Quantitative convergence rates of Markov chains: A simple account

Jeffrey S. Rosenthal[1]

(Appeared in *Electronic Communications in Probability* **7** (2002), 123–128.)

1 Introduction

Let P be the transition kernel for a Markov chain defined on a state space $\mathcal{X}$. Suppose we run two different copies of the chain, $\{X_n\}$ and $\{X'_n\}$, started (independently or otherwise) from two different initial distributions $\mathcal{L}(X_0)$ and $\mathcal{L}(X'_0)$. We are interested in quantitative upper-bounds on the total variation distance between the two chains after k steps of the chain, which is defined by

$$\|\mathcal{L}(X_k) - \mathcal{L}(X'_k)\|_{TV} \;\equiv\; \sup_{A\subseteq\mathcal{X}} |P(X_k \in A) - P(X'_k \in A)|\,.$$

Such quantitative bounds on convergence rates of Markov chains have been studied in various forms by Meyn and Tweedie (1994), Rosenthal (1995), Roberts and Tweedie (1999), Jones and Hobert (2001), Douc et al. (2002), and others. These investigations have been motivated largely by interest in Markov chain Monte Carlo (MCMC) algorithms including the Gibbs sampler and the Metropolis-Hastings algorithm (see e.g. Gilks et al., 1996), where convergence bounds provide useful information about how long the algorithms must be run to achieve a prescribed level of accuracy.

In this paper, we present one such quantitative bound result. This result is a simplified and improved version of the result in Rosenthal (1995), which also takes into account the ϵ-improvement (i.e., replacing αB_0 by B in the conclusion) of Roberts and Tweedie (1999). This result follows directly as a special case of the more complicated time-inhomogeneous results of Douc et al. (2002). However, the proof we present is very short and simple; and we feel that it is worthwhile to boil the proof down to its essence.

This paper is purely expository; no new results are presented.

2 Assumptions and Statement of Result

Our result requires a *minorisation condition* of the form

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

(i.e. $P(x,A) \ge \epsilon\nu(A)$ for all $x \in C$ and all measurable $A \subseteq \mathcal{X}$), for some probability measure $\nu(\cdot)$ on $\mathcal{X}$, some subset $C \subseteq \mathcal{X}$, and some $\epsilon > 0$.

It also requires a *drift condition* of the form

$$\overline{P}h(x,y) \;\le\; h(x,y)\,/\,\alpha\,, \qquad (x,y) \notin C \times C \tag{2}$$

for some function $h : \mathcal{X} \times \mathcal{X} \to [1,\infty)$ and some $\alpha > 1$, where

$$\overline{P}h(x,y) \;\equiv\; \int_{\mathcal{X}}\int_{\mathcal{X}} h(z,w)\,P(x,dz)\,P(y,dw)\,.$$

Finally, we let

$$B = \max[1,\, \alpha(1-\epsilon)\sup_{C\times C} \overline{R}h]\,, \tag{3}$$

where for $(x,y) \in C \times C$,

$$\overline{R}h(x,y) \;=\; \int_{\mathcal{X}}\int_{\mathcal{X}} (1-\epsilon)^{-2} h(z,w)\,(P(x,dz) - \epsilon\nu(dz))\,(P(y,dw) - \epsilon\nu(dw))\,.$$

[1] Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. jeff@math.toronto.edu. January, 2002; revised May, 2002. Supported in part by NSERC of Canada.

It is easily seen that $B \leq \max[1,\, \alpha(B_0 - \epsilon)]$ where $B_0 = \sup_{(x,y)\in C\times C} \hat{P}h(x,y)$; here $\hat{P} = \epsilon(\nu \times \nu) + (1-\epsilon)\overline{R}$ represents the joint updating of $\{(X_n, X'_n)\}$ in the proof below.

In terms of these assumptions, we state our result as follows.

