Surprising Convergence Properties of Some Simple Gibbs Samplers Under Various Scans

by

Gareth O. Roberts[1] and Jeffrey S. Rosenthal[2]

(February 10, 2015; last revised November 26, 2015.)

Abstract. We examine the convergence properties of some simple Gibbs sampler examples under various scans. We find some surprising results, including Gibbs samplers where deterministic-scan is much more efficient than random-scan, and other samplers where the opposite is true. We also present an example where the convergence takes precisely the same time with *any* fixed deterministic scan, but modifying the scan in any way leads to significantly slower convergence.

Keywords. Gibbs sampler, Markov chain Monte Carlo, convergence rate, random scan, deterministic scan

1. Introduction.

Since their introduction into the Bayesian statistical community by Gelfand and Smith (1990), Gibbs samplers and other Markov chain Monte Carlo (MCMC) algorithms have been very widely studied and used to approximately sample from probability distributions of statistical interest (see e.g. Brooks et al., 2011, and the references therein). One central and ongoing question is their *convergence rate*, i.e. the number of iterations they need to be run to make their distribution close to stationarity. This question has been considered by a number of authors (e.g. Liu et al., 1995; Rosenthal, 1995; Roberts and Sahu, 1997; Papaspiliopoulos and Roberts, 2008), but many questions remain unanswered.

The Gibbs sampler is usually defined for a target probability distribution π on a d-dimensional product state space such as $\mathbf{R}^d$ for $d \geq 2$. (The case $d = 2$ corresponds to the data augmentation algorithm of Tanner and Wong, 1987.) The algorithm proceeds by

[1]Department of Statistics, University of Warwick, CV4 7AL, Coventry, U.K. Email: g.o.roberts@lancaster.ac.uk. Supported in part by EPSRC grants EP/20620/01 and EP/S61577/01.

[2]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: jeff@math.toronto.edu. Web: http://probability.ca/jeff/ Supported in part by NSERC of Canada.

replacing, in turn, the current value of the i^{th} coordinate by a draw from the *conditional* distribution of that coordinate according to π conditional on the current values of all the other coordinates. The choices of the coordinate i to replace can be chosen either in deterministic order (usually by replacing the first coordinate, then the second coordinate, ..., then the d^{th} coordinate), or in random order (usually where the coordinate to replace next is chosen uniformly from $\{1, 2, \ldots, d\}$). The replacements then continue indefinitely. Under mild conditions (see e.g. Tierney, 1994; Roberts and Rosenthal, 2004), the resulting d-dimensional vector will asymptotically converge in distribution to the target stationary distribution π.

In this paper, we consider some very simple and artificial examples of Gibbs samplers, on both discrete and continuous state spaces. We investigate their rates of convergence to stationarity, and find some surprising results, in particular regarding how this rate is affected by various algorithm choices (especially random- versus deterministic-scan).

This question has previously been studied mostly for Gaussian target distributions. In that case, Amit and Grenander (1991) give analytic bounds for rates of convergence for both random and deterministic scan Gibbs samplers, and tentatively conclude in favour of the random scan implementation. However, Roberts and Sahu (1997) give a more detailed comparison for the (statistically very important) family of positively associated Gaussian distributions, showing that for these distributions, the deterministic scan has a uniformly faster rate of convergence than its random scan competitor. In fact this result extends trivially to all partial correlation structures which can be reduced to the positive association case by a succession of transformations which just switch the sign of a given co-ordinate. Thus, even for the Gaussian case, we do not have a full understanding of the problem.

Other work has been successful in comparing random and deterministic scans for specific classes of problems, for instance see Diaconis and Ram (2000) and Diaconis et al. (2008). See also Andrieu (2015) for some recent results in the special case of $d = 2$ coordinates. However, it appears very difficult to postulate general conditions under which one scan outperforms the other. In this paper we shall contribute towards the general heuristic emanating from the Roberts and Sahu study: namely that distributions exhibiting positive associations (which will need to be defined carefully) tend to be explored more rapidly using deterministic scans,

whereas distributions outside this class often prefer random scans. Our results, all from outside the Gaussian framework, support this heuristic, although a completely a general result still eludes us. Nevertheless, we believe our work will be practically beneficial in providing useful rules of thumb for Gibbs sampling implementation.

In this paper, to avoid confusion when comparing different scans, we will always measure convergence times in terms of the total number of individual coordinate *updates* which are required. Thus, one complete iteration (i.e., *sweep*) of a deterministic-scan Gibbs sampler corresponds to d individual updates. This allows for fair comparison between deterministic- and random-scan Gibbs sampler algorithms.

This paper begins (Section 2) by considering the case of target distributions with independent components, where we see that random-scan takes about $\log d$ times as many updates to converge. We next look (Section 3) at a continuous simplex example with pairwise updates whose convergence was recently analysed by Smith (2014), where we show that the first coordinate process converges in $O(d)$ iterations, as opposed to Smith's $O(d \log d)$ full-process convergence time. In Section 4 we present a related discrete-simplex example where random-scan converges much faster ($O(d^2)$) than deterministic-scan. In Section 5 we present a discrete-pyramid example where again random-scan converges much faster ($O(d)$) than deterministic-scan. In Section 6 we present a different discrete-staircase example where the opposite holds, i.e. discrete-scan converges faster (by at least a factor of 2); furthermore, any fixed deterministic-scan coordinate ordering converges equally quickly, but *changing* the coordinate ordering at any stage slows down the convergence dramatically. We close with a discussion (Section 7) of how our simple examples relate to general convergence principles of Roberts and Sahu (1997), and what lessons can be learned regarding the use of Gibbs samplers in more realistic applications.

2. The Independent Case.

We first consider the special case where the stationary distribution π consists of independent components, i.e. has probability density function of the form $\pi(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i)$. In this case, once each coordinate has been updated at least once, the chain has converged to

stationarity.

For a deterministic-scan Gibbs sampler, this necessarily happens after one complete scan, i.e. after precisely d individual updates.

