Recent progress on computable bounds and the simple slice sampler

by

Gareth O. Roberts* and Jeffrey S. Rosenthal**

(May, 1999.)

This paper discusses general quantitative bounds on the convergence rates of Markov chains. It then discusses application of these results to simple slice sampler algorithms. It is explained how, in some cases, very useful quantitative results are available, for large classes of target distributions.

1. Introduction.

Markov chain Monte Carlo (MCMC) algorithms – such as the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) – are an extremely popular tool in statistics (see for example the reviews by Smith and Roberts, 1993; Tierney, 1994; Gilks, Richardson, and Spiegelhalter, 1996).

A fundamental problem in MCMC implementation is the question of convergence rate. Specifically, how quickly does the algorithm converge to its target stationary distribution? That is, how long must the algorithm be run before it is approximately distributed in stationarity?

Here we will review recent results on computable bound techniques using Foster-Lyaponov drift inequalities and minorisation conditions. These methods can be applied to MCMC algorithms; here we present an example, the simple slice sampler, where useful bounds can be achieved for a wide variety of different target densities.

* Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, U.K. Internet: G.O.Roberts@lancaster.ac.uk.

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: jeff@math.toronto.edu. Supported in part by NSERC of Canada.

The simple slice sampler is a specialised type of MCMC auxiliary variable method (Swendsen and Wang, 1987; Edwards and Sokal, 1988; Besag and Green, 1993; Higdon, 1998) that has been popularised by Neal (1997), Fishman (1996), and Damien, Wakefield, and Walker (1999). This sampler involves just a single auxiliary variable. It is presented in Section 4.

All of the results described in this paper are proved elsewhere. Most of them can be found in Roberts and Rosenthal (1999a, 1999b), and Roberts and Tweedie (1998, 1999).

2. Convergence rates of Markov chains using drift and minorisation conditions.

Suppose that $X = (X_0, X_1, \ldots)$ is a Markov chain on a state space $\mathcal{X}$ with transition probabilities given by $P(\cdot,\cdot)$. Assuming suitable irreducibility and aperiodicity, and the existence of a stationary probability density π (often known only up to a normalisation constant), what hope have we of assessing the rate of convergence of the Markov chain with minimal information?

Suppose that there exists a *small set* C such that

$$P^{n_0}(x,\cdot) \geq \epsilon\nu(\cdot)\,, \qquad x \in C\,, \tag{1}$$

for some $n_0 \in \mathbf{N}$ and probability measure ν. (This equation is usually applied with $n_0 = 1$.) Suppose also that the Markov chain can be shown to satisfy a *Foster-Lyapunov drift condition*: there exists a function $V : \mathcal{X} \to [1,\infty)$ with

$$\mathbf{E}_x[V(X_1)] \leq \lambda V(x) + b\mathbf{1}_C(x) \tag{2}$$

for some constants $\lambda < 1$ and $0 < b < \infty$.

The existence of a small set and corresponding drift condition implies (see e.g. Meyn and Tweedie, 1993) that the chain is geometrically ergodic:

$$\|P^n(x,\cdot) - \pi\|_{\text{var}} \leq M(x)\rho^n\,, \tag{3}$$

for some constants $\rho < 1$ and $M(x) < \infty$. Here $P^n(\mathbf{x},\cdot)$ is shorthand for $\mathbf{P}\big(\mathbf{X}_n \in \cdot \,|\, \mathbf{X}_0 = \mathbf{x}\big)$; and

$$\|P^n(\mathbf{x},\cdot) - \pi(\cdot)\|_{\text{var}} \;\equiv\; \sup_{A\subseteq\mathbf{R}^d} \left|\mathbf{P}\big(\mathbf{X}_n \in A \,|\, \mathbf{X}_0 = \mathbf{x}\big) - \pi(A)\right|$$

is the *total variation distance* to the stationary distribution π after n steps, when starting from the point $\mathbf{x}$.

This is a useful result. However, as stated it establishes only the *existence* of constants ρ and $M(x)$, without provided any quantitative information about them. Such quantitative information is very important for practical applications.

An initial attempt to provide useful quantitative bounds using the equations (1) and (2) was presented by Meyn and Tweedie (1994), using analytic techniques. Tighter bounds were obtained by Rosenthal (1995), using coupling methodology. Roberts and Tweedie (1999, 1998) later gave refined results, which were of the form

$$\|P^n(x,\cdot) - \pi\|_{\text{var}} \leq (A + Bn)\rho^n ,$$

where A and B are computable functions of λ, b, ϵ, n_0, $V(x)$, and $d \equiv \sup_{y \in C} V(y)$. It is guaranteed that $\rho < 1$ provided that e.g. $2(1-\lambda)(d+1) \geq b$. This result proves geometric ergodicity for these chains. Furthermore, the bound provides quantitative values of the corresponding quantities $M(x)$ and ρ.