Theorem 1. Consider a Markov chain on a state space $\mathcal{X}$, having transition kernel P. Suppose there is $C \subseteq \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1,\infty)$, a probability distribution $\nu(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, and $\epsilon > 0$, such that (1) and (2) hold. Define B by (3). Then for any joint initial distribution $\mathcal{L}(X_0, X'_0)$, and any integers $1 \leq j \leq k$, if $\{X_n\}$ and $\{X'_n\}$ are two copies of the Markov chain started in the joint initial distribution $\mathcal{L}(X_0, X'_0)$, then

$$\|\mathcal{L}(X_k) - \mathcal{L}(X'_k)\|_{TV} \;\leq\; (1-\epsilon)^j + \alpha^{-k} B^{j-1}\, E[h(X_0, X'_0)]\,.$$

3 Proof of Result

The proof uses a coupling approach. We begin by constructing $\{X_n\}$ and $\{X'_n\}$ simultaneously using a "splitting technique" (Athreya and Ney, 1978; Nummelin, 1984; Meyn and Tweedie, 1993) as follows.

Let X_0 and X'_0 be drawn jointly from their given initial distribution. We shall let d_n be the "bell variable" indicating whether or not the chains have coupled by time n. Begin with $d_n = 0$. For $n = 0, 1, 2, \ldots$, proceed as follows. If $d_n = 1$, then choose $X_{n+1} \sim P(X_n, \cdot)$, and set $X'_{n+1} = X_{n+1}$ and $d_{n+1} = 1$. If $d_n = 0$ and $(X_n, X'_n) \in C \times C$, then flip (independently) a coin with probability of heads ϵ. If the coin comes up heads, then choose a point $x \in \mathcal{X}$ from the distribution $\nu(\cdot)$, and set $X_{n+1} = X'_{n+1} = x$, and set $d_{n+1} = 1$. If the coin comes up tails, then choose X_{n+1} and X'_{n+1} independently according to the residual kernels $(1-\epsilon)^{-1}(P(X_n, \cdot) - \epsilon\nu(\cdot))$ and $(1-\epsilon)^{-1}(P(X'_n, \cdot) - \epsilon\nu(\cdot))$, respectively, and set $d_{n+1} = 0$. Finally, if $d_n = 0$ and $(X_n, X'_n) \notin C \times C$, then draw $X_{n+1} \sim P(X_n, \cdot)$ and $X'_{n+1} \sim P(X'_n, \cdot)$, independently, and set $d_{n+1} = 0$.

It is then easily checked that X_n and X'_n are each marginally updated according to the transition kernel P. Also, $X'_n = X_n$ whenever $d_n = 1$. Hence, by the *coupling inequality* (e.g. Pitman, 1976; Lindvall, 1992), we have

$$\|\mathcal{L}(X_k) - \mathcal{L}(X'_k)\|_{TV} \;\leq\; P[X_k \neq X'_k] \;\leq\; P[d_k = 0]\,. \tag{4}$$

Now, let

$$N_k \;=\; \#\{m \;:\; 0 \leq m \leq k,\; (X_m, X'_m) \in C \times C\}\,,$$

and let $\tau_1, \tau_2, \ldots$ be the times of the successive visits of $\{(X_n, X'_n)\}$ to $C \times C$. Then for any integer j with $1 \leq j \leq k$,

$$P[d_k = 0] \;=\; P[d_k = 0,\; N_{k-1} \geq j] \;+\; P[d_k = 0,\; N_{k-1} < j]\,. \tag{5}$$

Now, the event $\{d_k = 0,\; N_{k-1} \geq j\}$ is contained in the event that the first j coin flips all came up tails. Hence, $P[d_k = 0,\; N_{k-1} \geq j] \leq (1-\epsilon)^j$, which bounds the first term in (5).

To bound the second term in (5), let

$$M_k \;=\; \alpha^k B^{-N_{k-1}} h(X_k, X'_k)\mathbf{1}(d_k = 0)\,, \qquad k = 0, 1, 2, \ldots$$

(where $N_{-1} = 0$). We claim that

$$E[M_{k+1} \,|\, X_0, \ldots, X_k, X'_0, \ldots, X'_k, d_0, \ldots, d_k] \;\leq\; M_k\,,$$

i.e. that $\{M_k\}$ is a *supermartingale.* Indeed, from the Markov property,

$$E[M_{k+1} \,|\, X_0, \ldots, X_k, X'_0, \ldots, X'_k, d_0, \ldots, d_k] \;=\; E[M_{k+1} \,|\, X_k, X'_k, d_k]\,.$$

