Convergence properties of perturbed Markov chains

by

Gareth O. Roberts
Statistical Laboratory
University of Cambridge

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

and
Peter O. Schwartz
Department of Mathematics
University of Toronto

Technical Report No. 9515, November 15, (1995)

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Convergence properties of perturbed Markov chains

by

Gareth O. Roberts*, Jeffrey S. Rosenthal**, and Peter O. Schwartz***

(November 13, 1995.)

Acknowledgements. We thank Neal Madras, Radford Neal, Peter Rosenthal, and Richard Tweedie for helpful conversations. This work was partially supported by EPSRC of the U.K., and by NSERC of Canada.

1. Introduction.

Markov chain Monte Carlo algorithms – such as Gibbs sampler and Metropolis-Hastings – are now widely used in statistics (Gelfand and Smith, 1990; Smith and Roberts, 1993), physical chemistry (Sokal, 1989), and computer science (Sinclair, 1992; Neal, 1993). To explore a complicated probability distribution $\pi(\cdot)$, a Markov chain $P(x, \cdot)$ is defined such that $\pi(\cdot)$ is stationary for the Markov chain. Hopefully, the Markov chain will converge in distribution to $\pi(\cdot)$, allowing for inferences to be drawn.

One potential shortcoming is that the Markov chain is not run analytically but rather by computer simulation. This creates several potential limitations: computers have finite precision and finite range; they use pseudo-randomness rather than true randomness; and they sometimes use algorithms which involve approximations. Thus, rather than running the original chain $P(x, \cdot)$, the computer is in fact running a slightly perturbed chain $\widetilde{P}(x, \cdot)$. This difference is potentially serious, as it might alter the chain's convergence properties,

* Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, U.K. Internet: `G.O.Roberts@statslab.cam.ac.uk`.
** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. Internet: `jeff@utstat.toronto.edu`.
*** Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. Internet: `schwartz@math.toronto.edu`.

convergence rate, and stationary distribution. It is reasonable to ask whether these changes will be significant. To the best of our knowledge, these questions are largely unexplored.

In this paper we begin an analysis of these questions. We are motivated largely by problems associated with finite precision computation (i.e. "roundoff error"). We shall often concentrate on the following model. Let $P(x,\cdot)$ be a family of transition probabilities for a Markov chain on $\mathcal{X}$, where $\mathcal{X}$ is a measurable separable metric space, with metric $\operatorname{dist}(\cdot,\cdot)$. Let $h:\mathcal{X}\to\mathcal{X}$ be a *roundoff function*, so that $h(x)$ is close to x for each $x\in\mathcal{X}$. Let $\widetilde{P}(x,\cdot)$ be a *rounded off chain* defined by

$$\widetilde{P}(x,A) \;=\; P(x,h^{-1}(A))\,. \tag{1}$$

Intuitively, $\widetilde{P}$ proceeds by running the original chain P correctly for one iteration, but then rounding off the result according to the function h. We note that such a $\widetilde{P}$ is not a perturbation of P in any natural operator norm, however we shall still refer to $\widetilde{P}$ as a *perturbation* of P.

(The main example to keep in mind is where $\mathcal{X}=\mathbf{R}^d$, and where h is the function which rounds each coordinate of $\mathbf{x}\in\mathbf{R}^d$ down to the nearest smaller multiple of δ. That is, $h(\mathbf{x})_i=\delta\lfloor x_i/\delta\rfloor$, where $\delta>0$; perhaps $\delta=2^{-31}$.)

We begin with an example (Proposition 1) to show that even if the original chain is geometrically ergodic, the new chain defined by (1) may be transient (and hence not converge at all), no matter how small $\sup_{x\in\mathcal{X}}\operatorname{dist}(h(x),x)$ is. Thus, geometric ergodicity is not preserved in general under small roundoff error.

We then begin proving positive results. We largely concentrate on the case in which P is geometrically ergodic, with stationary distribution $\pi(\cdot)$, which is implied by (Meyn and Tweedie, 1993, Theorems 15.0.1 and 5.5.7) the existence of a π-a.e.-finite function $V:\mathcal{X}\to[1,\infty]$, a subset $C\subseteq\mathcal{X}$, and finite positive numbers β and b, such that the (geometric) *drift condition*

$$\Delta V(x)\le -\beta V(x)+b\mathbf{1}_C(x)\,,\qquad x\in\mathcal{X} \tag{2}$$

holds, where

$$\Delta V(x)\equiv PV(x)-V(x)\equiv\int V(y)P(x,dy)-V(x)\,,$$

and where C is *small* for P, i.e. there is a non-zero measure ν on $\mathcal{X}$, and a positive integer n_0, such that $P^{n_0}(x, A) \geq \nu(A)$, for all $x \in C$ and $A \subseteq \mathcal{X}$. (This condition is equivalent to V-uniform ergodicity, a slightly stronger version of geometric convergence than geometric ergodicity, see Meyn and Tweedie, 1993, Theorem 16.0.1.)