The resulting quantitative bounds are very tight for certain Markov chains, but they are less tight (i.e., much too large) for other, more complicated chains.

Remarks.

1. In these results, the notion of small sets can be weakened to that of *pseudo-small* sets. A set C is pseudo-small if we can find $n_0 \in \mathbf{N}$ and a constant $\delta > 0$ (the *coupling constant*) such that for all $x, y \in C$, there is $n_{xy} \leq n_0$ with

   $$\int_{\mathcal{X}} [P^{n_{xy}}(x, dz) \wedge P^{n_{xy}}(y, dz)] \;\geq\; \delta \,.$$

   (For further details see Roberts and Rosenthal, 1997; Roberts and Tweedie, 1998.) Clearly all small sets are pseudo-small, however it is not clear whether the converse holds. It is clear, however, that small sets satisfying (1) can often be pseudo-small for the same n_0 but with a coupling constant δ which is substantially larger than ϵ. This has implications for the tightness of the computable bounds.

2. Bounds on convergence of ergodic averages, of the form

   $$\|\frac{1}{n}\sum_{i=1}^{n} P^n(x,\cdot) - \pi\|_{\text{var}} \;\leq\; A/n \,,$$

can be obtained using only a bound on the shift coupling time

$$T^* = \inf\Big\{t;\ \{X_s, 0 \le s \le t\} \cap \{X'_s, 0 \le s \le t\} \ne \emptyset\Big\}.$$

This approach is typically much easier to apply than are the full coupling techniques. For further details see Roberts and Rosenthal (1996), and Roberts and Tweedie (1998).

3. Bounds for stochastically monotone chains.

To proceed, recall that a Markov chain is *stochastic monotone* (with respect to an ordering $\preceq$ on the state space $\mathcal{X}$) if, for all fixed $z \in \mathcal{X}$, we have:

$$P(X_1 \preceq z | X_0 = x) \ \ge\ P(X_1 \preceq z | X_0 = x') \qquad \text{whenever} \qquad x \preceq x' \,.$$

In the presence of stochastic monotonicity, considerably improved convergence rate bounds are possible. Indeed, suppose that P is stochastically monotone on a totally ordered space $\mathcal{X}$, with the following drift condition holding:

$$PV(x) \le \lambda V(x) + b\mathbf{1}_{x_0}(x)$$

where $x_0 = \min\{x\,;\ x \in \mathcal{X}\}$, with $\lambda < 1$ and $V(\cdot) \ge 1$. Now the coupling construction is much easier, and it can be shown (Lund and Tweedie, 1996) that

$$\rho \le \lambda\ .$$

That is, the convergence rate is bounded directly by the drift parameter λ. Of course, this result makes heavy use of the fact that the drift is down to the atom $\{x_0\}$.

Similarly, suppose

$$PV(x) \le \lambda V(x) + b\mathbf{1}_{x \le c}(x)$$

and $(-\infty, c]$ is small with parameter ϵ as in (1). Reasonable bounds are available in this case as well. Indeed, we have (Roberts and Tweedie, 1998) that

$$\|P^n(x,\cdot) - \pi\|_{\text{var}} \le (A + Bn)\rho^n$$

where A, B, and ρ are computable functions of b, V, ϵ, n_0, and c. In particular we have (taking $n_0 = 1$) that

$$\log \rho = \max\left(\frac{-\log \lambda \log(1-\epsilon)}{\log(\frac{\lambda d + b - \epsilon}{\lambda}) - \log(1-\epsilon)}, \log \lambda\right).$$

Thus, once again, stochastic monotonicity allows for substantial improvement in the resulting quantitative convergence rate results. This fact shall assist us in our analysis of the simple slice sampler.

4. The simple slice sampler.

Let $\pi : \mathbf{R}^d \to \mathbf{R}^{\geq 0}$ be a non-negative integrable function. Our target distribution will be assumed to have density proportional to π. The simple slice sampler begins by choosing a factorisation of π, of the form

$$\pi(\mathbf{x}) = f_0(\mathbf{x})\, f_1(\mathbf{x})\,. \tag{4}$$

By renormalising π as necessary, we can (and do) assume without loss of generality that

$$\sup_{\mathbf{x} \in \mathbf{R}^d} f_1(\mathbf{x}) = 1\,. \tag{5}$$

The f_0-*simple slice sampler* proceeds as follows. Given $\mathbf{X}_n$, we sample a random variable Y_{n+1}, uniformly over the interval $(0, f_1(\mathbf{X}_n))$. We then sample $\mathbf{X}_{n+1}$ from the truncated probability distribution having density proportional to $f_0(\cdot)\, \mathbf{1}_{L(Y_{n+1})}(\cdot)$, where

$$L(y) \;=\; \left\{\mathbf{x} \in \mathbf{R}^d\,;\, f_1(\mathbf{x}) \geq y\right\}\,.$$

