Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much

Bryan He, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré

Stanford University

{bryanhe,cdesa,imit,chrismre}@stanford.edu

Abstract

Gibbs sampling is a Markov Chain Monte Carlo sampling technique that iteratively samples variables from their conditional distributions. There are two common scan orders for the variables: random scan and systematic scan. Due to the benefits of locality in hardware, systematic scan is commonly used, even though most statistical guarantees are only for random scan. While it has been conjectured that the mixing times of random scan and systematic scan do not differ by more than a logarithmic factor, we show by counterexample that this is not the case, and we prove that that the mixing times do not differ by more than a polynomial factor under mild conditions. To prove these relative bounds, we introduce a method of augmenting the state space to study systematic scan using conductance.

1 Introduction

Gibbs sampling, or Glauber dynamics, is a Markov chain Monte Carlo method that draws approximate samples from multivariate distributions that are difficult to sample directly [9; 15, p. 40]. A major use of Gibbs sampling is marginal inference: the estimation of the marginal distributions of some variables of interest [8]. Some applications include various computer vision tasks [9, 23, 24], information extraction [7], and latent Dirichlet allocation for topic modeling [11]. Gibbs sampling is simple to implement and quickly produces accurate samples for many models, so it is widely used and available in popular libraries such as OpenBUGS [16], FACTORIE [17], JAGS [18], and MADlib [14].

Algorithm 1 Gibbs sampler

---
input Variables x_i for $1 \leq i \leq n$, and target distribution π
   Initialize $x_1, \ldots, x_n$
  loop
     Select variable index s from $\{1, \ldots, n\}$
     Sample x_s from the conditional distribution $\mathbf{P}_\pi \left(X_s \mid X_{\{1,\ldots,n\} \setminus \{s\}} \right)$
  end loop

Gibbs sampling (Algorithm 1) iteratively selects a single variable and resamples it from its conditional distribution, given the other variables in the model. The method that selects the variable index to sample (s in Algorithm 1) is called the *scan order*. Two scan orders are commonly used: random scan and systematic scan (also known as deterministic or sequential scan). In random scan, the variable to sample is selected uniformly and independently at random at each iteration. In systematic scan, a fixed permutation is selected, and the variables are repeatedly selected in that order. The existence of these two distinct options raises an obvious question—which scan order produces accurate samples more quickly? This question has two components: hardware efficiency (how long does each iteration take?) and statistical efficiency (how many iterations are needed to produce an accurate sample?).

From the hardware efficiency perspective, systematic scans are clearly superior [21, 22]. Systematic scans have good spatial locality because they access the variables in linear order, which makes their iterations run faster on hardware. As a result, systematic scans are commonly used in practice.

Comparing the two scan orders is much more interesting from the perspective of statistical efficiency, which we focus on for the rest of this paper. Statistical efficiency is measured by the *mixing time*, which is the number of iterations needed to obtain an accurate sample [15, p. 55]. The mixing times of random scan and systematic scan have been

studied, and there is a longstanding conjecture [3; 15, p. 300] that systematic scan (1) never mixes more than a constant factor slower than random scan and (2) never mixes more than a logarithmic factor faster than random scan. This conjecture implies that the choice of scan order does not have a large effect on performance.

Recently, Roberts and Rosenthal [20] described a model in which systematic scan mixes more slowly than random scan by a polynomial factor; this disproves direction (1) of this conjecture. Independently, we constructed other models for which the scan order has a significant effect on mixing time. This raises the question: what are the true bounds on the difference between these mixing times? In this paper, we address this question and make the following contributions.

- In Section 3, we study the effect of the variable permutation chosen for systematic scan on the mixing time. In particular, in Section 3.1, we construct a model for which a systematic scan mixes a polynomial factor faster than random scan, disproving direction (2) of the conjecture, and in Section 3.2, we construct a model for which the systematic scan with the worst-case permutation results in a mixing time that is slower by a polynomial factor than both the best-case systematic scan permutation and random scan.

- In Section 4, we empirically verify the mixing times of the models we construct, and we analyze how the mixing time changes as a function of the permutation.

- In Section 5, we prove a weaker version of the conjecture described above, providing relative bounds on the mixing times of random and systematic scan. Specifically, under a mild condition, different scan orders can only change the mixing time by a polynomial factor. To obtain these bounds, we introduce a method of augmenting the state space of Gibbs sampling so that the method of conductance can be applied to analyze its dynamics.

2   Related Work

Recent work has made progress on analyzing the mixing time of Gibbs sampling, but there are still some major limitations to our understanding. In particular, most known results are only for specific models or are only for random scan. For example, mixing times are known for Mallow's model [1, 4], and colorings of a graph [5] for both random and systematic scan, but these are not applicable to general models. On the other hand, random scan has been shown to mix in polynomial time for models that satisfy structural conditions – such as having close-to-modular energy functions [10] or having bounded hierarchy width and factor weights [2] – but corresponding results for for systematic scan are not known. The major exception to these limitations is Dobrushin's condition, which guarantees $O(n \log n)$ mixing for both random scan and systematic scan [6, 13]. However, there are many models of interest with close-to-modular energy functions or bounded hierarchy width that do not satisfy Dobrushin's condition.

A similar choice of scan order appears in stochastic gradient descent (SGD), where the standard SGD algorithm uses random scan, and the incremental gradient method (IGM) uses systematic scan. In contrast to Gibbs sampling, avoiding "bad permutations" in the IGM is known to be important to ensure fast convergence [12, 19]. In this paper, we bring some intuition about the existence of bad permutations from SGD to Gibbs sampling.

3   Models in Which Scan Order Matters

Despite a lack of theoretical results regarding the effect of scan order on mixing times, it is generally believed that scan order only has a small effect on mixing time. In this section, we first define relevant terms and state some common conjectures regarding scan order. Afterwards, we give several counterexamples showing that the scan order can have asymptotic effects on the mixing time.

The *total variation distance* between two probability distributions μ and ν on Ω is [15, p. 47]

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|.$$

The *mixing time* is the minimum number of steps needed to guarantee that the total variation distance between the true and estimated distributions is below a given threshold ϵ from any starting distribution. Formally, the *mixing time* of a stochastic process P with transition matrix $P^{(t)}$ after t steps and stationary distribution π is [15, p. 55]

$$t_{\mathrm{mix}}(P, \epsilon) = \min \left\{ t : \max_{\mu} \|P^{(t)} \mu - \pi\|_{\mathrm{TV}} \leq \epsilon \right\},$$

Table 1: Models and Approximate Mixing Times

Model	$t_{\text{mix}}(R)$	$\min\limits_{\alpha} t_{\text{mix}}(S_\alpha)$	$\max\limits_{\alpha} t_{\text{mix}}(S_\alpha)$
Sequence of Dependencies	n^2	n	n^2
Two Islands	2^n	2^n	$n2^n$
Discrete Pyramid [20]	n	n^3	n^3
Memorize and Repeat	n^3	n^2	n^2
Soft Dependencies	$n^{3/2}$	n	n^2

where the maximum is taken over the distribution μ of the initial state of the process. When comparing the statistical efficiency of systematic scan and random scan, it would be useful to establish, for any systematic scan process S and random scan process R on the same n-variable model, a relative bound of the form

$$F_1(\epsilon, n, t_{\text{mix}}(R, \epsilon)) \leq t_{\text{mix}}(S, \epsilon) \leq F_2(\epsilon, n, t_{\text{mix}}(R, \epsilon)) \tag{1}$$

for some functions F_1 and F_2. Similarly, to bound the effect that the choice of permutation can have on the mixing time, it would be useful to know, for any two systematic scan processes S_α and S_β with different permutations on the same model, that for some function F_3,

$$t_{\text{mix}}(S_\alpha, \epsilon) \leq F_3(\epsilon, n, t_{\text{mix}}(S_\beta, \epsilon)). \tag{2}$$

Diaconis [3] and Levin et al. [15, p. 300] conjecture that systematic scan is never more than a constant factor slower or a logarithmic factor faster than random scan. This is equivalent to choosing $F_1(\epsilon, n, t) = C_1(\epsilon) \cdot t \cdot (\log n)^{-1}$ and $F_2(\epsilon, n, t) = C_2(\epsilon) \cdot t$ in the inequality in (1), for some functions C_1 and C_2. It is also commonly believed that all systematic scans mix at the same asymptotic rate, which is equivalent to choosing $F_3(\epsilon, n, t) = C_3(\epsilon) \cdot t$ in (2).