Then, if $(X_k, X'_k) \notin C \times C$, then $N_k = N_{k-1}$ and $d_{k+1} = d_k$, so

$$\begin{aligned} E[M_{k+1} \,|\, X_k, X'_k, \{T > k\}] \;&=\; \alpha^{k+1} B^{-N_{k-1}} E[h(X_{k+1}, X'_{k+1}) \,|\, X_k, X'_k]\mathbf{1}(d_k = 0) \\ &=\; M_k\, \alpha E[h(X_{k+1}, X'_{k+1}) \,|\, X_k, X'_k] \,/\, h(X_k, X'_k) \\ &\leq\; M_k\,, \end{aligned}$$

by (2). Similarly, if $(X_k, X'_k) \in C \times C$, then $N_k = N_{k-1} + 1$, so assuming $d_k = 0$ (since if $d_k = 1$ then $d_{k+1} = 1$ so the result is trivial), we have

$$\begin{aligned} E[M_{k+1} \,|\, X_k, X'_k, d_k] &= \alpha^{k+1} B^{-N_{k-1}-1} E[h(X_{k+1}, X'_{k+1}) \, \mathbf{1}(d_{k+1} = 0) \,|\, X_k, X'_k, d_k] \\ &= \alpha^{k+1} B^{-N_{k-1}-1} (1-\epsilon)(\overline{R}h)(X_k, X'_k) \\ &= M_k \, \alpha \, B^{-1} (1-\epsilon)(\overline{R}h)(X_k, X'_k) \,/\, h(X_k, X'_k) \\ &\leq M_k \,, \end{aligned}$$

by (3) since $h \geq 1$. Hence, $\{M_k\}$ is a supermartingale. Then, since $B \geq 1$,

$$\begin{aligned} P[d_k = 0,\ N_{k-1} < j] \;&=\; P[d_k = 0,\ N_{k-1} \leq j-1] \;\leq\; P[d_k = 0,\ B^{-N_{k-1}} \geq B^{-(j-1)}] \\ &= P[\mathbf{1}(d_k = 0) \, B^{-N_{k-1}} \geq B^{-(j-1)}] \\ &\leq B^{j-1} E[\mathbf{1}(d_k = 0) \, B^{-N_{k-1}}] \qquad \text{(by Markov's inequality)} \\ &\leq B^{j-1} E[\mathbf{1}(d_k = 0) \, B^{-N_{k-1}} \, h(X_k, X'_k)] \qquad \text{(since } h \geq 1\text{)} \\ &= \alpha^{-k} B^{j-1} E[M_k] \qquad \text{(by defn of } M_k\text{)} \\ &\leq \alpha^{-k} B^{j-1} E[M_0] \qquad \text{(since } \{M_k\} \text{ is supermartingale)} \\ &= \alpha^{-k} B^{j-1} E[h(X_0, X'_0)] \qquad \text{(by defn of } M_0\text{)} \,. \end{aligned}$$

Theorem 1 now follows from combining these two bounds with (5) and (4).

4 Extensions and Applications

If P has a stationary distribution $\pi(\cdot)$, then in Theorem 1 we can choose $\mathcal{L}(X'_0) = \pi(\cdot)$, so that $\mathcal{L}(X'_k) = \pi(\cdot)$ for all k. Theorem 1 then implies that

$$\|\mathcal{L}(X_k) - \pi(\cdot)\|_{TV} \;\leq\; (1-\epsilon)^j + \alpha^{-k} B^{j-1} \, E[h(X_0, X'_0)] \,,$$

where the expectation is now taken with respect to $X'_0 \sim \pi(\cdot)$. Furthermore, we can allow j to grow with k, for example by setting $j = \lfloor rk \rfloor$ where $0 < r < 1$, to make $(1-\epsilon)^j \to 0$ as $k \to \infty$.

The minorisation condition (1) can be relaxed to a *pseudo-minorisation* condition, where the measure $\nu = \nu_{x,x'}$ may depend upon the pair $(x, x') \in C \times C$ (Roberts and Rosenthal, 2000). More generally, the set $C \times C$ can be replaced by a non-rectangular *ϵ-coupling set* $\overline{C} \subseteq \mathcal{X} \times \mathcal{X}$ (Bickel and Ritov, 2002; Douc et al., 2002). Also, $\overline{P}$ and $\overline{R}$ need not update the two components *independently* as they do above; it is required only that they have the correct marginal distributions (Douc et al., 2002).