The key to this slice sampler is that the joint chain $(\mathbf{X}_n, Y_n)$ has stationary density proportional to $f_0(\mathbf{x})\, \mathbf{1}_{f_1(\mathbf{x}) \geq y}$. Hence, the marginal stationary distribution of $\mathbf{X}_n$ is exactly proportional to π, the target distribution. In this way, the slice sampler can be used to generate an approximate sample from the target distribution. It is thus an example of an MCMC sampling algorithm. (For an interactive simulation of the slice sampler over the internet, see Rosenthal, 1998.)

It is known that the slice sampler has good qualitative convergence properties. For example, the following facts are known.

Proposition 1. *The simple slice sampler satisfies the following:*

(a) (Roberts and Rosenthal, 1999b, Proposition 1) The convergence properties of a simple slice sampler depend only on the quantities

$$Q(y) \equiv Q_{f_0,f_1}(y) = \int_{\mathbf{R}^d} f_0(\mathbf{z}) \mathbf{1}_{\{f_1(\mathbf{z}) \geq y\}} d\mathbf{z} \,. \tag{6}$$

(where $d\mathbf{z}$ is Lebesgue measure on $\mathbf{R}^d$). More precisely, the sequence $\{f_1(\mathbf{X}_n)\}$ is itself a Markov chain, whose transition probabilities depend only on Q, and which is "sufficient" for $\{\mathbf{X}_n\}$ (in the sense of Roberts and Rosenthal, 1998), so that the convergence properties of $\{\mathbf{X}_n\}$ are governed by those of $\{f_1(\mathbf{X}_n)\}$.

(b) (Roberts and Rosenthal, 1999a, Proposition 2) The simple slice sampler is stochastically monotone with respect to the ordering $\preceq$ defined by $\mathbf{x}_1 \preceq \mathbf{x}_2$ if and only if $f_1(\mathbf{x}_1) \leq f_1(\mathbf{x}_2)$.

(c) (Mira and Tierney, 1997, Theorem 6) If f_1 is bounded and $f_1 \in L^1$, then the corresponding uniform slice sampler is uniformly ergodic, *i.e. there is $M < \infty$ and $\rho < 1$ with*

$$\|\mu_k - \pi\| \leq M\rho^k \,, \tag{7}$$

uniformly over choice of the initial distribution μ_0.

(d) (Roberts and Rosenthal, 1999a, Theorem 7) More generally, if π is bounded, and has tails at least as light as $x^{-\alpha}$ for some $\alpha > 1$ (formally, if $Q'(y)y^{1+\frac{1}{\alpha}}$ is non-increasing for sufficiently small y), then the corresponding slice sampler is geometrically ergodic.

Remarks.

1. Another use of the stochastic monotonicity property (b) above is in the construction of *perfect samplers* based upon slice samplers. See Mira, Møller, and Roberts (1999).
2. Mira and Tierney (1997, Theorem 2) have also proved that the simple slice sampler performs uniformly better than the corresponding independence sampler.
3. Simulation studies have shown (Roberts and Rosenthal, 1999b, Section 6) that slice samplers appear to perform very well in specific examples.
4. Whilst the convergence problems of the simple slice sampler are extremely promising, the major problem for the algorithm is the fact that its implementation is restricted by

the need to simulate from f_0 restricted to $L(y)$. This is often problematic, sometimes needing problem specific rejection algorithm techniques to perform the simulation.

Results (c) and (d) of Proposition 1 are very nice general convergence properties, and require only very mild conditions on the target distribution π. However, they do not provide any *quantitative* information about the convergence rate. We consider this question presently.

5. Quantitative convergence rate bounds for the slice sampler.

We shall require the following condition on Q, for some $0 < Y \leq 1$:

$$yQ'(y) \text{ is non-increasing for } y \leq Y\,. \tag{8}$$

In terms of this condition, we can state a general result (Roberts and Rosenthal, 1999b, Theorem 7) about slice sampler convergence. [This result uses general quantitative bound results developed in Roberts and Tweedie (1999, 1998), building on the work of Rosenthal (1995) and Lund and Tweedie (1996).] Recall that we are assuming (5), i.e. that $\sup_{\mathbf{x}\in\mathbf{R}^d} f_1(\mathbf{x}) = 1$.