These conjectures imply that using systematic scan instead of random scan will not result in significant consequences, at least asymptotically, and that the particular permutation used for systematic scan is not important. However, we show that neither conjecture is true by constructing models (listed in Table 1) in which the scan order has substantial asymptotic effects on mixing time.

In the rest of this section, we go through two models in detail to highlight the diversity of behaviors that different scan orders can have. First, as a warm-up, we construct a model, which we call the *sequence of dependencies* model, for which a single "good permutation" of systematic scan mixes faster, by a polynomial factor, than both random scan and systematic scans using most other permutations. This serves as a counterexample to the conjectured lower bounds (i.e. the choice of F_1 and F_3) on the mixing time of systematic scan. Second, we construct a model, the *two islands* model, in which there is a small set of "bad permutations" that mix very slowly in comparison to random scan and most other systematic scans. This contradicts the conjectured upper bounds (i.e. the choice of F_2 and F_3). For completeness, we also discuss the *discrete pyramid* model introduced by Roberts and Rosenthal [20] (which contradicts the conjectured choice of F_2). Table 1 also lists the mixing times of a few additional models we constructed: these models further explore the space of asymptotic comparisons among scan orders, but for brevity we defer them to the appendix.

3.1   Sequence of Dependencies

The first model we will describe is the sequence of dependencies model (Figure 1a). The goal of this model is to explore the question of how fast systematic scan can be, by constructing a model for which systematic scan mixes rapidly for one particular good permutation. The sequence of dependencies model achieves this by having the property that, at any time, progress towards mixing is only made if a particular variable is sampled; this variable is always the one that is chosen by the good permutation. As a result, while a systematic scan using a good permutation makes progress at every step, both random scan and other systematic scans often fail to make progress, which leads to a gap between their mixing times. Thus, this model exhibits two surprising behaviors: (1) one systematic scan is polynomially better than random scan and (2) systematic scans using different permutations have polynomial differences in mixing times. We now describe this model in detail.
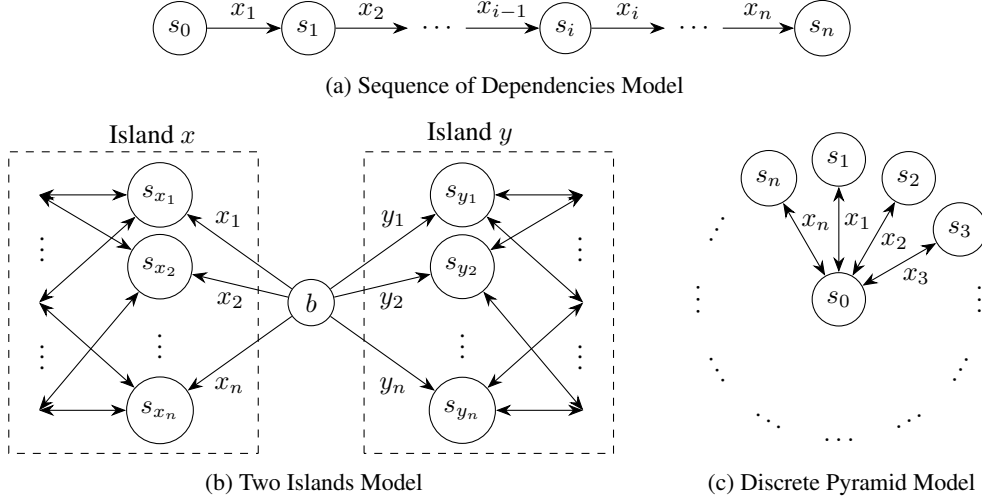
(a) Sequence of Dependencies Model

Island x Island y

(b) Two Islands Model

(c) Discrete Pyramid Model

Figure 1: State space of the models.

Variables   There are n binary variables $x_1, \ldots, x_n$. Independently, each variable has a very strong prior of being true. However, variable x_i is never true unless x_{i-1} is also true. The unnormalized probability distribution is the following, where M is a very large constant.

$$P(x) \propto \begin{cases} 0 & \text{if } x_i \text{ is true and } x_{i-1} \text{ is false for some } i \in \{2, \ldots, n\} \\ M^{|x|} & \text{otherwise} \end{cases}$$

State Space   There are $n + 1$ states with non-zero probability: $s_0, \ldots, s_n$, where s_i is the state where the first i variables are true and the remaining $n - i$ variables are false. In the stationary distribution, s_n has almost all of the mass due to the strong priors on the variables, so reaching s_n is essentially equivalent to mixing. Notice that sampling x_i will almost always move the state from s_{i-1} to s_i, very rarely move it from s_i to s_{i-1}, and can have no other effect. The worst-case starting state is s_0, where the variables must be sampled in the order $x_1, \ldots, x_n$ for this model to mix.

Random Scan   The number of steps needed to transition from s_0 to s_1 is distributed as a geometric random variable with mean n (variables are randomly selected, and specifically x_1 must be selected). Similarly, the number of steps needed to transition from s_{i-1} to s_i is distributed as a geometric random variable with mean n. In total, there are n transitions, so $O(n^2)$ steps are needed to mix.

Best Systematic Scan   The best systematic scan uses the order $x_1, x_2, \ldots, x_n$. For this scan, one sweep will reach s_n no matter what the starting state is, so the mixing time is n.

Worst Systematic Scan   The worst systematic scan uses the order $x_n, x_{n-1}, \ldots, x_1$. The first sweep only uses x_1, the second sweep only uses x_2, and in general, any sweep only makes progress using one transition. Finally, in the n-th sweep, x_n is used in the first step. Thus, this process mixes in $n(n-1) + 1$ steps, which is $O(n^2)$.

3.2   Two Islands

With the sequence of dependencies model, we showed that a single good permutation can mix much faster than other scan orders. Next, we describe the two islands model (Figure 1b), which has the reverse behavior: it has bad permutations that yield much slower mixing times. The two islands model achieves this by having two disjoint blocks of variables such that consecutively sampling two variables from the same block accomplishes very little. As a result, a systematic scan that uses a permutation that frequently consecutively samples from the same block mixes a polynomial factor slower than both random scan and most other systematic scans. We now describe this model in detail.

Variables There are $2n$ binary variables grouped into two blocks: $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Conditioned on all other variables being false, each variable is equally likely to be true or false. However, the x variables and the y variables contradict each other. As a result, if any of the x's are true, then all of the y's must be false, and if any of the y's are true, then all of the x's must be false. The unnormalized probability distribution for this model is the following.

$$P(x, y) \propto \begin{cases} 0 & \text{if } \exists x_i \text{ true and } \exists y_j \text{ true} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

This model can be interpreted as a machine learning inference problem in the following way. Each variable represents whether the reasoning in some sentence is sound. The sentences corresponding to $x_1, \ldots, x_n$ and the sentences corresponding to $y_1, \ldots, y_n$ reach contradicting conclusions. If any variable is true, its conclusion is correct, so all of the sentences that reached the opposite conclusion must be not be sound, and their corresponding variables must be false. However, this does not guarantee that all other sentences that reached the same conclusion have sound reasoning, so it is possible for some variables in a block to be true while others are false. Under these assumptions alone, the natural way to model this system is with the two islands distribution in (3).

State Space We can think of the states as being divided into three groups: states in island x (at least one of the x variables are true), states in island y (at least one of the y variables are true), and a single *bridge state* b (all variables are false). The islands are well-connected internally, but it is impossible to directly move from one island to the other – the only way to move from one island to the other is through the bridge. To simplify the analysis, we assume that the bridge state has very low mass. This assumption allows us to assume that islands mix rapidly in comparison to the time required to move onto the bridge and that chains always move off of the bridge when a variable is sampled. The same asymptotic behavior results when the bridge state has the same mass as the other states.

The bridge is the only way to move from one island to the other, so it acts as a bottleneck. As a result, the efficiency of bridge usage is critical to the mixing time. We will use *bridge efficiency* to refer to the probability that the chain moves to the other island when it reaches the bridge. Under the assumption that mixing within the islands is rapid in comparison to the time needed to move onto the bridge, the mixing time will be inversely proportional to the bridge efficiency of the chain.