We show that if $\widetilde{P}$ is a perturbation which is close to P in a certain V-related sense, then $\widetilde{P}$ will also be geometrically ergodic, with a stationary distribution and convergence rate bounds close to those for P. Thus, many important convergence properties are preserved in this case.

We also prove that if the drift function V above can be chosen so that $\log V$ is uniformly continuous (which will often be the case in practice, but is not the case for the example of Proposition 1), then for sufficiently small roundoff errors of the form (1), we will again have many important convergence properties preserved. Specifically, $\widetilde{P}$ will again be geometrically ergodic, and will have similar convergence rate bounds to those for P. (This increases the applicability of previous work of Meyn and Tweedie (1994) and Rosenthal (1995), which provides theoretical bounds on convergence times for Markov chain Monte Carlo algorithms on continuous state spaces; specifically, it argues that bounds for exact algorithms will also have application to the slightly perturbed computer-simulated algorithms used in practice.)

Finally, we consider (Section 6) the question of proximity of the stationary distribution for $\widetilde{P}$ to that of P. We show that under suitable uniformity conditions, the two stationary distributions may be made arbitrarily close in total variation norm. For roundoff perturbations this is not the case, however we still show that if $\log V$ is uniformly continuous, then the stationary distributions may be made arbitrarily close in the Prohorov metric (i.e. in the sense of weak convergence).

Many of the results in this paper concern perturbations arising from sufficiently small roundoff error; such results thus depend on an underlying metric on the state space. Other results we give are metric free, relating perturbations which are small in total variation distance to total variation distance proximity of stationary distributions. These results are stronger, but require more restricted conditions on the perturbation being considered. In this context we are motivated by the following example which occurs naturally in the implementation of Metropolis-Hastings algorithms. Suppose that a Markov chain kernel is constructed according to an equation:

$$P(x, dy) = \alpha(x, y)Q(x, dy) + \mathbf{1}_x(y)r(x).$$

Here Q is a Markov chain kernel, α is the *acceptance probability* of a move from x to y, and r is chosen to make P be stochastic. It is common, in the implementation of such algorithms, that the calculation of Q and α is difficult so that errors might occur. Total variation perturbations are relevant to this application.

This paper is organized as follows. Section 2 contains our cautionary example. Section 3 contains convergence results for V-specific perturbations. Section 4 contains convergence results related to roundoff errors. Section 5 extends previous convergence rate bounds to rounded off chains. Finally, Section 6 considers closeness of the stationary distributions of $\widetilde{P}$ and P, using (among other techniques) the notion of *regeneration times* of the two chains.

2. An example of what can go wrong.

We begin with an example of a Markov chain with many nice properties, including geometric convergence, but for which arbitrarily small roundoff errors can lead to transient chains.

Proposition 1. *There exists a Feller continuous, geometrically ergodic Markov chain $P(x, \cdot)$ on the positive real numbers $\mathcal{X} = \mathbf{R}^{>0}$, such that for any $\delta > 0$, there is a one-to-one, onto, continuous function $h : \mathcal{X} \to \mathcal{X}$ with $\sup_{x \in \mathcal{X}} |h(x) - x| = \delta$, such that for the chain $\widetilde{P}(x, \cdot)$ on $\mathcal{X}$ defined by (1), every point $x \in \mathcal{X}$ is transient.*

Proof. Let the Markov chain transition probabilities be defined as follows. For $x \in \mathcal{X}$, let $P(x, \cdot) = \mathcal{U}[a_x, b_x]$ be the uniform distribution on the interval $[a_x, b_x]$, where $a_x = \max(0, x - \frac{4}{x})$ and $b_x = x + \min(\frac{1}{2}, \frac{1}{x})$. Then P is easily seen to be λ-irreducible (where λ is Lebesgue measure), aperiodic, and strong Feller continuous, hence (Meyn and Tweedie, 1993, Theorems 6.2.5 (ii) and 5.5.7) every compact set is small. Furthermore, if we set

$V(x) = xe^{x^2}$, then it is computed that, for $x > 2$,

$$PV(x) \equiv \int_{\mathcal{X}} V(y)P(x,dy) = V(x) \left(\frac{e^{2+x^{-2}} - e^{-8+16x^{-2}}}{10} \right).$$

Now, for $x > 2$ (say), this is less than $0.95\,V(x)$. It follows that P satisfies (2) with $C = [0,2]$. Hence (Meyn and Tweedie, 1993), since P is aperiodic, it has a stationary distribution π and is geometrically ergodic.