Proposition 2. *Consider the simple slice sampler for the target density $\pi(\mathbf{x})$, with any factorisation $\pi(\mathbf{x}) = f_0(\mathbf{x})\, f_1(\mathbf{x})$. If (8) holds for some $0 < Y \leq 1$, then for all $\mathbf{x} \in \mathbf{R}^d$, and for all $n \geq \xi$, we have*

$$\|P^n(\mathbf{x}_0, \cdot) - \pi(\cdot)\|_{\text{var}} \;\leq\; K(n + \eta - \xi)\rho^n\ .$$

Here

$$K = \frac{e\, y_*(1-y_*)^{-\xi/\eta}}{\eta} \qquad (\text{where } e = 2.71828\ldots)\,,$$

$$\xi = \frac{\log\left(f_1(\mathbf{x})^{-\beta} + \frac{b}{1-\lambda} - 1\right)}{\log(\lambda^{-1})}, \qquad \eta = \frac{\log\left(\frac{\lambda s + b - y_*}{\lambda(1-y_*)}\right)}{\log(\lambda^{-1})}\,,$$

$s = y^{-\beta}$, *and* $\rho = (1-y_*)^{\eta^{-1}}$. *Furthermore,*

$$\lambda \;\equiv\; \frac{1}{(1-\beta)(1+\alpha\beta)} + \frac{\alpha\beta(y_*/Y)^{\beta}}{1+\alpha\beta}$$

and

$$b \;=\; \frac{Y^{-\beta}(1+\alpha\beta(1-\beta))}{(1-\beta)(1+\alpha\beta)} - \lambda \; .$$

The quantities α, β, and y_* may be chosen freely, provided that $\alpha > 1$, that $0 < \beta < \min\left(\frac{\alpha-1}{\alpha},\ \frac{1}{\alpha}\right)$, and that $y_* \in (0, Y)$.

Indeed, this result follows from choosing the small set $C = \{\mathbf{x} \in \mathbf{R}^d;\ f_1(\mathbf{x}) \geq y_*\}$. This allows us to take $n_0 = 1$ and $\epsilon = y_*$ in (1). The values of λ and b in (2) are bounded by a stochastic comparison argument, taken from Roberts and Rosenthal (1999a, Proposition 6).

This result has many parameters in it, making it difficult to understand. To simplify it, it is suggested (Roberts and Rosenthal, 1999b) to choose $y_* = Y/10$, $\beta = 0.1$, and $\alpha = 10$. Using these choices, we obtain (Roberts and Rosenthal, 1999b, Corollary 4) the following result.

Proposition 3. *If (8) holds for some value of Y with $0 < Y \leq 1$ (i.e., if $yQ'(y)$ is non-increasing for $y \leq Y$), then for all $\mathbf{x}$ such that $f_1(\mathbf{x})/\sup_{\mathbf{w}\in\mathbf{R}^d} f_1(\mathbf{w}) \geq 0.01$ and for all $n \geq n_*(Y)$ with $n_*(Y)$ as in Table 1 below, we have that*

$$\|P^n(\mathbf{x}_0, \cdot) - \pi(\cdot)\|_{\mathrm{var}} \leq 0.01 \; .$$

Y	$n_*(Y)$
1.0	525
0.5	1,400
0.1	10,850
0.01	160,000
0.001	2,075,000

Table 1. Convergence times $n_*(Y)$ as a function of Y.

Now, in general it is not clear when (8) will hold. However, it is proved in Roberts and Rosenthal (1999a) that, in one dimension (i.e. if $d = 1$), if π is log-concave, then the uniform slice sampler will satisfy (8) with $Y = 1$. Hence, from Proposition 3, in such cases the uniform slice sampler will provably converge in just $n_*(1) = 525$ iterations.

This result was extended in Roberts and Rosenthal (1999b), where the *polar factorisation* was proposed, given by $f_0(\mathbf{x}) = |\mathbf{x}|^{-(d-1)}$ and $f_1(\mathbf{x}) = |\mathbf{x}|^{d-1}\pi(\mathbf{x})$. It was proven there that, with this factorisation, if π were log-concave, then condition (8) would be satisfied, with

$$Y \;=\; A(f_1) \;\equiv\; \left[\inf_{|\theta|=1} \sup_{r>0} f_1(r\theta)\right] \Big/ \left[\sup_{|\theta|=1} \sup_{r>0} f_1(r\theta)\right] ;$$

here (r, θ) reproesents a polar-coordinate representation of a vector in $\mathbf{R}^d$. The quantity $A(f_1)$ is thus a measure of the *spherical asymmetry* of the function f_1.