Random Scan In random scan, the variable selected after getting on the bridge is independent of the previous variable. As a result, with probability $1/2$, the chain will move onto the other island, and with probability $1/2$, the chain will return to the same island, so the bridge efficiency is $1/2$.

Best Systematic Scan Several different systematic scans achieve the fastest mixing time. One such scan is $x_1, y_1, x_2, y_2, \ldots, x_n, y_n$. Since the sampled variables alternate between the blocks, if the chain moves onto the bridge (necessarily by sampling a variable from the island it was previously on), it will always proceed to sample a variable from the other block, which will cause it to move onto the other island. Thus, the bridge efficiency is 1. More generally, any systematic scan that alternates between sampling from x variables and sampling from y variables will have a bridge efficiency of 1.

Worst Systematic Scan Several different systematic scans achieve the slowest mixing time. One such scan is $x_1, \ldots, x_n, y_1, \ldots, y_n$. In this case, if the chain moves onto the bridge, it will almost always proceed to sample a variable from the same block, and return to the same island. In fact, the only way for this chain to move across islands is if it moves from island x to the bridge using transition x_n and then moves to island y using transition y_1, or if it moves from island y to the bridge using transition y_n and then moves to island x using transition x_1. Thus, only 2 of the $2n$ transitions will cross the bridge, and the bridge efficiency is $1/n$. More generally, any systematic scan that consecutively samples all x variables and then all y variables will have a bridge efficiency of $1/n$.

Comparison of Mixing Times The mixing times of the chains are inversely proportional to the bridge efficiency. As a result, random scan takes twice as long to mix as the best systematic scan, and mixes $n/2$ times faster than the worst systematic scan.

3.3 Discrete Pyramid

In the discrete pyramid model (Figure 1c) introduced by Roberts and Rosenthal [20], there are n binary variables x_i, and the mass is uniformly distributed over all states where at most one x_i is true. In this model, the mixing time of random scan, $O(n)$, is asymptotically better than that of systematic scan for any permutation, which all have the same mixing time, $O(n^3)$.

4 Experiments

In this section, we run several experiments to illustrate the effect of scan order on mixing times. First, in Figure 2a, we plot the mixing times of the models from Section 3 as a function of the number of variables. These experiments validate our results about the asymptotic scaling of the mixing time, as well as show that the scan order can have a significant effect on the mixing time for even small models. (Due to the exponential state space of the two islands model, we modify it slightly to make the computation of mixing times feasible: we simplify the model by only considering the states that are adjacent to the bridge, and assume that the states on each individual island mix instantly.)

In the following experiments, we consider a modified version of the two islands model, in which the mass of the bridge state is set to 0.1 of the mass of the other states to allow the effect of scan order to be clear even for a small number of variables. Figure 2b illustrates the rate at which different scan orders explore this modified model. Due to symmetry, we know that half of the mass should be on each island in the stationary distribution, so getting half of the mass onto the other island is necessary for mixing. This experiment illustrates that random scan and a good systematic scan move to the other island quickly, while a bad systematic scan requires many more iterations.

Figure 2c illustrates the effect that the permutation chosen for systematic scan can have on the mixing time. In this experiment, the mixing time for each permutation was found and plotted in sorted order. For the sequence of dependencies model, there are a small number of good permutations which mix very quickly compared to the other permutations and random scan. However, no permutation is bad compared to random scan. In the two islands model, as we would expect based on the analysis in Section 3, there are a small number of bad permutations which mix very slowly compared to the other permutations and random scan. Some permutations are slightly better than random scan, but none of the scan orders are substantially better. In addition, the mixing times for systematic scan approximately discretized due to the fact that mixing time depends so heavily on the bridge efficiency.

5 Relative Bounds on Mixing Times via Conductance

In Section 3, we described two models for which a systematic scan can mix a polynomial factor faster or slower than random scan, thus invalidating conventional wisdom that the scan order does not have an asymptotically significant effect on mixing times. This raises a question of how different the mixing times of different scans can be. In this section, we derive the following weaker – but correct – version of the conjecture stated by Diaconis [3] and Levin et al. [15].
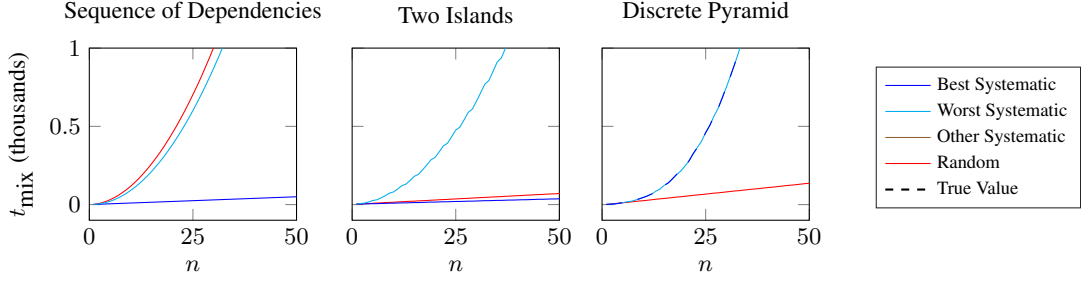
One of the obstacles to proving this result is that the systematic scan chain is not *reversible*. A standard method of handling non-reversible Markov chains is to study a lazy version of the Markov chain instead [15, p. 9]. In the lazy version of a Markov chain, each step has a probability of $1/2$ of staying at the current state, and acts as a normal step otherwise. This is equivalent to stopping at a random time that is distributed as a binomial random variable. Due to the fact that systematic scan is not reversible, our bounds are on the lazy systematic scan, rather than the standard systematic scan.

Theorem 1. *For any random scan Gibbs sampler R and lazy systematic scan sampler S with the same stationary distribution π, their relative mixing times are bounded as follows.*
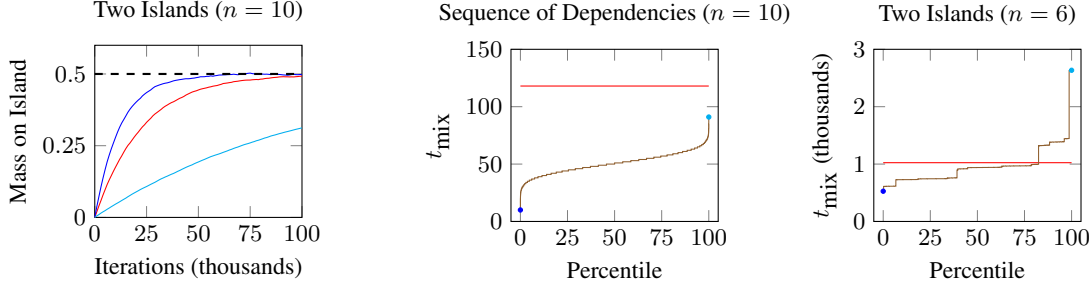
$$(1/2 - \epsilon)^2\, t_{mix}(R, \epsilon) \le 2t_{mix}^2(S, \epsilon) \log\left(\frac{1}{\epsilon \pi_{min}}\right)$$

$$(1/2 - \epsilon)^2\, t_{mix}(S, \epsilon) \le \frac{8n^2}{(\min_{x,i} P_i(x,x))^2} t_{mix}^2(R, \epsilon) \log\left(\frac{1}{\epsilon \pi_{min}}\right),$$

where P_i is the transition matrix corresponding to resampling just variable i, and π_{min} is the probability of the least likely state in π.

Sequence of Dependencies   Two Islands   Discrete Pyramid

(a) Mixing times for $\epsilon = 1/4$.

Two Islands ($n = 10$)   Sequence of Dependencies ($n = 10$)   Two Islands ($n = 6$)

(b) Marginal island mass over time.

(c) Sorted mixing times of different permutations ($\epsilon = 1/4$).

Figure 2: Empirical analysis of the models.

Under mild conditions, namely ϵ being fixed and the quantities $\log(\pi_{\min}^{-1})$ and $(\min_{x,i} P_i(x,x))^{-1}$ being at most polynomial in n, this theorem implies that the choice of scan order can only affect the mixing time by up to polynomial factors in n and t_{mix}. We now outline the proof of this theorem.

In the two islands models, the mixing time of a scan order was determined by its ability to move through a single bridge state that restricted flow. This suggests that a technique with the ability to model the behavior of this bridge state is needed to bound the relative mixing times of different scans. Conductance, also known as the bottleneck ratio, is a topological property of Markov chains used to bound mixing times by considering the flow of mass around the model [15, p. 88]. This ability to model bottlenecks in a Markov chain makes conductance a natural technique both for studying the two islands model and bounding mixing times in general.