For a random-scan Gibbs sampler, this corresponds to the *coupon collector's problem*, i.e. to how many i.i.d. choices of a coordinate from $\{1, 2, \ldots, d\}$ must be made before each coordinate has been chosen at least once. This is well known to take approximately $d \log d$ updates, and indeed for any $\beta < 1$ the probability of achieving success after $\lfloor \beta\, d \log d \rfloor$ updates goes to 0 as $d \to \infty$ (see e.g. Erdos and Renyi, 1961). This shows:

Conclusion: *If the target distribution has independent components, then the random-scan Gibbs sampler takes* $\log d$ *times as many updates to converge as does the deterministic-scan Gibbs sampler.*

3. Continuous Simplex Example.

We next consider the following simple Markov chain first presented in Aldous and Fill (2002), and later analysed by Smith (2014). Let $\mathcal{X} = \{(x_1, \ldots, x_d) : x_i \geq 0,\ \sum_i x_i = 1\}$ be a d-dimensional simplex in $\mathbf{R}^d$. Let $\{X_n\}$ be the Markov chain on $\mathcal{X}$ defined as follows. Given a state $X_n = (X_{n,1}, X_{n,2}, \ldots, X_{n,d}) \in \mathcal{X}$, first select distinct indices i and j uniformly at random from $\{1, 2, \ldots, d\}$, then choose $\lambda \sim \text{Uniform}[0, 1]$, and then set $X_{n+1,i} = \lambda(X_{n,i} + X_{n,j})$ and $X_{n+1,j} = (1-\lambda)(X_{n,i} + X_{n,j})$, with $X_{n+1,k} = X_{n,k}$ for $k \neq i, j$. These Markov chain dynamics are easily seen to be reversible with respect to $\pi := \text{Uniform}(\mathcal{X})$, so that π is the (unique) stationary distribution for $\{X_n\}$. (This Markov chain $\{X_n\}$ is described in the literature as a "Gibbs sampler"; strictly speaking it is a "block Gibbs sampler", with overlapping blocks – which is necessary since the rigid condition $\sum_i x_i = 1$ does not allow any single coordinate x_i to be updated by itself.)

Smith (2014) analyses this Markov chain in detail. His Theorem 1.1 implies that for any fixed $\epsilon > 0$, there is $c_\epsilon < \infty$ such whenever $n \geq c_\epsilon\, d\, \log d$, the distribution of X_n will be within ϵ of π in total variation distance, i.e. $\|\mathcal{L}(X_n) - \pi\|_{TV} := \sup_A |\mathbf{P}(X_n \in A) - \pi(A)| < \epsilon$. (Total variation distance is a standard way of measuring MCMC convergence; see e.g. Tierney, 1994

or Roberts and Rosenthal, 2004.) Hence, this process converges to stationary in $O(d \log d)$ iterations as $d \to \infty$. (That is, the number of iterations required to get within a fixed $\epsilon > 0$ of stationarity, divided by $d \log d$, remains bounded as $d \to \infty$.)

3.1. The high-dimensional limiting process $\{Y_m\}$.

Since the convergence bounds of Smith (2014) are most relevant as $d \to \infty$, we consider the limiting dynamics of $\{X_n\}$ as $d \to \infty$. As in the MCMC diffusion limits of e.g. Roberts et al. (1997) and Roberts and Rosenthal (1998), we focus on a rescaled version of the first coordinate process $\{X_{n,1}\}$ when the remaining coordinates are in stationarity. Specifically, we let Y_m be d times the value of the first coordinate of X_n after N_m iterations, i.e. $Y_m = d\, X_{N_m,1}$. where N_m is the m^{th} time that the first coordinate of $\{X_n\}$ is updated (i.e. the m^{th} time that one of the coordinate choices i and j is equal to 1). In particular, since at each iteration coordinate 1 is chosen with probability $2/d$, it follows that for large m, N_m is approximately $m/(2/d) = md/2$. That is, $\{Y_m\}$ follows the first coordinate of $\{X_n\}$, except multiplied by a factor of d, and sped up by a factor of approximately $d/2$.

We claim that as $d \to \infty$, the dynamics of this re-scaled process $\{Y_m\}$ are described by $Y_{m+1} = U_m(Y_m + Z_m)$, where Z_m and U_m follow the probability distributions $Z_m \sim$ Exponential(1) and $U_m \sim$ Uniform$[0, 1]$ and are independent. (So, the $\{Y_m\}$ process is similar to an autoregressive process.) Indeed, the logic is as follows. The uniform distribution $\pi = \text{Uniform}(\mathcal{X})$ on a symplex correspondes to a Dirichlet$(1, 1, \ldots, 1)$ distribution, whose one-dimensional marginal distributions are each Beta$(1, d-1)$ with density function $(1-x)^{d-2}$ for $0 < x < 1$. That is, if $X_n \sim \pi$, then each component $X_{n,j}$ has density function $(1-x)^{d-2}$ for $0 < x < 1$. It follows that $d\, X_{n,j}$ has density function proportional to $(1 - x/d)^{d-2}$ for $0 < x < d$. This last density function converges, as $d \to \infty$, to e^{-x} for $x > 0$. That is, as $d \to \infty$, the distribution under π of $d\, X_{n,j}$ converges to the Exponential(1) distribution. Now, when coordinate 1 of $\{X_n\}$ is updated, the update is of the form $X_{n+1,1} = \lambda(X_{n,1} + X_{n,j})$. Hence, $d\, X_{n+1,1} = \lambda(d\, X_{n,1} + d\, X_{n,j})$, i.e. $Y_{m+1} = \lambda(Y_m + d\, X_{n,j})$. Here $\lambda \sim$ Uniform$[0, 1]$, and $d\, X_j \approx$ Exponential(1). This corresponds to the above claimed limiting dynamics, with $U_m = \lambda$ and $Z_m = d\, X_{n,j}$.

The stationary distribution of this process $\{Y_m\}$ is then given by $\pi_Y = \text{Exponential}(1)$. Indeed, this can be checked directly: if $Y_m \sim \pi_Y = \text{Exponential}(1)$ and $Z_m \sim \text{Exponential}(1)$, then since the Exponential(1) probability distribution is the same as the Gamma$(1, 1)$ probability distribution, therefore by a basic property of the Gamma distribution, $Y_m + Z_m \sim \text{Gamma}(2, 1)$ which has density function $g(x) = x\, e^{-x}\, \mathbf{1}_{x>0}$. Also $U_m \sim \text{Uniform}[0, 1]$ with density $h(x) = x\, \mathbf{1}_{0<x<1}$, so by the usual convolution formula for densities, $U_m(Y_m + Z_m)$ has density function given by $f(x) = \int_{-\infty}^{\infty} h(s)\, g(x/s)\, \frac{1}{|s|}\, ds = \int_0^1 g(x/s)\, \frac{1}{|s|}\, ds = e^{-x}$ for $x > 0$, i.e. $U_m(Y_m + Z_m) \sim \text{Exponential}(1) = \pi_Y$, as it should.