From this observation, it followed that:

Proposition 4. *Suppose that π is a d-dimensional density which is log-concave (at least along rays emanating from the origin). Set $f_0(\mathbf{x}) = |\mathbf{x}|^{-(d-1)}$ and $f_1(\mathbf{x}) = |\mathbf{x}|^{d-1}\pi(\mathbf{x})$, and let $Y = A(f_1)$ be the asymmetry parameter of f_1 as above. Then for any initial value $\mathbf{x}$ such that $f_1(\mathbf{x}) / \sup_{\mathbf{w}\in\mathbf{R}^d} f_1(\mathbf{w}) \geq 0.01$, and for all $n \geq n_*(Y)$ (with $n_*(Y)$ as in Table 1), the polar slice sampler satisfies that*

$$\|P^n(\mathbf{x}_0, \cdot) - \pi(\cdot)\|_{\text{var}} \leq 0.01 \; .$$

i.e. it will be within 1% of its target distribution after at most $n_*(Y)$ iterations.

6. Conclusion.

We see from this review that substantial progress has been made in recent years on the subject of computable quantitative bounds for convergence rates of general Markov chains.

Furthermore, simple slice samplers (including the uniform slice sampler and the polar slice sampler) are particularly amenable to theoretical analysis. A number of interesting theorems are now available which give rather sharp information about their convergence rates, and they appear to be promising algorithms for future applications.

REFERENCES

Besag, J.E. and Green, P.J. (1993), Spatial statistics and Bayesian computation (with discussion). J. Royal Stat. Soc. Ser. B **55**, 25–38.

Damien, P., Wakefield, J.C., and Walker, S. (1999), Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. J. Royal Stat. Soc. Ser. B **61**, 331–344.

Edwards, R.G. and Sokal, A.D. (1988), Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. Phys. Rev. Lett. **38**, 2009–2012.

Fishman, G. (1996), An analysis of Swendsen-Wang and Related Sampling Methods. Technical Report, Department of Operations Research, University of North Carolina.

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

Geman S. and Geman D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721-741.

Gilks W.R. Richardson S. and Spiegelhalter D. J., ed. (1996), Markov chain Monte Carlo in practice. Chapman and Hall, London.

Hastings W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97-109.

Higdon, D.M. (1998), Auxiliary variable methods for Markov chain Monte Carlo with applications. J. Amer. Stat. Assoc. **93**, 585–595.

Lund, R.B. and Tweedie, R.L. (1996), Geometric convergence rates of stochastically ordered Markov chains. Math. Oper. Research **21**, 182–194.

Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., and Teller E. (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087-1091.

Meyn S.P and Tweedie R.L. (1993), Markov chains and stochastic stability. Springer-Verlag, London.

Meyn S.P and Tweedie R.L. (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981-1011.

Mira, A., Møller, J., and Roberts, G.O. (1999), Perfect slice samplers. In preparation.

Mira, A. and Tierney, L. (1997), On the use of auxiliary variables in Markov chain Monte Carlo sampling. Technical Report, School of Statistics, University of Minnesota.

Neal, R. (1997), Markov chain Monte Carlo methods based on 'slicing' the density function. Technical Report, Department of Statistics, University of Toronto.

Roberts, G.O. and Rosenthal, J.S. (1996), Quantitative bounds for convergence rates of continuous time Markov processes. Elec. J. Prob. **1**, paper 9.

Roberts, G.O. and Rosenthal, J.S. (1997), Shift-coupling and convergence rates of ergodic averages. Comm. in Stat. – Stoch. Models **13**, 147–165.

Roberts, G.O. and Rosenthal, J.S. (1998), Sufficient Markov chains. Preprint.

Roberts, G.O. and Rosenthal, J.S. (1999a), Convergence of slice sampler Markov chains. J. Royal Stat. Soc., Series B, to appear.

Roberts, G.O. and Rosenthal, J.S. (1999b), The polar slice sampler. preprint

Roberts, G.O. and Tweedie, R.L. (1998), Rates of convergence for stochastically monotone stochastic processes. Preprint.

Roberts, G.O. and Tweedie, R.L. (1999), Bounds on regeneration times and convergence rates for Markov chains. Stoch. Proc. Appl. **80**, 211–229.

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

Rosenthal, J.S. (1998), Slice Sampler Java Applet.
`http://markov.utstat.toronto.edu/jeff/java/slice.html`

Swendsen, R.H. and Wang, J.S. (1987), Nonuniversal critical dynamics in Monte Carlo simulations. Phys. Rev. Lett. **58**, 86–88.

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