More formally, consider a Markov chain on state space Ω with transition matrix P and stationary distribution π. The *conductance* of a set S and of the whole chain are respectively defined as

$$\Phi(S) = \frac{\sum_{x \in S, y \notin S} \pi(x) P(x,y)}{\pi(S)} \qquad\qquad \Phi_\star = \min_{S:\pi(S) \le \frac{1}{2}} \Phi(S).$$

Conductance can be directly applied to analyze random scan. Let P_i be the transition matrix corresponding to sampling variable i. The state space Ω is used without modification, and the transition matrix is $P = \frac{1}{n} \sum_{i=1}^{n} P_i$. The stationary distribution is the expected target distribution π.

On the other hand, conductance cannot be directly applied to systematic scan. Systematic scan is not a true Markov chain because it uses a sequence of transition matrices rather than a single transition matrix. One standard method of converting systematic scan into a true Markov chain is to consider each full scan as one step of a Markov chain. However, this makes it difficult to compare with random scan because it completely changes which states are connected by single steps of the transition matrix. To allow systematic and random scan to be compared more easily, we introduce an alternative way of converting systematic scan to a true Markov chain by augmenting the state space. The augmented state space is $\Psi = \Omega \times [n]$, which represents an ordered pair of the normal state and the index of the variable to be sampled. The transition probability is $P((x,i),(y,j)) = P_i(x,y)s(i,j)$, where $s(i,j) = \mathbb{I}[i+1 \equiv j \pmod{n}]$ is an indicator that shows if the correct variable will be sampled next.

Additionally, augmenting the state space for random scan allows easier comparison with systematic scan in some cases. For augmented random scan, the state space is also $\Psi = \Omega \times [n]$, the same as for systematic scan. The transition

probability is $P\left((x,i),(y,j)\right) = \frac{1}{n}P_i(x,y)$, which means that the next variable to sample is selected uniformly. The stationary distributions of the augmented random scan and systematic scan chains are both $\pi\left((x,i)\right) = n^{-1}\pi(x)$. Because the state space and stationary distribution are the same, augmented random scan and augmented systematic scan can be compared directly, which lets us prove the following lemma.

Lemma 1. *For any random scan Gibbs sampler and systematic scan sampler with the same stationary distribution π, let Φ_{RS} denote the conductance of the random scan process, let $\Phi_{RS\text{-}A}$ denote the conductance of the augmented random scan process, and let $\Phi_{SS\text{-}A}$ denote the conductance of the augmented systematic scan process. Then,*

$$\frac{1}{2n} \cdot \min_{x,i} P_i(x,x) \cdot \Phi_{RS\text{-}A} \le \Phi_{SS\text{-}A} \le \Phi_{RS}.$$

In Lemma 1, the upper bound states that the conductance of systematic scan is no larger than the conductance of random scan. We use this in the next section to show that systematic scan cannot mix too much more quickly than random scan. To prove this lemma, we show that for any set S under random scan, the set $\hat{S}$ containing the corresponding augmented states for systematic scan will have the same conductance under systematic scan as S had under random scan.

The lower bound in Lemma 1 states that the conductance of systematic scan is no smaller than a function of the conductance of augmented random scan. This function depends on the number of variables n and $\min_{x,i} P_i(x,x)$, which is the minimum holding probability of any state. To prove this lemma, we show that for any set S under augmented systematic scan, we can bound its conductance under augmented random scan.

There are well-known bounds on the mixing time of a Markov chain in terms of its conductance, which we state in Theorem 2 [15, pp. 89, 235].

Theorem 2. *For any lazy or reversible Markov chain,*

$$\frac{1/2 - \epsilon}{\Phi_\star} \le t_{mix}(\epsilon) \le \frac{2}{\Phi_\star^2} \log\left(\frac{1}{\epsilon\pi_{min}}\right).$$

It is straightforward to prove the result of Theorem 1 by combining these bounds with the conductance bounds from the previous section.

6 Conclusion

We studied the effect of scan order on mixing times of Gibbs samplers, and found that for particular models, the scan order can have an asymptotic effect on the mixing times. These models invalidate conventional wisdom about scan order and show that we cannot freely change scan orders without considering the resulting changes in mixing times. In addition, we found bounds on the mixing times of different scan orders, which replaces a common conjecture about the mixing times of random scan and systematic scan.

Acknowledgments

References

[1] Itai Benjamini, Noam Berger, Christopher Hoffman, and Elchanan Mossel. Mixing times of the biased card shuffling and the asymmetric exclusion process. *Transactions of the American Mathematical Society*, 357(8):3013–3029, 2005.

[2] Christopher De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Rapidly mixing gibbs sampling for a class of factor graphs using hierarchy width. In *Advances in Neural Information Processing Systems*, 2015.

[3] Persi Diaconis. Some things we've learned (about markov chain monte carlo). *Bernoulli*, 19(4):1294–1305, 2013.

[4] Persi Diaconis and Arun Ram. Analysis of systematic scan metropolis algorithms using iwahori-hecke algebra techniques. *The Michigan Mathematical Journal*, 48(1):157–190, 2000.

[5] Martin Dyer, Leslie Ann Goldberg, and Mark Jerrum. Systematic scan for sampling colorings. *The Annals of Applied Probability*, 16(1):185–230, 2006.

[6] Martin Dyer, Leslie Ann Goldberg, and Mark Jerrum. Dobrushin conditions and systematic scan. *Combinatorics, Probability and Computing*, 17(06):761–779, 2008.

[7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[8] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[9] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[10] Alkis Gotovos, Hamed Hassani, and Andreas Krause. Sampling from probabilistic submodular models. In *Advances in Neural Information Processing Systems*, 2015.

[11] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[12] Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Convergence rate of incremental gradient and newton methods. *arXiv preprint arXiv:1510.08562*, 2015.

[13] Thomas P. Hayes. A simple condition implying rapid mixing of single-site dynamics on spin systems. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 2006.

[14] Joseph M Hellerstein, Christoper Ré, Florian Schoppmann, Daisy Zhe Wang, Eugene Fratkin, Aleksander Gorajek, Kee Siong Ng, Caleb Welton, Xixuan Feng, Kun Li, et al. The madlib analytics library: or mad skills, the sql. *Proceedings of the VLDB Endowment*, 5(12):1700–1711, 2012.

[15] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2009.

[16] David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

[17] Andrew McCallum, Karl Schultz, and Sameer Singh. Factorie: Probabilistic programming via imperatively defined factor graphs. In *Advances in Neural Information Processing Systems*, 2009.

[18] Martyn Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, 2003.

[19] Benjamin Recht and Christopher Ré. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

[20] Gareth O. Roberts and Jeffrey S. Rosenthal. Surprising convergence properties of some simple gibbs samplers under various scans. *International Journal of Statistics and Probability*, 5(1):51–60, 2015.

[21] Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1):703–710, 2010.

[22] Ce Zhang and Christopher Ré. Towards high-throughput gibbs sampling at scale: A study across storage managers. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013.

[23] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[24] Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.

A Additional Models in Which Scan Order Matters

Table 2: Models Classified by Relative Mixing Times

		$\min_\alpha t_{\mathrm{mix}}(S_\alpha)$ versus $t_{\mathrm{mix}}(R)$		
		$\ll$	$\approx$	$\gg$
$\max_\alpha t_{\mathrm{mix}}(S_\alpha)$	$\ll$	Memorize and Repeat	$\times$	$\times$
versus	$\approx$	Sequence of Dependencies	Known Cases	$\times$
$t_{\mathrm{mix}}(R)$	$\gg$	Soft Dependencies	Two Islands	Discrete Pyramid [20]

Table 2 classifies the models based on the mixing times of the best systematic scan and worst systematic scan relative to random scan. For models in the first column, the best systematic scan mixes asymptotically faster than random scan. For models in the second column, they differ only up to logarithmic factors. For models in the third column, the best systematic scan mixes asymptotically slower than random scan. The rows are classified in the same way based on the worst systematic scan instead of the best systematic scan.