3.2. The convergence rates of $\{Y_m\}$ and $\{X_n\}$.

We next observe that the limiting process $\{Y_m\}$ converges to π_Y in $O(1)$ iterations as $d \to \infty$. Indeed, it eventually converges to π_Y since it is ϕ-irreducible and aperiodic (see e.g. Tierney, 1994; Roberts and Rosenthal, 2004). Furthermore the convergence time must be $O(1)$ since the quantity d does not appear in the above description of the dynamics of $\{Y_m\}$.

Now, since index 1 is only selected with probability $2/d$, this means that $N_m \approx md/2$, and in particular time in the N_m scale is $O(d)$ times as large as in the original scale. Hence, in the original time scale, convergence takes $O(d)$ times as long. That is:

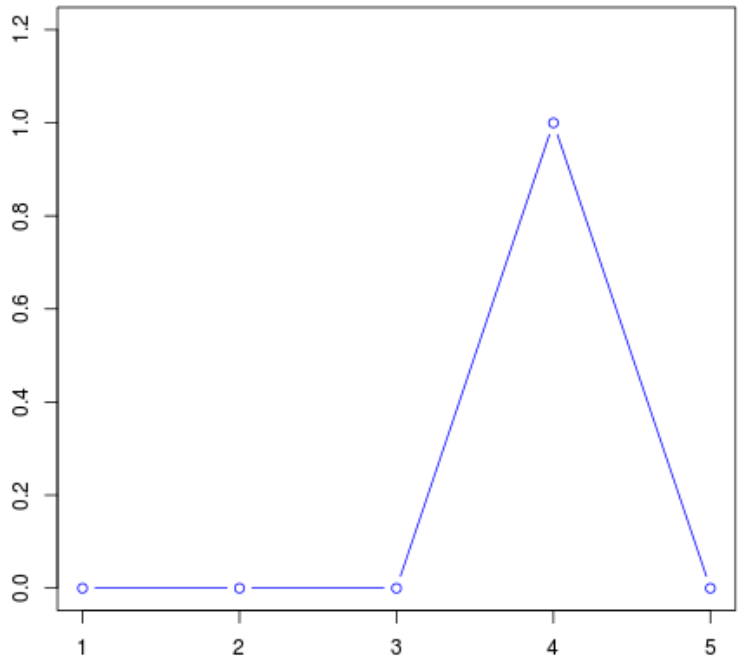
Conclusion: *For the continuous simplex example, in the original time scale, the first coordinate process $\{X_{n,1}\}$ converges to its stationary distribution in $O(d)$ iterations.*

Now, compared to this $O(d)$ convergence result, the $O(d \log d)$ convergence bound of Smith (2014) has an extra factor of $\log d$. This extra factor arises since Smith considers the entire process $\{X_n\}$, while we consider just the first coordinate process $\{X_{n,1}\}$. Indeed, the coordinate are approximately independent as $d \to \infty$. And, in the independent case,

$$\|\mu_1 \times \ldots \times \mu_d - \nu_1 \times \ldots \times \nu_d\|_{TV} \;=\; 1 - \prod_i (1 - \|\mu_i - \nu_i\|_{TV}) \;\lesssim\; \sum_i \|\mu_i - \nu_i\|_{TV}\,.$$

(Here the equality follows by recalling that $1 - \|\mu - \nu\|$ is the maximal probability that $X = Y$ where $X \sim \mu$ and $Y \sim \nu$, and the final approximation follows since if r is small then $1 - r \approx e^{-r}$.) This means that each individual coordinate's total variation distance to stationarity should be about ϵ/d to make the overall total variation distance equal ϵ. This

Figure 1: A Typical State for the Discrete Simplex Example, d=5



requires an additional factor of $\log d$ iterations to achieve. (Another way to think about this is that, by the coupon-collector's problem, about $O(d \log d)$ iterations, not just $O(d)$ iterations, are required to ensure that each coordinate gets selected $O(1)$ times, see e.g. Erdos and Renyi, 1961.)

This simple example thus provides an alternative perspective on the convergence rate results of Smith (2014). However, we wish to focus more on the comparison of different Gibbs samplers with different updating schemes. To do so, we next consider a discrete version of this example.

4. Discrete Simplex Example.

We next consider a discrete version of the previous example. Specifically, let $\mathcal{X} = \{(x_1, \ldots, x_d) \in \{0,1\}^d : \sum_i x_i = 1\}$ be a discrete simplex, so that $|\mathcal{X}| = d$ (see Figure 1). Define a Markov chain on $\mathcal{X}$ as follows. Given a state $X_n = (X_{n,1}, X_{n,2}, \ldots, X_{n,d}) \in \mathcal{X}$, first

select distinct indices i and j uniformly at random from $\{1, 2, \ldots, d\}$. Then, with probability $1/2$ set $X_{n+1} = X_n$. Otherwise, with probability $1/2$, "swap" the i^{th} and j^{th} coordinates by setting $X_{n+1,i} = X_{n,j}$ and $X_{n+1,j} = X_{n,i}$, with $X_{n+1,k} = X_{n,k}$ for $k \neq i, j$. These Markov chain dynamics are again easily seen to be reversible with respect to $\pi := \text{Uniform}(\mathcal{X})$, so that π is the (unique) stationary distribution for $\{X_n\}$.