The joint drift condition (2) can be derived from univariate drift conditions of the form $PV \leq \lambda V + b$ or $PV \leq \lambda V + b \, \mathbf{1}_C$ in various ways (see e.g. Rosenthal, 2001, Proposition 9); such univariate drift conditions may be easier to identify in specific examples.

Extensions of Theorem 1 have been developed for *stochastically monotone chains* (Lund et al., 1996; Roberts and Tweedie, 2000), for *time-inhomogeneous chains* (Douc et al., 2002; Bickel and Ritov, 2002), for *nearly-periodic chains* (Rosenthal, 2001), and in the context of *shift-coupling* (Aldous and Thorisson, 1993; Roberts and Rosenthal, 1997; Roberts and Tweedie, 1999).

Versions of Theorem 1 have been applied to a number of simple Markov chain examples in Meyn and Tweedie (1994), Rosenthal (1995), and Roberts and Tweedie (1999). They have also been applied to more substantial examples of the Gibbs sampler, including a hierarchical Poisson model (Rosenthal, 1995), a version of the variance components model (Rosenthal, 1996), and some other MCMC examples (Jones and Hobert, 2001). Furthermore, with the aid of auxiliary simulation to only *approximately* verify (1) and (2), approximate versions of Theorem 1 have been applied successfully to more complicated Gibbs sampler examples (Cowles and Rosenthal, 1998; Cowles, 2001).

In spite of these successes in particular applications, it remains true that verifying (1) and (2) for complicated Markov chains is usually a difficult task. Nevertheless, it is of clear theoretical, and sometimes practical, importance to be able to identify convergence bounds solely in terms of drift and minorisation conditions, as in Theorem 1.

Acknowledgement. I am grateful to Randal Douc and Eric Moulines for allowing me to make use of the unpublished work of Douc et al. (2002). I thank the referees for very helpful comments.

REFERENCES

D.J. Aldous and H. Thorisson (1993), Shift-coupling. Stoch. Proc. Appl. **44**, 1-14.

K.B. Athreya and P. Ney (1978), A new approach to the limit theory of recurrent Markov chains. Trans. Amer. Math. Soc. **245**, 493-501.

P.J. Bickel and Y. Ritov (2002), Ergodicity of the conditional chain of general state space HMM. Work in progress.

M.K. Cowles (2001). MCMC Sampler Convergence Rates for Hierarchical Normal Linear Models: A Simulation Approach. Statistics and Computing, to appear.

M.K. Cowles and J.S. Rosenthal (1998), A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. Statistics and Computing **8**, 115–124.

R. Douc, E. Moulines, and J.S. Rosenthal (2002), Quantitative convergence rates for inhomogeneous Markov chains. Preprint.

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

G.L. Jones and J.P. Hobert (2001), Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Statistical Science, to appear.

T. Lindvall (1992), Lectures on the Coupling Method. Wiley & Sons, New York.

R.B. Lund, S.P. Meyn, and R.L. Tweedie (1996), Computable exponential convergence rates for stochastically ordered Markov processes. Ann. Appl. Prob. **6**, 218-237.

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

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

E. Nummelin (1984), General irreducible Markov chains and non-negative operators. Cambridge University Press.

J.W. Pitman (1976), On coupling of Markov chains. Z. Wahrsch. verw. Gebiete **35**, 315–322.

G.O. Roberts and J.S. Rosenthal (1997), Shift-coupling and convergence rates of ergodic averages. Communications in Statistics – Stochastic Models, Vol. **13**, No. **1**, 147–165.

G.O. Roberts and J.S. Rosenthal (2000), Small and Pseudo-Small Sets for Markov Chains. Communications in Statistics – Stochastic Models, to appear.

G.O. Roberts and R.L. Tweedie (1999), Bounds on regeneration times and convergence rates for Markov chains. Stoch. Proc. Appl. **80**, 211–229. See also the corrigendum, Stoch. Proc. Appl. **91** (2001), 337–338.

G.O. Roberts and R.L. Tweedie (2000), Rates of convergence of stochastically monotone and continuous time Markov models. J. Appl. Prob. **37**, 359–373.

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

J.S. Rosenthal (1996), Analysis of the Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput. **6**, 269–275.

J.S. Rosenthal (2001), Asymptotic Variance and Convergence Rates of Nearly-Periodic MCMC Algorithms. Preprint.