The models in Section 3 display possible problems that can occur when changing scan orders. However, two behaviors were not shown by the models in Section 3: a case where random scan is asymptotically worse than all systematic scans, and a case where random scan is asymptotically better than a systematic scan and asymptotically worse than another systematic scan. This section gives two additional models that exhibit these behaviors, along with a more detailed explanation of the discrete pyramid model. These two models are more complicated than the models in Section 3 due to the fact that they require the number of states for each variable to grow with n. Combined with the models in Section 3, these models show that all logically consistent asymptotic behaviors (that is, behaviors where the worst systematic scan is no better than the best systematic scan) are possible.

A.1 Discrete Pyramid

The discrete pyramid model (Figure 1c) was first described by Roberts and Rosenthal [20] and is included for completeness to show that it is possible for random scan to mix asymptotically faster than all systematic scans.

Variables There are n binary variables $x_1, \ldots, x_n$. Conditioned on all other variables being false, each variable is equally likely to be true or false. However, the variables are all contradictory, and at most one variable can be true. This model can also be interpreted as an n islands model, where there are n well-connected regions (which consist of a single state) connected by a single bridge state.

State Space There are $n + 1$ states $s_0, s_1, \ldots, s_n$ with nonzero probability. s_0 is the state where all variables are false, and for all other i, s_i is the state where variable x_i is true and all other variables are false.

Random Scan The worst-case total variation occurs when the starting state is not x_0. Suppose that the starting state is s_k. x_k will be selected with probability $1/n$, and in this case, the state will change to x_0 with probability $1/2$. Thus, the number of samples needed to leave x_k is distributed as a Geometric random variable with a mean of $2n$. This means that within $O(n)$ steps, the chain will have left the initial state with high probability. Once the chain has reached x_0, each step has a probability of $1/2$ of leaving x_0 and uniformly going to another state. Thus, $O(1)$ steps are sufficient for the chain to mix after it reaches x_0. In total, the chain mixes in $O(n)$ updates.

All Systematic Scans Once again, the worst-case total variation occurs when the starting state is not x_0. Suppose that the starting state is s_k. A systematic scan step will change nothing until x_k is reached. Then, with probability $1/2$, the state will remain s_k, and with probability $1/2$ the state will change to x_0. If it does move to x_0, the scan will continue, and move away from x_0 with probability $1/2$ at each step. Thus, each time the scan reaches the current state, the state will advance Z steps, where Z is a Geometric random variable with rate $1/2$. This corresponds to a weighted random walk on a circle, which is known to have a mixing time of $O(n^2)$. Notice that each step of the random walk actually requires one full sweep, so $O(n^3)$ steps are needed for any systematic scan to mix.
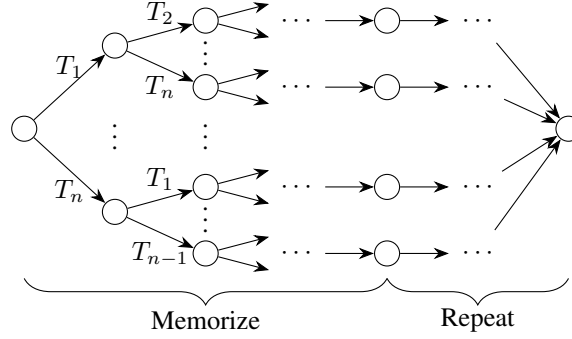
Figure 3: Memorize and Repeat Model

A.2 Memorize and Repeat

We introduce the memorize and repeat model to show that it is possible for all systematic scans to be asymptotically better than random scan.

Description   In this example (Figure 3), there are two types of states: states memorizing a permutation, and states requiring the memorized permutation to be repeated. In the stationary distribution, almost all of the mass is on the state where a permutation has been memorized and repeated, and the probability of the states increase exponentially from left to right. There are n variables $x_1, \ldots, x_n$. In the memorize phase, sampling a variable will allow the state to change as long as the variable has not been sampled before. In the repeat phase, a sampling a variable will only allow the state to change if it is the next variable in the memorized permutation. The repeat phase is similar to the sequence of dependencies example, with the variables rearranged to match the memorized permutation.

In the worst-case starting distribution, the wrong permutation can already be memorized, which results in no asymptotic gap between mixing times. To allow the difference in mixing time to appear even in the worst-case analysis, we repeat the memorize and repeat process n times.

Random Scan   Random scan encounters the coupon collector problem to get through the memorize phase, so $O(n \log n)$ steps are needed. Afterwards, the repeat phase is equivalent to a sequence of dependencies, which requires $O(n^2)$ steps. Thus, to get through one memorize and repeat chain, $O(n^2)$ steps are needed. This mechanism is repeated n times, so $O(n^3)$ steps are needed.

All Systematic Scans   If a systematic scan starts at the beginning of a memorize sequence, $2n$ steps are sufficient to move through the memorize and repeat chain (n steps to memorize, n steps to repeat). However, when analyzing the worst-case total variation, we must consider the case where the wrong permutation has already been memorized, which requires $O(n^2)$ steps to move through. However, this can only happen once, so in the worst case $O(n^2)$ steps are needed to move through the one incorrect permutation, and $O(n)$ steps are needed for the remaining $n - 1$ memorize and repeat chains. Thus, $O(n^2)$ steps are needed in total.

State Space   To simplify the explanation of the state space and formulation as a Gibbs sampler, this explanation is for only one memorize and repeat cycle. In the memorize phase, the states are labeled by the part of the permutation that has already been memorized. In the repeat phase, the states are labeled by the remainder of the permutation that still has to be repeated.

Formulation as a Gibbs Sampler   There are n integer-valued variables $x_1, \ldots, x_n$ with values ranging from 0 to $n + 1$. In the initial state that does not have anything memorized, all variables have a value of 0. In the memorize phase, any variable that has not yet been memorized has a value of 0, and any variable that has been memorized stores its index within the permutation. This means that for states in the memorize phase with i variables memorized, exactly one variable will have a value of j, for each integer j from 1 to i, and the remaining $n - i$ variables that have not been

memorized have a value of 0. As more variables are memorized, the states become exponentially more likely. Next, in the repeat phase, when a variable is used, its value changes to $n + 1$. This means that for states in the repeat phase with i variables repeated, the i variables that have been repeated have a value of $n + 1$, and the remaining $n - i$ variables that still have to be repeated have a unique integer from $i + 1$ to n. As more variables are repeated, the states become exponentially more likely. The log probability distribution is the following, where M is a very large constant, Z is the normalizing constant, and $\sigma(a_1, \ldots, a_n)$ denotes the set of all permutations of $(a_1, \cdots, a_n)$.

$$\log_M P(x) = -\log_M Z + \begin{cases} \max(x_1, \ldots, x_n) & \text{if } (x_1, \ldots, x_n) \in \sigma(0, \ldots, 0, \ 1, \ldots, i) \\ n + \sum_{i=1}^{n} \mathbb{I}[x_i = n + 1] & \text{if } (x_1, \ldots, x_n) \in \sigma(i + 1, \ldots, n, \ n + 1, \ldots, n + 1) \\ -\infty & \text{otherwise} \end{cases}$$

$$= -\log_M Z + \begin{cases} \text{(Number of Memorized Variables)} & \text{if valid memorize state} \\ n + \text{(Number of Repeated Variables)} & \text{if valid repeat state} \\ -\infty & \text{otherwise} \end{cases}$$

First, for states in the memorize phase, if a variable that has not yet been memorized is sampled, then its value will almost always change to $i + 1$, where i is the number of variables already memorized, due to the large value of M. However, sampling a variable that has already been memorized will not change its value because all other values will either result in an invalid state (and have a probability of 0) or be significantly less likely due to the large value of M. Next, for states in the repeat phase, if the correct variable is sampled, then its value will almost always change to $n + 1$. However, sampling the wrong variable will not change its value because all other values result in an invalid state or decrease the probability of the state.

A.3 Soft Dependencies

In the soft dependencies model, some systematic scans mix asymptotically faster than random scan, and some systematic scans mix asymptotically slower.

Description This example resembles the memorize portion of the previous example. However, no repeat phase is included, and only some permutations are accepted. In particular, a permutation is accepted only if each element of the permutation is followed by an element that is in the next $\sqrt{n}$ unused variables that come after it (mod n).