4.1. Convergence rate.

We next consider the rate of convergence of this process. Suppose it begins with $x_k = 1$. Then the index k is considered for a swap with probability $2/d$ at each iteration, and is then swapped with probability $1/2$. So, the time T until the process first has $x_k = 0$ is distributed as a Geometric random variable with mean about $1/\left[(2/d)(1/2)\right] = d$. But once $x_k = 0$, then the 1 is equally likely to be at any of the other $d-1$ coordinates, so the process is within $\frac{1}{d}$ of stationarity. We conclude that the $\|\mathcal{L}(X_n) - \pi\|_{TV} \leq \mathbf{P}(T > n) + \frac{1}{d} = (1 - \frac{1}{d})^n + \frac{1}{d} \approx e^{-n/d} + \frac{1}{d}$, which is small (as $d \to \infty$) if n is a large multiple of d. That is:

Conclusion: *The discrete simplex example converges in $O(d)$ updates.*

4.2. Deterministic scan modified version.

We next consider a "deterministic scan" version of the above process in which the pairs (i, j) are not chosen at random, but rather are chosen in sequence to be first $(1, 2)$, then $(2, 3)$, then $(3, 4)$, then $\ldots$, then $(d-1, d)$, and then finally $(d, 1)$, before returning to $(1, 2)$ and repeating.

This deterministic-scan version has rather different dynamics. One full deterministic-scan "sweep" of the algorithm now consists of a sequence of $d - 2$ long "wasted" moves where both sites i and $i + 1$ are 0 and thus cannot be changed; followed by a random sequence of moves which involve a possible change. In each of these 'possible move sequences', the algorithm will move the "1" one coordinate backwards with probability $1/2$, leave it unchanged with probability $1/4$, move it one forwards with probability $1/8$, move it two forwards with probability $1/16$, etc. That is, if Z_r is the position of the "1" after n complete

'possible move sequences', then $Z_{r+1} = Z_r - 1 + G_r$ where G_r is a Geometric random variable with mean 1 (and where the arithmetic is done modulo d, to wrap around the circle).

The movement of the "1" in this version thus follows a mean 0 random walk around the circle. Such random walks are well-known (e.g. Diaconis, 1988) to require $O(d^2)$ iterations to converge, as $d \to \infty$. So, we conclude that the deterministic-scan version of this process converges in $O(d^2)$ complete sweeps. Since each sweep corresponds to d individual updates, this implies:

Conclusion: *The deterministic-scan modified version of the discrete simplex example converges in $O(d^3)$ individual updates.*

This indicates that in this example the deterministic-scan version is much less efficient than the random-scan version. Indeed, here random-scan converges in $O(d)$ updates, while deterministic-scan converges in $O(d^3)$ updates. That is, random-scan is more efficient than deterministic-scan by a factor of $O(d^2)$, which is a very substantial improvement. We shall discuss this issue further below.

5. Discrete Pyramid Example.

One limitation of the above examples is that they are not conventional Gibbs samplers which update the coordinates one at a time. Rather, the coordinates had to be updated in blocks of two coordinates at a time, which was necessitated by the rigid condition that $\sum_i x_i = 1$. We now present a modified version of the previous example, which has the same general conclusions, but is a "true" Gibbs sampler which updates the coordinates one at a time.

We now let $\mathcal{X} = \{(x_1, \ldots, x_d) \in \{0,1\}^d : \sum_i x_i \leq 1\}$, so that $\mathcal{X}$ is sort of a discrete "pyramid" rather than a simplex. (Thus, $\mathcal{X}$ contains all states like the one in Figure 1, but also contains the state $(0, 0, \ldots, 0)$.) We again let $\pi = \text{Uniform}(\mathcal{X})$. We then consider the usual (true) Gibbs sampler dynamics, where each coordinate i is updated, one at a time, from its conditional distribution given the current value of all the other coordinates. In this case, to update coordinate i, we proceed as follows: if any other $x_j = 1$ for $j \neq i$ then we must

keep $x_i = 0$, while if all the other $x_j = 0$ then we set $x_i = 1$ or $x_i = 0$ with probability $1/2$ each. Such updates are all reversible with respect to π, so π is again the (unique) stationary probability distribution for this process.

5.1. Random-scan version.

For this Gibbs sampler, the usual random-scan version proceeds at each iteration by choosing i uniformly from $\{1, 2, \ldots, d\}$, and then updating x_i as above. We now analyse the convergence of this random-scan version.

Suppose this version begins with $x_k = 1$. Then at each iteration, the index k is selected for update with probability $1/d$, and if it is selected then x_k is set to 0 with probability $1/2$. So, the time U until the process first has $x_k = 0$ is distributed as a Geometric random variable with mean about $1/\left[(1/d)(1/2)\right] = 2d$. Furthermore, once $x_k = 0$, then the chain is in the state $(0, 0, \ldots, 0)$. From there, the time V until the process leaves the state $(0, 0, \ldots, 0)$ is distributed as a Geometric random variable with mean 1. Furthermore, after $U + V$ iterations the process is uniformly distributed on $\mathcal{X} \setminus \{(0, 0, \ldots, 0)\}$, and hence again is within $\frac{1}{d}$ of stationarity in total variation distance. We conclude, similar to the above, that $\|\mathcal{L}(X_n) - \pi\|_{TV} \leq \mathbf{P}(U + V > n) + \frac{1}{d}$. Furthermore, since U is $O(d)$, and V is $O(1)$, it follows that the process converges to within ϵ of stationarity after $O(d)$ updates:

Conclusion: *The random-scan Gibbs sampler for the discrete pyramid example, starting with $x_k = 1$ for some k, converges in $O(d)$ individual updates.*

(By contrast, if the process happens to *begin* in the state $(0, 0, \ldots, 0)$, then $U = 0$ above, and the convergence then occurs in just $O(1)$ individual updates.)

Remark. Another way to see the above result is to let I be independent of $\{X_n\}$ with $\mathbf{P}(I = 1) = \frac{1}{d+1} = 1 - \mathbf{P}(I = 0)$, and let $T = IU + (1 - I)(U + V) = U + (1 - I)V$, i.e. T is usually equal to $U + V$ but has probability $\frac{1}{d+1}$ of just equalling U. In this case, we have $\mathcal{L}(X_T) = \pi$ exactly. That is, this T is a *strong stationary time* in the sense of Aldous and Diaconis (1987) and Diaconis and Fill (1990). It follows that $\|\mathcal{L}(X_n) - \pi\|_{TV} \leq \mathbf{P}(T > n) \leq \mathbf{P}(U + V > n)$, thus giving a slightly stronger (and still $O(d)$) convergence time bound.

