1 - x]$ for all $0 \leq x \leq \frac{1}{2}$. Then the above arguments show that the total mixing time is $O(n \log n \frac{\|F''\|_\infty + 1}{1 - 2E[\lambda^2]})$, essentially without modifcation.

It is also possible to apply this argument to the discrete analogue of the simplex, in which M indistinguishable balls are stored in n boxes; these are known as M-compositions of n. The analogous Markov chain involves choosing two boxes, holding N balls between them, at every step, and putting $0 \leq k \leq N$ of them in

the first box with probability $\frac{1}{N+1}$, and the remainder in the second box. The arguments given above apply to the discrete chain, giving a mixing bound of order $O(n \log n)$, but there need to be enough balls for the continuous approximation to be good at each step. A straightforward step-through of the argument gives a bound of $O(n \log n)$ for $M > n^{18.5}$ above. Aldous' greedy argument, which gives an upper bound of $O(n^2 \log n)$, holds for $M > n^{5.5}$.

The follow-up paper [22] will discuss a wider variety of related walks, requiring larger modifcations.

7. Perfect sampling on the simplex. In this section, we discuss how the two-chain coupling described above can be modified into a grand coupling, and how to use this fact to create a perfect sampling algorithm. Before describing the algorithm, we mention that it is not a practical way to obtain uniform points on the simplex. However, the same algorithm can be used to obtain samples from the other distributions on the simplex mentioned in Section 6, many of which are a priori much harder to sample from. The method is also of some interest as a relatively rare instance of a coupling from the past (CFTP) algorithm which does not use monotonicity or anti-monotonicity.

To begin, we recall the CFTP algorithm, described in greater detail in [15]. First, choose some large time T, and start a copy of the Markov chain X^{ω}_{-T} for each ω in the sample space Ω. Next, couple all of the chains from time $-T$ to time 0. If the chains have coalesced by time 0, the resulting single value is distributed according to the stationary distribution of the chain. If not, we couple chains started at all points from $-2T$ to T and keep the evolution from $-T$ to 0, then from $-3T$ to $-2T$ keeping the evolution from $-2T$ to 0, and so on until coalescence at 0 has occurred.

For Markov chains on a finite state space, it is easy in theory to construct a grand coupling that will eventually coalesce, though bad couplings are very inefficient. In practice, even on finite chains, CFTP is only used if the chain has some very special properties. The most popular properties are monotonicity and its twin anti-monotonicity. Briefly, we introduce a partial order $\leq$ on Ω, and say that a coupling of two chains X_t, Y_t is monotone if $X_0 \leq Y_0$ implies $X_t \leq Y_t$ for all $t > 0$. It is then easy to see that if our grand coupling is monotone, it is sufficient to keep track of chains started at maximal and minimal elements of the poset. If they have coupled, all states have coupled.

For Markov chains on infinite state spaces, many grand couplings will never coalesce, and of course we cannot keep track of all of the starting values on a computer. Some chains have a monotonicity property, but such a property is not obvious for the simplex model. Despite this, there is a fairly efficient perfect sampling algorithm that requires tracking only $n + 1$ points (and a little extra overhead each time an epoch of length T fails to coalesce).

Let X^v_t be a copy of the Markov chain started at $v = (v[1], v[2], \ldots, v[n])$ at time 0, and let e_j be the jth standard unit basis vector. We construct a grand

coupling of the chains X_t^v as follows. For time $0 < t < T_1$, do a proportional coupling. That is, at each time t, choose coordinates $i(t)$, $j(t)$ and parameter $\lambda(t)$, and update all chains using these three numbers. We claim that for each t, there exists a matrix $M_t[i, j]$ such that for any v, $X_t^v[i] = \sum_{j=1}^{n} M_t[i, j]v[j]$. To see this, observe that $X_{t+1} = M_{i(t), j(t), \lambda(t)} X_t$, where $M_{i(t), j(t), \lambda(t)}[i(t), i(t)] = M_{i(t), j(t), \lambda(t)}[i(t), j(t)] = \lambda(t)$, $M_{i(t), j(t), \lambda(t)}[j(t), i(t)] = M_{i(t), j(t), \lambda(t)}[j(t), j(t)] = (1 - \lambda(t))$, $M_{i(t), j(t), \lambda(t)}[k, k] = 1$ for $k \notin \{i(t), j(t)\}$ and all other entries are 0. We can then write $M_t[i, j] = \prod_{s<t} M_{i(s), j(s), \lambda(s)}$. Thus, for $v, w \in \Delta_n$,

$$
\begin{aligned}
\|X_t^v - X_t^w\|_1 &= \sum_{i=1}^{n} |X_t^v[i] - X_t^w[i]| \\
&= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} M_t[i, j](v[j] - w[j]) \right| \\
&\le \sum_{i=1}^{n} n \sup_j M_t[i, j] \\
&\le \sum_{i=1}^{n} n \sum_{j,k} |X_t^{e_j}[i] - X_t^{e_k}[i]|,
\end{aligned}
$$

which gives

$$
(7.1) \qquad \|X_t^v - X_t^w\|_1 \le n \sum_{j,k} \|X_t^{e_j} - X_t^{e_k}\|_1.
$$