More formally, consider a permutation $(a_1, \ldots, a_n)$ of $(1, \ldots, n)$. The permutation is accepted only if the following holds for all $i \in \{1, \ldots, n - 1\}$.

$$\sqrt{n} \geq |\{ j \in \{1, \ldots, n\} \mid j > i + 1 \text{ and } a_j - a_i \ (\text{mod } n) < a_{i+1} - a_i \ (\text{mod } n) \}|$$

In this condition, the requirement that $j > i + 1$ means that a_j is an unused element, and the requirement that $a_j - a_i \ (\text{mod } n) < a_{i+1} - a_i \ (\text{mod } n)$ means that starting from a_i, a_j is reached before a_{i+1}. These two requirements imply that a_{i+1} is within the next $\sqrt{n}$ unused variables following a_i. This condition is equivalent to the following.

$$\sqrt{n} \geq \begin{cases} \sum_{j=i+2}^{n} \mathbb{I}[a_i < a_j < a_{i+1}] & \text{if } a_i < a_{i+1} \\ \sum_{j=i+2}^{n} \mathbb{I}[a_i < a_j \text{ or } a_j < a_{i+1}] & \text{if } a_i > a_{i+1} \end{cases}$$

Random Scan First, consider when there are at least $\sqrt{n}$ remaining transitions. The first $n - \sqrt{n} = O(n)$ transitions will have this property. During this period, random scan advances to the next state with probability $n/\sqrt{n} = \sqrt{n}$. Thus, $O(n\sqrt{n})$ steps are needed to get until there are fewer than $\sqrt{n}$ remaining transitions. Once there are fewer than $\sqrt{n}$ remaining transitions, random scan needs $O(n)$ steps to make each transition, so $O(n\sqrt{n})$ more steps are needed. In total, $O(n\sqrt{n})$ steps are needed.

Best Systematic Scan The best systematic scan uses the order $x_1, \ldots, x_n$. This uses n steps to give a valid permutation, and thus mixes in n steps.

Worst Systematic Scan The worst systematic scan uses the order $x_n, x_{n-1}, \ldots, x_1$. After a transition is taken, nearly a full scan is needed until the next valid transition will be reached, so $O(n^2)$ steps are needed.

State Space The state space has a similar form to the memorize portion of the previous example. The states are labeled by the portion of the permutation that has already been given, but only prefixes to valid permutations are accepted.

Formulation as a Gibbs Sampler There are n integer-valued variables $x_1, \ldots, x_n$ with values ranging from 0 to n. In the initial state that does not have anything memorized, all variables have a value of 0. Then, as the permutation is memorized, the variables that have not been used yet have a value of 0, and any variable that has been used stores its index within the permutation.

In addition, to merge the permutations together into one state, all variables need to be sampled one more time (note that this does not require the variables to be sampled in a particular order). This extra sampling process only adds lower order terms and constant factors to the mixing time. Once each variable is sampled again, its value is changed to n.

The log probability distribution is the following, where M is a very large constant, Z is the normalizing constant, and $\sigma(a_1, \ldots, a_n)$ denotes the set of all permutations of $(a_1, \cdots, a_n)$.

$$\log_M P(x) = -\log_M Z + \begin{cases} \max(x_1, \ldots, x_n) & \text{if } (x_1, \ldots, x_n) \in \sigma(0, \ldots, 0,\ 1, \ldots, i) \text{ and is valid} \\ n - 1 + \sum_{i=1}^{n} \mathbb{I}[x_i = n] & \text{if at least two of } (x_1, \ldots, x_n) \text{ have a value of } n \\ -\infty & \text{otherwise} \end{cases}$$

$$= -\log_M Z + \begin{cases} \text{(Number of Memorized Variables)} & \text{if valid memorize state} \\ n - 1 + \text{(Number of Merged Variables)} & \text{if valid merge state} \\ -\infty & \text{otherwise} \end{cases}$$

First, in the memorize phase, sampling a valid unused variable will almost always change its value to $i + 1$, where i is the number of variables already memorized, due to the large value of M. However, sampling a variable that has already been memorized or sampling a variable that least to an invalid permutation will not change its value. Then, when the last variable in the permutation is sampled, its value changes to n. Afterwards, sampling any variable will also change its value to n.

B   Priors in the Sequence of Dependencies

To simplify the analysis of the sequence of dependencies, we assumed that the priors on the variables were strong enough such sampling the correct variable essentially always caused the state to advance. We now analyze how large M (the strength of the priors) needs to be in order to allow the best systematic scan to mix in $O(n)$ time. In this section, we show that if $M = \Omega(n)$, then the best systematic scan mixes in $O(n)$ time, and if $M = o(n)$, then the best systematic scan cannot mix $O(n^2)$ time.

First, suppose that $M = \Omega(n)$. This means that $M \geq cn$ for some $c > 0$ and sufficiently large n. The probability of transitioning from s_{i-1} to s_i when variable x_i is sampled is

$$\frac{M}{1 + M} \geq \frac{cn}{1 + cn} = 1 - \frac{1}{1 + cn} > 1 - \frac{1}{cn}.$$

The probability of transitioning from s_0 to s_n after sampling the sequence $x_1, \ldots, x_n$ is at least $(1 - 1/cn)^n$, which limits to $e^{-1/c}$ for large n. In other words, if $M = \Omega(n)$, then a single sweep of the best systematic scan will reach s_n with a probability that does not approach 0. Thus, a constant number of sweeps is sufficient to reach s_n with high probability, which is equivalent to mixing.

On the other hand, suppose that $M = o(n)$. Then, for any $c > 0$, $M < cn$ for sufficiently large n. This means that as n increases, the probability that a single sweep of the best systematic scan will reach s_n will go to 0, so no constant number of sweeps will be sufficient to mix.

C   Bridge Efficiency with Normal Mass on Bridge

In the analysis of the two islands model, we assumed that the bridge has negligible mass. We now analyze the mixing times without the assumption that the bridge has small mass. The same asymptotic behavior still results.

Even when the bridge has the same mass as the other states, it still acts as a bottleneck to the model. For sufficiently large n, the islands will mix rapidly in comparison reaching the bridge, so the mixing time is still inversely proportional to the bridge efficiency. However, when the bridge has the same mass as the other states, sampling a variable while on the bridge will only have a $1/2$ chance of moving off of the bridge.

Random Scan   In random scan, the variables that are sampled are completely independent, so the bridge efficiency is still $1/2$.

Best Systematic Scan   Consider the scan $x_1, y_1, x_2, y_2, \ldots$. Suppose the sampling x_1 changes the state to the bridge state – a similar analysis will apply for any other variable. The next variable is y_1, which will change the state to the other island with probability $1/2$. Afterwards, x_2 is sampled, which will change the state to the same island with probability $1/4$. Then, y_2 is sampled, which will change the state to the other island with probability $1/8$. Thus, the probability of moving to the other island is $1/2 + 1/8 + 1/32 + \ldots = 2/3$.

Worst Systematic Scan   Consider the scan $x_1, \ldots, x_n, y_1, \ldots, y_n$. Suppose that sampling x_n changes the state to the bridge state. For large n, the probability of leaving the bridge state onto island y approaches 1. Next, suppose that sampling x_{n-1} changes the state to the bridge state. There is a probability of $1/2$ of moving back to island x via x_n, but the chain will move onto island y otherwise. In general, moving onto the bridge via x_i or y_i will result in moving to the island with probability 2^{-n+i}. The average probability of moving onto the other island is then $2/n$.

D   Proofs for Section 5

In this section, we prove our relative bounds on mixing times (Theorem 1), along with related claims and lemmas.

Claim 1. *The stationary distribution of augmented random and systematic scan is*

$$\pi\left((x,i)\right) = \frac{1}{n}\pi(x)$$

Proof. We prove this claim by showing that applying the transition matrix for augmented random scan or augmented systematic scan does not change this distribution.