5.2. Deterministic-scan version.

The usual deterministic-scan version of this Gibbs sampler proceeds by updating first x_1, then x_2, then x_3, ..., and then x_d, before returning to x_1 and repeating. We now analyse the convergence of this deterministic-scan version.

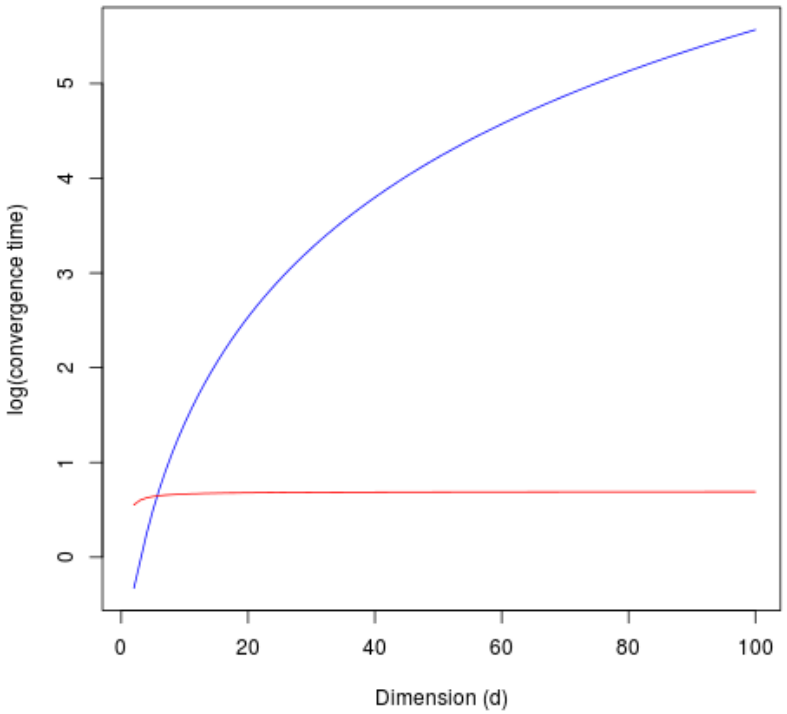
We shall use an argument analogous to that in Subsection 4.2. In this case a "possible move sequence" consists of a sequence of consecutive update steps in which a move might possibly have been made. We describe the effect of a single "possible move sequence". Suppose we begin with $x_k = 1$ for some k, then the updates will change nothing until they reach coordinate k. At this point, x_k will either remain equal to 1 with probability 1/2, or will be changed to 0 with probability 1/2. If it is changed to 0, then the "possible move sequence" will continue until it changes some other coordinate to 1, after which the remaining updates will change nothing. That is, each "possible move sequence" will advance the "1" some distance Z around the circle, where Z is a Geometric random variable with $\mathbf{P}(Z = m) = 2^{-m-1}$ for $m = 0, 1, 2, \ldots$ (and where arithmetic is again done modulo d so it wraps around the circle).

The process thus again corresponds to a random walk on the circle, though this time a walk with positive-mean. However, by subtracting off the mean at each iteration (which does not affect the convergence since π is uniform), it is easily seen that the convergence time for this positive-mean random walk will again be $O(d^2)$ complete sweeps, i.e. $O(d^3)$ individual updates, just like in the mean-0 case above:

Conclusion: *The deterministic-scan Gibbs sampler for the discrete pyramid example converges in $O(d^3)$ individual updates.*

We thus conclude that in this example, as in the previous one, convergence requires $O(d^3)$ updates in the deterministic-scan case, but just $O(d)$ updates in the random-scan case. Hence, even for this true Gibbs sampler, the random-scan version is more efficient by the very large factor of $O(d^2)$.

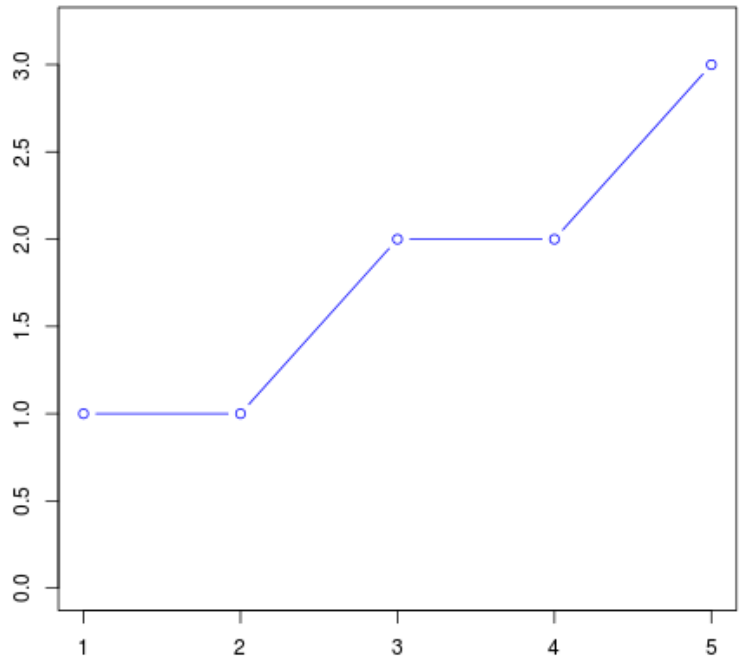
Figure 2: Discrete Pyramid Example: Deterministic v. Random



5.3. Spectral Radius Comparisons.

We have also verified the above comparison directly with numerical computations of the corresponding Markov operator spectral gaps. Specifically, for $1 \leq d \leq 100$, we computed the Markov transition probability matrix for both the random-scan and deterministic-scan versions of the Discrete Pyramid Gibbs sampler example. We then computed the spectral radius of each of these matrices. Finally, we transformed these spectral radius values ρ into convergence time estimates $-1/log(\rho)$, and normalised them so each convergence time was measured in units of d individual updates. Figure 2 shows a plot of the logarithms of the corresponding normalised convergence times for deterministic (top, blue) and random (bottom, red) scans, as a function of the dimension d; it is clear that the convergence time for random-scan is remaining constant on this scale (corresponding to convergence in $O(d)$ individual updates), while the deterministic-scan is growing quickly.