Applying Lemma 3.1 to the expectation of $\|X_t^{e_j} - X_t^{e_k}\|_2$ for all distinct pairs j, k, using Markov's inequality to bound the probability that the L^2 norm is large, and finally noting that $\|X_t^{e_j} - X_t^{e_k}\|_1 \le n \|X_t^{e_j} - X_t^{e_k}\|_2$, we find that for $t > \frac{3}{2} dn \log n$,

$$
P[\|X_t^{e_j} - X_t^{e_k}\|_1 > n^{-k}] < 2n^{2k+1-d},
$$

and so taking a union bound and applying the inequality proved just above,

$$
P\left[\sup_{v, w \in \Delta_n} \|X_t^v - X_t^w\|_1 > n^{3-k} \right] < 2n^{2k+3-d}.
$$

This tells us that after $O(n \log n)$ steps, the L^1 distance between any pair of points is extremely small with high probability. The second step of the coupling is almost identical to the algorithm given in Section 4 of this note. Run $X_t^{(n^{-1}, \ldots, n^{-1})}$ from time T_1 to time T, recording all choices of $i(t)$, $j(t)$ and $\lambda(t)$ from representation (1.1). Then form the same partition process, and use it to attempt subset couplings of all variables to this special chain. We will perform these couplings in such a way that with high probability, all chains simultaneously have successful subset couplings, rather than merely having a high probability of a substantial fraction of the subset couplings succeeding.

At each subset coupling stage, use the update variable $\lambda(t)$ for the chain $X_t^{(n^{-1},\ldots,n^{-1})}$. For each other chain X_t^v, there will be some probability $p(t, v)$ that X_t^v performs a successful subset coupling with $X_t^{(n^{-1},\ldots,n^{-1})}$. Let p be a known lower bound on $\inf_{v \in \Delta_n} p(t, v)$. This can be obtained from Lemma 4.4 and inequality (7.1). To determine the update value of X_t^v, choose a single uniform random variable U. If $U < p$, let X_t^v have a successful subset coupling, in which case the change to X_{t+1}^v depends only on $i(t)$, $j(t)$ and $X_{t+1}^{(n^{-1},\ldots,n^{-1})}$, not the particular value of U. Otherwise, update with λ taken from the $\frac{U-p}{1-p}$'th quantile of the remainder distribution. When $\frac{X_t^v[i]+X_t^v[j]}{X_t^{(n^{-1},\ldots,n^{-1})}[i]+X_t^{(n^{-1},\ldots,n^{-1})}[j]} \leq 1$, this has density

$$f(\lambda) = C\left(1 - \frac{X_t^v[i] + X_t^v[j]}{X_t^{(n^{-1},\ldots,n^{-1})}[i] + X_t^{(n^{-1},\ldots,n^{-1})}[j]}\mathbf{1}_{g^{-1}(\lambda)\in[0,1]}\right)$$

for

$$g(\lambda) = \lambda\frac{X_t^{(n^{-1},\ldots,n^{-1})}[i] + X_t^{(n^{-1},\ldots,n^{-1})}[j]}{X_t^v[i] + X_t^v[j]}$$

$$+ \frac{1}{X_t^v[i] + X_t^v[j]} \sum_{s \in S\setminus\{i\}} (X_t^{(n^{-1},\ldots,n^{-1})}[s] - X_t^v[s])$$

and C a normalizing constant. An analogous formula holds when

$$\frac{X_t^v[i] + X_t^v[j]}{X_t^{(n^{-1},\ldots,n^{-1})}[i] + X_t^{(n^{-1},\ldots,n^{-1})}[j]} > 1.$$

Under this grand coupling, all subset couplings succeed together with probability at least p. As long as the n points $X_{T_1}^{e_j}$ are close as measured in L^1 metric, and $X_t^{(n^{-1},\ldots,n^{-1})}[i]$ remains far from 1 and 0, the proof of Lemma 4.4 tells us that all of the subset couplings succeed with high probability. Finally, if a single subset coupling fails at time t, then all chains should be coupled according to the proportional coupling for time $s > t$.