For augmented random scan,

$$\sum_{x,i} \pi((x,i)) P((x,i),(y,j)) = \sum_{x,i} \frac{\pi(x)}{n} \cdot P((x,i),(y,j)) = \frac{1}{n}\sum_{x,i} \pi(x) \cdot \frac{1}{n} P_i(x,y)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[\sum_{x\in\Omega} (\pi(x)P_i(x,y)) \cdot \frac{1}{n} \right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[\pi(y) \cdot \frac{1}{n} \right] = \frac{1}{n}\pi(y)$$

$$= \pi((y,j))$$

For augmented systematic scan,

$$\sum_{x,i} \pi((x,i)) P((x,i),(y,j)) = \sum_{x,i} \frac{\pi(x)}{n} \cdot P((x,i),(y,j)) = \frac{1}{n}\sum_{x,i} \pi(x) \cdot P_i(x,y) s(i,j)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[\sum_{x\in\Omega} (\pi(x)P_i(x,y)) \, s(i,j) \right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} [\pi(y)s(i,j)] = \frac{1}{n}\pi(y)$$

$$= \pi((y,j))$$

$\square$

Lemma 1. *For any random scan Gibbs sampler and systematic scan sampler with the same stationary distribution π, let Φ_{RS} denote the conductance of the random scan process, let $\Phi_{RS\text{-}A}$ denote the conductance of the augmented random scan process, and let $\Phi_{SS\text{-}A}$ denote the conductance of the augmented systematic scan process. Then,*

$$\frac{1}{2n} \cdot \min_{x,i} P_i(x,x) \cdot \Phi_{RS\text{-}A} \leq \Phi_{SS\text{-}A} \leq \Phi_{RS}.$$

Proof.

Upper Bound: The conductance of the whole chain is the smallest conductance of any set with mass no larger than $\frac{1}{2}$. Then, to prove that this inequality holds, we will show that for any set $S \in \Omega$ with mass no larger than $\frac{1}{2}$, there exists a set $T \in \Psi$ with mass no larger than $\frac{1}{2}$ such that the conductance of S under random scan is the same as the conductance of T under augmented systematic scan.

From the standpoint of random scan, consider a set $S \in \Omega$ with mass no larger than $\frac{1}{2}$. The conductance is

$$\Phi_{RS}(S) = \frac{\sum_{x \in S} \sum_{y \in S^c} \pi(x) P(x,y)}{\pi(S)}$$

$$= \frac{\frac{1}{n} \sum_{i=1}^{n} \sum_{x \in S} \sum_{y \in S^c} \pi(x) P_i(x,y)}{\pi(S)}$$

Then, for augmented systematic scan, consider the set $T = \{(x,i) : x \in S, i \in [n]\}$. First, notice that $\pi(T) = \pi(S) \leq \frac{1}{2}$. The conductance is

$$\Phi_{SS\text{-}A}(T) = \frac{\sum_{(x,i) \in T} \sum_{(y,j) \in T^c} \pi((x,i)) P((x,i),(y,j))}{\pi(T)}$$

$$= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{x \in S} \sum_{y \in S^c} \pi((x,i)) P((x,i),(y,j))}{\pi(S)}$$

$$= \frac{1}{n} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{x \in S} \sum_{y \in S^c} \pi(x) P_i(x,y) s(i,j)}{\pi(S)}$$

$$= \frac{1}{n} \frac{\sum_{i=1}^{n} \sum_{x \in S} \sum_{y \in S^c} \pi(x) P_i(x,y)}{\pi(S)}$$

$$= \Phi_{RS}(S)$$

This implies that for any $S \in \Omega$ with $\pi(S) \leq \frac{1}{2}$, there exists $T \in \Psi$ with $\pi(T) \leq \frac{1}{2}$ such that $\Phi_{RS}(S) = \Phi_{SS}(T)$. Therefore, $\Phi_{SS\text{-}A} \leq \Phi_{RS}$.

Lower Bound: In this proof, we will work with the *flow* between two sets

$$Q(A,B) = \sum_{x \in A, y \in B} \pi(x) P(x,y).$$

Notice that the conductance of a set can then be defined as

$$\Phi(S) = \frac{Q(S, S^c)}{\pi(S)}.$$

To prove that this inequality holds for the whole chain, we will show that the same inequality holds for any set $S \in \Psi$. Consider some arbitrary state $x \in \Omega$. Flow can leave from the corresponding augmented states in two ways: flowing from some $(x,i) \in S$ to $(x,j) \in S^c$ or flowing from $(x,i) \in S$ to $(y,j) \in S^c$, where $y \neq x$ (x and y differ in only variable i). Let $S_x = \{(x,i) \in S\}$, and let $S_x^c = \{(x,i) \in S^c\}$. These two components can be written as $Q(S_x, S_x^c)$ and $\sum_{y \neq x} Q(S_x, S_y^c)$.

Now, we will find upper bounds for the random scan flows and lower bounds for the systematic scan flows. In the following statements, it is implicit that $y \neq x$, and $\gamma = \min_{x,i} P_i(x,x)$ will denote the minimum holding probability.

First, we bound the amount of flow from $(x, i) \in S$ to $(x, j) \in S^c$. For augmented random scan, the following upper bound holds.

$$
\begin{aligned}
Q_{\text{RS}}(S_x, S_x^c) &= \sum_{(x,i) \in S_x, (x,j) \in S_x^c} \pi((x,i)) P((x,i),(x,j)) \\
&= \sum_{(x,i) \in S_x, (x,j) \in S_x^c} \frac{1}{n} \pi(x) \cdot \frac{1}{n} P_i(x,x) \\
&\leq \sum_{(x,i) \in S_x, (x,j) \in S_x^c} \frac{1}{n} \pi(x) \cdot \frac{1}{n} \\
&= \frac{|S_x|}{n} \pi(x) \frac{n - |S_x|}{n} \\
&\leq \begin{cases} \frac{1}{4} \pi(x) & \text{if } |S_x| \neq 0, n \\ 0 & \text{if } |S_x| = 0, n \end{cases}
\end{aligned}
$$

For augmented systematic scan, the following lower bound holds.

$$
\begin{aligned}
Q_{\text{SS}}(S_x, S_x^c) &= \sum_{(x,i) \in S_x, (x,j) \in S_x^c} \pi((x,i)) P((x,i),(x,j)) \\
&= \sum_{(x,i) \in S_x, (x,j) \in S_x^c} \frac{1}{n} \pi(x) \cdot P_i(x,x) s(i,j) \\
&\geq \sum_{(x,i) \in S_x, (x,j) \in S_x^c} \frac{1}{n} \pi(x) \cdot \gamma \cdot s(i,j) \\
&\geq \begin{cases} \frac{1}{n} \pi(x) \gamma & \text{if } |S_x| \neq 0, n \\ 0 & \text{if } |S_x| = 0, n \end{cases}
\end{aligned}
$$

Similarly,

$$
Q_{\text{SS}}(S_y, S_y^c) \geq \begin{cases} \frac{1}{n} \pi(y) \gamma & \text{if } |S_y| \neq 0, n \\ 0 & \text{if } |S_y| = 0, n \end{cases}
$$

Now, we bound the amount of flow from x to y for $y \neq x$. Note that $P_i(x,y) = 0$ for all i if x and y differ in more than one variable. As a result, we will assume that x and y differ in only variable i for the next two bounds. For augmented random scan, the following upper bound holds.

$$
\begin{aligned}
Q_{\text{RS}}(S_x, S_y^c) &= \begin{cases} \frac{1}{n} \pi(x) P_i(x,y) \frac{n - |S_y|}{n} & \text{if } (x,i) \in S \\ 0 & \text{if } (x,i) \notin S \end{cases} \\
&\leq \begin{cases} \frac{1}{n} \pi(x) P_i(x,y) & \text{if } (x,i) \in S \text{ and } |S_y^c| \neq 0 \\ 0 & \text{if } (x,i) \notin S \text{ or } |S_y^c| = 0 \end{cases}
\end{aligned}
$$

In the derivation of the next bound, note that if $|S_y^c| = n$, then we are guaranteed that $(y, i + 1 (\text{mod } n)) \in S_y^c$. For augmented systematic scan, the following lower bound holds.

$$
\begin{aligned}
Q_{\text{SS}}(S_x, S_y^c) &= \begin{cases} \frac{1}{n} \pi(x) P_i(x,y) & \text{if } (x,i) \in S \text{ and } (y, i+1) \in S^c \\ 0 & \text{otherwise} \end{cases} \\
&\geq \begin{cases} \frac{1}{n} \pi(x) P_i(x,y) & \text{if } (x,i) \in S \text{ and } |S_y^c| = n \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

17

Now, we can derive several inequalities between the augmented random scan flow and the augmented systematic scan flow as direct consequences of the bounds we just found. First, we bound the relative flows from S_x to S_x^c.

$$Q_{\text{SS}}(S_x, S_x^c) \geq \frac{4\gamma}{n} Q_{\text{RS}}(S_x, S_x^c) \geq \frac{\gamma}{n} Q_{\text{RS}}(S_x, S_x^c)$$