Figure 3: A Typical State for the Discrete Staircase Example, d=5



6. Discrete Staircase Example.

We now instead let

$$\mathcal{X} \;=\; \{(x_1, \ldots, x_d) \in \mathbf{Z}^d : x_1 \in \{0, 1\},\; x_i \in \{x_{i-1}, x_{i-1} + 1\} \text{ for } 2 \leq i \leq d\}\,,$$

so that $\mathcal{X}$ consists of different possible "staircases" which each begin at the origin and move either up one or straight ahead as the index i increases (see Figure 3). We further let $\pi(x) \propto \exp\left(d \sum_i x_i\right)$, so that π gives much more weight to staircases which ascend more quickly, and gives almost all of its weight to the maximal staircase with $x_i = i$ for all i.

We then consider usual (true) Gibbs samplers for this $\mathcal{X}$ and π. In this case, if we attempt to update coordinate i, then if $x_{i-1} = x_i$ and $x_{i+1} = x_i + 1$ then we will increase the value of x_i to $x_i + 1$ with probability nearly 1, while if these conditions are not both satisfied then we will leave x_i unchanged. (For convenience here we take $x_0 = 0$, and ignore x_{d+1} if it arises.)

To fix ideas, we will imagine starting this process in the minimal staircase state $(0, 0, \ldots, 0)$. The question then becomes, how quickly will this process increase from this minimal staircase

$(0, 0, \ldots, 0)$ to the maximal staircase state $(1, 2, \ldots, d)$.

6.1. Usual deterministic scan version.

For this Gibbs sampler, the usual deterministic-scan version proceeds by updating first x_1, then x_2, then x_3, etc. This version benefits from some "cascading effect". Indeed, if we start at the minimal staircase $(0, 0, \ldots, 0)$, then the first complete sweep increases only x_d, the second complete sweep increases x_{d-1} and x_d, the third complete sweep increases x_{d-2} and x_{d-1} and x_d, etc. It follows that in precisely d complete sweeps, i.e. after precisely d^2 individual updates, x_d reaches the value d, and hence the process is in the maximal staircase state and has converged (disregarding exponentially-small probabilities). That is:

Conclusion: *For the discrete staircase example, the usual deterministic-scan Gibbs sampler converges in precisely d^2 individual updates.*

6.2. Other deterministic scan orderings.

We next consider other versions of the deterministic scan Gibbs sampler, with other orderings of the indices.

For the reverse index ordering $d, d-1, \ldots, 1$, the first scan increases each of $x_1, x_2, \ldots, x_d$, the second scan increases $x_2, \ldots, x_d$, the third scan increases $x_3, \ldots, x_d$, etc. Hence, again, in precisely d complete sweeps, i.e. after precisely d^2 individual updates, x_d reaches the value d so the process reaches the maximal staircase giving near-convergence to stationarity.

It turns out, surprisingly, that all other index orderings have updates which remain "sandwiched" between these two extremes, and thus still converge in precisely d complete sweeps, i.e. precisely d^2 iterations. To make this more precise, let $c = (c_1, c_2, \ldots, c_d)$ be any index ordering (i.e., any permutation of $\{1, 2, \ldots, d\}$). Let $z_i = x_i - x_{i-1}$ (with $z_1 = x_1$) record the differences, so each z_i equals either 0 or 1. In this notation, the minimal staircase corresponds to $z_1 = z_2 = \ldots = z_d = 0$, and the maximal staircase corresponds to $z_1 = z_2 = \ldots = z_d = 1$.

Suppose we start this process with all $z_i = 0$, and run a Gibbs sampler with the deterministic scan order c. Then the sampler actually moves from having all $z_i = 0$ to having all

$z_i = 1$ in precisely d complete sweeps, in the following manner. First, find the value d within the scan order c. Then, move right from there within c, decreasing your value to $d-1$ if $d-1$ is to the right of d, and then to $d-2$ if $d-2$ is to the right of $d-1$, etc. Whatever value i you end up with is equal to the first coordinate i to get $z_i = 1$. Then, find the value $i-1$ within the scan order c, and repeat the process, i.e. move right from there and decrease to $i-1$ if $i-1$ is to the right of i, then to $i-2$ if $i-2$ is to the right of $i-1$, etc. Whatever value j you end up with is equal to the second coordinate j to get $z_j = 1$. And so on. Then, once you reach the coordinate 1, then on subsequent sweeps the other z_i become 1 in sequence from left to right. In particular, on each complete deterministic sweep, one additional value z_i changes from 0 to 1, and surprisingly, no such value ever changes from 1 back to 0. So, after d complete sweeps, the process has converged. That is:

Conclusion: *For the discrete staircase example, any fixed-order deterministic-scan Gibbs sampler still converges in precisely d^2 individual updates.*

An example will help clarify the above. Suppose $d = 5$, and the scan is $c = (2, 1, 5, 3, 4)$. Then starting from 5 and moving right, it decreases to 4 (but not to 3), so after one scan, $z_4 = 1$. Then, starting from 3 and moving right, it does not decrease to 2, so after two sweeps $z_2 = z_4 = 1$. Then, starting from 2 and moving right, it decreases to 1, so after three sweeps $z_1 = z_2 = z_4 = 1$. Then after four sweeps $z_3 = z_1 = z_2 = z_4 = 1$, and after five sweeps $z_5 = z_3 = z_1 = z_2 = z_4 = 1$. In terms of the original x_i variables, they progress in turn to: (0,0,0,0,0), (0,0,0,1,1), (0,1,1,2,2), (1,2,2,3,3), (1,2,3,4,4), (1,2,3,4,5), and thus converges in $d = 5$ iterations.

As mentioned above, what is notable about this process is that once $z_i = 1$ for some i, it never returns to 0, which is not at all obvious a priori. Furthermore, this property is *not* preserved if we *change* the scan ordering as we go. That is, using any fixed deterministic scan order, the sampler converges in precisely d iterations. However, if we change scan order as we go (e.g. from the usual order to the inverse order) then it will converge much slower. This suggests that when using deterministic-scan samplers, in this example at least, it is important to keep the update order *consistent*.