It remains to determine what to do if one of the above subset couplings fails. In order to obtain a perfect sample, it will be necessary to look at a grand coupling for the epoch $-2T \leq t \leq -T$. Assume for now that the grand coupling described above succeeds for the chain started at $-2T$. It is necessary to determine the value at time 0 of the chain started from $(\frac{1}{n}, \ldots, \frac{1}{n})$ at time $-2T$. Assume that at time $-T$, this chain is at $v \in \Delta_n$. Then our sample will be X_0^v. Fortunately, from the above description, it is possible to calculate this value from v and the values of i, j, λ and U used during the first epoch. Thus, it is sufficient to record those $O(T)$ pieces of information in each failed epoch. A longer discussion of this algorithm, with pseudocode, may be found in [23].

It should be noted that, for other target distributions on the simplex, such as those in Section 6, the above algorithm can also be used without a rigorous bound on the mixing time and can be used to rigorously check an estimated bound of time T. Simply run the algorithm with epoch size T; the number of failed runs k out of a total of N runs is distributed as a binomial random variable with some unknown probability q, where q is an upper bound on the total variation distance to stationarity at time T.

REFERENCES

[1] ALDOUS, D. and FILL, J. (1994). Reversible Markov chains and random walks on graphs Available at http://www.stat.berkeley.edu/~aldous/RWG/book.html.

[2] BLUMBERG, O. (2011). A coupling proof for random transpositions. Preprint.

[3] BOLLOBÁS, B. (2001). *Random Graphs*, 2nd ed. *Cambridge Studies in Advanced Mathematics* **73**. Cambridge Univ. Press, Cambridge. MR1864966

[4] BURTON, R. and KOVCHEGOV, Y. (2011). Mixing times via super-fast coupling. Preprint.

[5] DIACONIS, P. (2009). The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 179–205. MR2476411

[6] ERDÖS, P. and RÉNYI, A. (1961). On a classical problem of probability theory (in English). *Publ. Math. Inst. Hung. Acad. Sci.*, *Ser. A* **6** 215–220.

[7] HAYES, T. and VIGODA, E. (2003). A non-Markovian coupling for randomly sampling colorings. In *FOCS Proceedings*.

[8] JIANG, J. (2012). Total variation bound for Kac's random walk. Preprint.

[9] JIANG, J. (2012). Polynomial mixing time of the Kac random walk on the orthogonal group. Preprint.

[10] JONES, G. L. and HOBERT, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16** 312–334. MR1888447

[11] LEVIN, D. A., PERES, Y. and WILMER, E. L. (2009). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI. MR2466937

[12] LOVÁSZ, L. (1999). Hit-and-run mixes fast. *Math. Program.* **86** 443–461. MR1733749

[13] LOVÁSZ, L. and VEMPALA, S. (2003). Hit and run is fast and fun. Technical report, Microsoft Research.

[14] OLIVEIRA, R. I. (2009). On the convergence to equilibrium of Kac's random walk on matrices. *Ann. Appl. Probab.* **19** 1200–1231. MR2537204

[15] PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252.

[16] RANDALL, D. and WINKLER, P. (2005). Mixing points on an interval. In *Proceedings of ANALCO*. SIAM, Philadelphia.

[17] RANDALL, D. and WINKLER, P. (2005). Mixing points on a circle. In *Approximation*, *Randomization and Combinatorial Optimization*. *Lecture Notes in Computer Science* **3624** 426–435. Springer, Berlin. MR2193706

[18] ROSENTHAL, J. S. (1994). Random rotations: Characters and random walks on SO(N). *Ann. Probab.* **22** 398–423. MR1258882

[19] ROSENTHAL, J. S. (1995). On generalizing the cut-off phenomenon for random walks on groups. *Adv. in Appl. Math.* **16** 306–320. MR1342831

[20] ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90** 558–566. MR1340509

[21] RUBINSTEIN, R. Y. and MELAMED, B. (1998). *Modern Simulation and Modeling.* Wiley, New York. MR1607871

[22] SMITH, A. (2012). Analysis of convergence rates of some Gibbs samplers on continuous state spaces. Preprint.

[23] SMITH, A. (2012). Some analyses of Markov chains by the coupling method. Ph.D. thesis, Stanford Univ., Stanford, CA.

[24] YUEN, W. K. (2001). Applications of geometric bounds to convergence rates of Markov chains and Markov processes on R^n. Ph.D. thesis, Univ. Toronto. MR2702043

ICERM
BROWN UNIVERSITY
121 SOUTH MAIN STREET
PROVIDENCE, RHODE ISLAND 02912
USA
E-MAIL: asmith3@math.stanford.edu