Next, we bound the relative flows from x to y, where x and y differ in exactly variable i.

$$Q_{\text{SS}}(S_x, S_y^c) + \frac{1}{n} P_i(y, x) Q_{\text{SS}}(S_y, S_y^c)$$

$$\geq \begin{cases} \frac{1}{n}\pi(x)P_i(x,y) & \text{if } (x,i) \in S \text{ and } |S_y^c| = n \\ \frac{1}{n} P_i(y,x) \cdot \frac{1}{n}\pi(y)\gamma & \text{if } (x,i) \in S \text{ and } |S_y^c| \neq 0, n \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{1}{n}\pi(x)P_i(x,y) & \text{if } (x,i) \in S \text{ and } |S_y^c| = n \\ \frac{1}{n^2}\pi(x)P_i(x,y)\gamma & \text{if } (x,i) \in S \text{ and } |S_y^c| \neq 0, n \\ 0 & \text{otherwise} \end{cases}$$

$$\geq \begin{cases} \frac{1}{n^2}\pi(x)P_i(x,y)\gamma & \text{if } (x,i) \in S \text{ and } |S_y^c| \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$= \frac{\gamma}{n} \begin{cases} \frac{1}{n}\pi(x)P_i(x,y) & \text{if } (x,i) \in S \text{ and } |S_y^c| \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\geq \frac{\gamma}{n} \Phi_{\text{RS}}(S_x, S_y^c)$$

Finally, we bound the relative flows.

$$Q_{\text{SS}}(S, S^c) = \sum_{x \in \Omega} \sum_{y \in \Omega} Q_{\text{SS}}(S_x, S_y^c) = \sum_{x \in \Omega} \left(Q_{\text{SS}}(S_x, S_x^c) + \sum_{y \neq x} Q_{\text{SS}}(S_x, S_y^c) \right)$$

$$\geq \frac{1}{2} \sum_{x \in \Omega} \left(2 Q_{\text{SS}}(S_x, S_x^c) + \sum_{y \neq x} Q_{\text{SS}}(S_x, S_y^c) \right)$$

$$\frac{1}{2} \sum_{x \in \Omega} \left(Q_{\text{SS}}(S_x, S_x^c) + \sum_{y \in \Omega} \sum_{i=1}^{n} \frac{1}{n} P_i(y, x) Q_{\text{SS}}(S_y, S_y^c) + \sum_{y \neq x} Q_{\text{SS}}(S_x, S_y^c) \right)$$

$$\geq \frac{1}{2} \sum_{x \in \Omega} \left(Q_{\text{SS}}(S_x, S_x^c) + \sum_{y \neq x} \left(Q_{\text{SS}}(S_x, S_y^c) + \sum_{i=1}^{n} \frac{1}{n} P_i(y, x) Q_{\text{SS}}(S_y, S_y^c) \right) \right)$$

$$\geq \frac{1}{2} \sum_{x \in \Omega} \left(\frac{\gamma}{n} Q_{\text{RS}}(S_x, S_x^c) + \frac{\gamma}{n} \sum_{y \neq x} Q_{\text{RS}}(S_x, S_y^c) \right)$$

$$\geq \frac{\gamma}{2n} \sum_{x \in \Omega} \left(Q_{\text{RS}}(S_x, S_x^c) + \sum_{y \neq x} Q_{\text{RS}}(S_x, S_y^c) \right)$$

$$= \frac{\gamma}{2n} Q_{\text{SS}}(S, S^c)$$

The mass of S is the same for augmented random scan and augmented systematic scan, so the same inequality holds for the conductances of the sets. Finally, because this inequality holds for any set S, the inequality also holds for the conductance of the whole chain. □

Theorem 2. *For any lazy or reversible Markov chain,*

$$\frac{1/2 - \epsilon}{\Phi_\star} \leq t_{mix}(\epsilon) \leq \frac{2}{\Phi_\star^2} \log\left(\frac{1}{\epsilon \pi_{min}} \right).$$

18

Proof. The lower bound of this inequality,

$$\frac{1/2 - \epsilon}{\Phi_\star} \leq t_{\mathrm{mix}}(\epsilon)$$

is is shown by Theorem 7.3 of [15] and holds for any Markov chain – that is, it does not actually require the Markov chain to be lazy or reversible.

The upper bound of this inequality,

$$t_{\mathrm{mix}}(\epsilon) \leq \frac{2}{\Phi_\star^2} \log\left(\frac{1}{\epsilon \pi_{\mathrm{min}}}\right),$$

is shown by Theorem 17.10 of [15]. □

Theorem 1. *For any random scan Gibbs sampler R and lazy systematic scan sampler S with the same stationary distribution π, their relative mixing times are bounded as follows.*

$$(1/2 - \epsilon)^2 \, t_{mix}(R, \epsilon) \leq 2 t_{mix}^2(S, \epsilon) \log\left(\frac{1}{\epsilon \pi_{min}}\right)$$

$$(1/2 - \epsilon)^2 \, t_{mix}(S, \epsilon) \leq \frac{8 n^2}{(\min_{x,i} P_i(x,x))^2} t_{mix}^2(R, \epsilon) \log\left(\frac{1}{\epsilon \pi_{min}}\right),$$

where P_i is the transition matrix corresponding to resampling just variable i, and π_{min} is the probability of the least likely state in π.

Proof.

Upper Bound for Random Scan: First, we upper bound the mixing time of random scan.

$$t_{\mathrm{mix}}(R, \epsilon) \leq \frac{2}{\Phi_{\mathrm{RS}}^2} \log\left(\frac{1}{\epsilon \pi_{\mathrm{min}}}\right)$$

Next, we lower bound the mixing time for systematic scan.

$$t_{\mathrm{mix}}(S, \epsilon) \geq \frac{1/2 - \epsilon}{\Phi_{\mathrm{SS\text{-}A}}} \geq \frac{1/2 - \epsilon}{\Phi_{\mathrm{RS}}}$$

This theorem results from algebraic manipulation of the previous two inequalities.

$$t_{\mathrm{mix}}^2(S, \epsilon) \geq \frac{(1/2 - \epsilon)^2}{\Phi_{\mathrm{RS}}^2}$$

$$(1/2 - \epsilon)^2 \, t_{\mathrm{mix}}(R, \epsilon) \leq 2 t_{\mathrm{mix}}^2(S, \epsilon) \log\left(\frac{1}{\epsilon \pi_{\mathrm{min}}}\right)$$

Upper Bound for Systematic Scan: First, we lower bound the mixing time for random scan.

$$t_{\mathrm{mix}}(R, \epsilon) \geq \frac{1/2 - \epsilon}{\Phi_{\mathrm{RS\text{-}A}}}$$

$$t_{\mathrm{mix}}(R, \epsilon)^2 \geq \frac{(1/2 - \epsilon)^2}{\Phi_{\mathrm{RS\text{-}A}}^2}$$

Next, we manipulate the lower bound of Lemma 1.

$$\Phi_{\mathrm{SS}} \geq \frac{1}{2n} \cdot \min_{x,i} P_i(x,x) \cdot \Phi_{\mathrm{RS\text{-}A}}$$

$$\frac{1}{\Phi_{\mathrm{SS}}^2} \leq \frac{4 n^2}{(\min_{x,i} P_i(x,x))^2} \cdot \frac{1}{\Phi_{\mathrm{RS\text{-}A}}^2}$$

Using this result, we upper bound the mixing time for systematic scan.

$$t_{\mathrm{mix}}(S, \epsilon) \leq \frac{2}{\Phi_{\mathrm{SS}}^2} \log\left(\frac{1}{\epsilon\pi_{\min}}\right) \leq \frac{8n^2}{(\min_{x,i} P_i(x,x))^2} \cdot \frac{1}{\Phi_{\mathrm{RS}}^2} \log\left(\frac{1}{\epsilon\pi_{\min}}\right)$$

This theorem results from the previous inequalities.

$$(1/2 - \epsilon)^2 \, t_{\mathrm{mix}}(S, \epsilon) \leq \frac{8n^2}{(\max_{x,i} P_i(x,x))^2} t_{\mathrm{mix}}(R, \epsilon)^2 \log\left(\frac{1}{\epsilon\pi_{\min}}\right)$$

$\square$