6.3. Random-scan version.

The random-scan version of this Gibbs sampler proceeds by choosing the index i uniformly from $\{1, 2, \ldots, d\}$, and then attempting to update x_i conditional on the current configuration. Now, index i can only increase if $x_{i-1} = x_i$ and $x_{i+1} = x_i + 1$, otherwise it does not change. The random-scan algorithm is thus rather inefficient.

To make this more precise, consider just the last two coordinates, x_{d-1} and x_d. In any given configuration, *at most one* of these two coordinates can increase (since if $x_{d-1} = x_d$ then x_{d-1} cannot increase, while if $x_{d-1} = x_d+1$ then x_d cannot increase). So, in expectation, the sum $x_{d-1} + x_d$ increases at most once every d random-scan updates. For convergence, we need to reach the maximal state where $x_{d-1} + x_d = (d-1) + d = 2d - 1$, so in expectation it takes $d(2d-1)$ updates to converge.

To get an actual convergence bound, we can use a simple large deviations principle, see e.g. Theorem 9.3.4 of Rosenthal (2006), to conclude that for any $\epsilon > 0$, the probability of reaching the maximal state after $(1-\epsilon)d(2d-1)$ updates is $\leq \rho^{2d-1}$ where $\rho = \inf_{s>0}[e^{-s(d(2d-1)+\epsilon)}\, M(s)] < 1$; here $M(s) = \mathbf{E}[e^{sG}] = \frac{\frac{1}{d}e^s}{1-(1-\frac{1}{d})e^s}$ is the moment generating function of $G \sim \text{Geometric}(1/d)$ corresponding to the time it takes to increase the value of $x_{d-1} + x_d$ by 1. The important point is that the probability of reaching the maximal state after $(1-\epsilon)d(2d-1)$ updates is exponentially small as a function of d and thus insignificant, i.e. it really does take $d(2d-1)$ updates to converge:

Conclusion: *For the discrete staircase example, the usual random-scan Gibbs sampler requires at least $d(2d-1)$ individual updates.*

Comparing this to the deterministic-scan convergence time, this shows that for this example as $d \to \infty$, random-scan converges more slowly than deterministic-scan by at least a factor of 2. This provides a non-trivial classical Gibbs sampler example where random-scan is clearly less efficient than deterministic-scan. And most interestingly, this conclusion remains for *any* fixed deterministic-scan ordering, but fails if the index ordering is modified in any way during the run.

6.4. Numerical Comparison.

The above shows that for the discrete staircase example, the convergence time for random-scan has a lower bound of about $d(2d-1)$, which is about twice the d^2 convergence time for deterministic-scan. That is, in this example, the ratio of the convergence time for random-scan versus for deterministic-scan is bounded below by $d(2d-1)/d^2 = 2 - (1/d)$, which for large d is about 2. We suspect that for large d the true ratio is actually larger than 2, perhaps growing as $O(\log d)$ as $d \to \infty$. To test this, we repeated 100 simulations of the random-scan sampler in each dimension from 1 to 200, and plotted the ratio of the number of updates required to converge, divided by the d^2 iterations required by random-scan. The results are shown in Figure 4. As expected, this ratio quickly increases above 2 (red line). Now, as the dimension d increases still larger up to 200, the ratio continues to increase very slightly, though it appears to mostly level off at approximately the value 3.78. We are unable to determine whether the ratio grows to infinity as $d \to \infty$, perhaps at the rate $O(\log d)$, or whether it remains bounded.

7. Discussion.

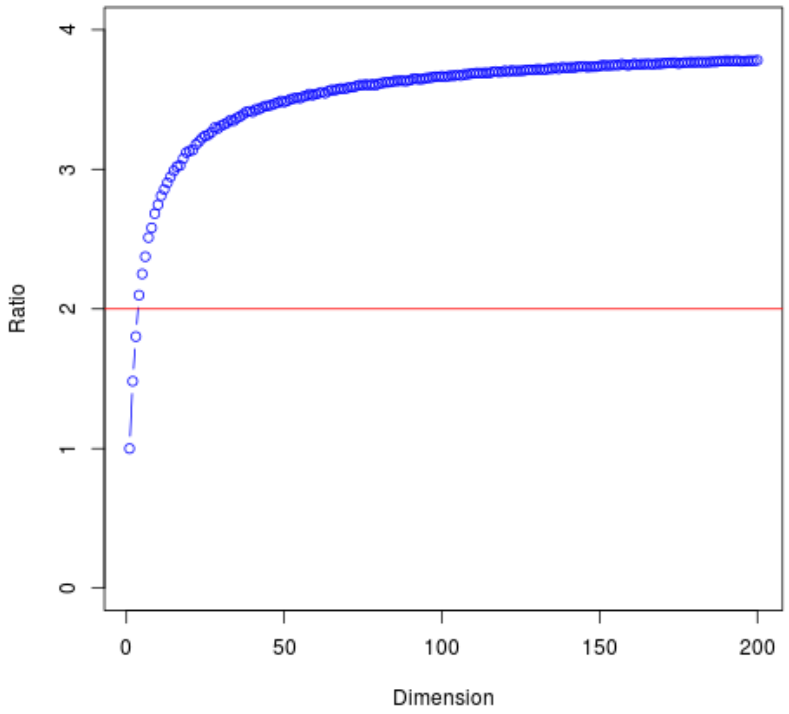
This paper has presented several different simple examples of Gibbs samplers, and considered their convergence times under different scans.

In particular, for the examples of Sections 4 and 5, the random-scan versions are orders-of-magnitude more efficient than deterministic-scan versions. Now, in these examples, $\sum_i x_i$ is constrained, so the different x_i values are *negatively correlated* with each other in stationarity, which may be relevant as we discuss below.

By contrast, for the example of Section 6, deterministic-scan versions (with any ordering) are significantly more efficient than the random-scan version. On the other hand, in that example, x_i is constrained to be within 1 of $x_{i-1},$, so the different x_i values are *positively correlated* with each other.

Taken together, these examples suggest that often random-scan is more efficient when dealing with negative correlations, while deterministic-scan is more efficient when dealing

Figure 4: Ratios of Convergence Times for Discrete Staircase



with positive correlations. This is consistent with the results of Roberts and Sahu (1997), e.g. their Theorem 6 shows that for Gaussian targets with positive correlations, random-scan has larger spectral radius (and hence smaller spectral gap, and hence slower convergence) than does deterministic-scan. Our examples thus provide further support for this rule of thumb, though we are unable to prove it more generality. (For some related results, see e.g. Liu et al., 1995, and Papaspiliopoulos and Roberts, 2008.)

Perhaps most interestingly, for the example of Section 6, *any* fixed deterministic-scan Gibbs sampler is very efficient, but mixing and matching different scan orderings is much worse. So, in this case, the scan ordering chosen doesn't really matter (they all converge in precisely the same number of iterations), but it is essential to keep the ordering consistent.

Other examples further muddy the waters, e.g. consider the following (suggested by A. Smith, personal communication). Let $G = (V, E)$ be a graph with two vertices A and B in the "middle", and d additional vertices in an outer ring, with A and B connected to each other and to each vertex in the outer ring (so A and B each have degree $2d + 1$, while the

other vertices each have degree 2). Consider the "hard-core model" with state space

$$\mathcal{X} = \{f : V \to \{0,1\} \text{ s.t. } f(u)\, f(v) = 0 \ \forall\ (u,v) \in E\}\,,$$

and with stationary distribution $\pi(f) \propto c^{\sum_{v \in V} f(v)}$ for some (large) $c > 0$. (That is, π is biased towards having as many 1's as possible, subject to no two 1's being connected.) Suppose we start at the state with $f(A) = 1$ and all other $f(v) = 0$, and consider the convergence time to reach the (modal) state where $f(v) = 1$ for all v in the outer ring. With the random-scan Gibbs sampler, it takes about cd updates to reduce $f(A)$ to 0, and then about $d \log d$ updates (by the coupon-collector's problem) to visit and hence set to 1 the entire outer ring, for a total time of $cd + d \log d$. For a deterministic-scan ordering in which A is followed by B, it again takes about cd updates to reduce $f(A)$ to 0, but then $f(B)$ is probably immediately set to 1, after which it takes another cd updates to reduce $f(B)$ to 0, and then a further d updates to systematically set the outer ring to 1, for a total time of $2cd + d$. So, if c grows as $O(d)$ or larger, then random-scan is faster by a factor of about 2. However, if $c = O(1)$, then random-scan is slower by a factor of $O(\log d)$. That is, depending on the relation of the parameters c and d, either scan can be superior.

We find all of these results to be surprising and interesting, but admittedly their implications for the Gibbs sampler practitioner are not completely clear. They do suggest that if the target distribution's partial correlation signs are known, then one should perhaps choose deterministic-scan for positive correlations or targets which can be reduced to such by simple sign-changing transformations of individual components. Rather more tentatively they also suggest the use of random-scan where we have at least some negative partial correlations which cannot be removed by sign-changing transformations. However even in the Gaussian case, there is as yet no rigorous theory to back this up.

If the partial correlation signs are unknown, then it is wise to try both versions, and then attempt to estimate (perhaps via convergence diagnostics, see e.g. Gelman and Rubin, 1992) which one is performing more efficiently. Furthermore, when using deterministic-scan algorithms, it may be that a fixed scan ordering should be chosen throughout a simulation, and not be modified during the run.

Our specific examples do not provide definitive information about how Gibbs samplers will perform in other, more complicated contexts. However, they do provide one more piece of information in the complex puzzle of how Gibbs samplers can be used more efficiently and effectively for different target distributions.

References

D. Aldous and P. Diaconis (1987), Strong uniform times and finite random walks. Adv. Appl. Math. **8**, 69–97.

D. Aldous and J.A. Fill (2002), Reversible Markov Chains and Random Walks on Graphs. Unfinished monograph, available at: http://www.stat.berkeley.edu/∼aldous/RWG/book.html

Y. Amit and U. Grenander (1991), Comparing sweep strategies for stochastic relaxation. J. Mult. Anal. **37**, 197–222.

C. Andrieu (2015), A note on one of the Markov chain Monte Carlo novice's questions. Preprint, available at: http://arxiv.org/abs/1504.03467

S. Brooks, A. Gelman, G.L. Jones, and X.-L. Meng, eds. (2011), Handbook of Markov chain Monte Carlo. Chapman & Hall / CRC Press.

P. Diaconis (1988), Group Representations in Probability and Statistics. IMS Lecture Series, volume **11**. Institute of Mathematical Statistics, Hayward, California.

P. Diaconis, K. Khare, and L. Saloff-Coste (2008), Gibbs sampling, exponential families and orthogonal polynomials. Statist. Sci. **23(2)**, 151–178.

P. Diaconis and A. Ram (2000), Analysis of systematic scan Metropolis algorithms using Iwahori-Hecke algebra techniques. Michigan Math. J. **48**, 157–190.

P. Diaconis and J.A. Fill (1990), Strong stationary times via a new form of duality. Ann. Prob. **8**, 1483–1522.

P. Erdos and A. Renyi (1961), On a classical problem of probability theory. Magyar Tudomanyos Akademia Matematikai Kutato Intezetenek Kozlemenyei **6**, 215–220.

A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.

A. Gelman and D.B. Rubin (1992), Inference from iterative simulation using multiple sequences. Stat. Sci. **7(4)**, 457–472.

J.S. Liu, W.H. Wong, and A. Kong (1995), Covariance structure and convergence rate of the Gibbs sampler with various scans. J. Roy. Stat. Soc. Ser. B **57(1)**, 157–169.

O. Papaspiliopoulos and G.O. Roberts (2008), Stability of the Gibbs sampler for Bayesian hierarchical models. Ann. Stat. **36(1)**, 95-117.

G.O. Roberts, A. Gelman, and W.R. Gilks (1997), Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Prob. **7**, 110–120.

G.O. Roberts and J.S. Rosenthal (1998), Optimal scaling of discrete approximations to Langevin diffusions. J. Roy. Stat. Soc. B **60**, 255–268.

G.O. Roberts and S. Sahu (1997), Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. J. Roy. Statis. Soc. Ser. B **59**, 291–317.

J.S. Rosenthal (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.

J.S. Rosenthal (2006), A First Look at Rigorous Probability Theory, 2nd ed. World Scientific Publishing, Singapore.

A. Smith (2014), A Gibbs sampler on the n-simplex. Ann. Appl. Prob. **24(1)**, 114–130.

M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528–550.

L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.