On the other hand, given $\delta > 0$, set $h(x) = x + \delta \min(x,1)$. Then h is one-to-one, onto, and continuous, with $\sup\limits_{x \in \mathcal{X}} |h(x) - x| = \delta$. However, if we define $\widetilde{P}$ by (1), then for $x > \max(2, \frac{5}{\delta})$ we will have $\widetilde{P}(x,\cdot)$ supported entirely on $[x + \frac{\delta}{5}, \infty)$. It follows that such x are transient for $\widetilde{P}$. Furthermore, it is easily seen that from any point $x \in \mathcal{X}$, it is possible to reach the set $[\max(2, \frac{5}{\delta}), \infty]$ in a finite number of steps. Hence, every point x is transient. ∎

Remark. In the above example, the perturbed chain $\widetilde{P}$ is not even ϕ-irreducible. On the other hand, if we modify h so that $h(x) = x + \frac{2}{x}$ for large x, then we will have $\widetilde{P}$ ϕ-irreducible but still transient. This shows that the ϕ-irreducibility of $\widetilde{P}$ is not itself sufficient to ensure the ergodicity of $\widetilde{P}$.

This proposition is significant in that it shows that arbitrarily small changes to a well-behaved, ergodic Markov chain may result in a perturbed Markov chain which is transient, and hence does not converge to any distribution at all (much less to a distribution close to the target stationarity distribution π of the original ergodic chain). This poses important questions for the standard computer-simulated use of Markov chain Monte Carlo, and suggests that we seek conditions under which small perturbations to a Markov chain will not alter its properties so drastically. Such is the subject of the remainder of this paper. Indeed, we shall show that the "problem" in the above example is that $\log V$ is not a uniformly continuous function on $\mathcal{X}$.

3. Robustness of geometric ergodicity under perturbations.

We begin with the following elementary V-specific criterion for robustness of the drift condition (2).

Lemma 2. *Suppose a Markov chain P on $\mathcal{X}$ satisfies (2) for some V, C, β, and b. Let $\widetilde{P}$ be a second Markov chain, with $|PV - \widetilde{P}V| \leq \delta V$ for some $\delta < \beta$. Then $\widetilde{P}$ also satisfies (2), for the same V, C, and b, but with β replaced by $\beta - \delta$.*

Proof. We have that

$$\widetilde{P}V \leq PV + \delta V \leq (1 - \beta + \delta)V + b\mathbf{1}_C .$$

The result follows. ∎

This lemma shows that perturbations of P, which have a sufficiently small effect on PV, preserve the drift condition (2) (with suitable modification of β). To study preservation of geometric ergodicity, one must also worry about preservation of ϕ-irreducibility, aperiodicity, and the smallness of C (Meyn and Tweedie, 1993). We have no control over these items in general. However for rounded off chains as given by (1), this is more feasible. Indeed we have

Lemma 3. *If P is a Markov chain on $\mathcal{X}$, and if $\widetilde{P}$ is defined by (1) for some function $h : \mathcal{X} \to \mathcal{X}$, then*

(a) if P is aperiodic, then $\widetilde{P}$ is also aperiodic.

(b) if a subset $C \subseteq \mathcal{X}$ is small for P, then it is also small for $\widetilde{P}$.

Proof. For (a), we note that if $\widetilde{P}$ were periodic, then we could partition $\mathcal{X}$ as $\mathcal{X} = \bigcup_{i=1}^{d} \mathcal{X}_i$, where $\widetilde{P}(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ (where $d \geq 2$, and where we identify $\mathcal{X}_{d+1}$ with $\mathcal{X}_1$). But then the partition $\mathcal{X} = \bigcup_{i=1}^{d} \mathcal{Y}_i$, where $\mathcal{Y}_i = h^{-1}(\mathcal{X}_i)$, would satisfy that $P(x, \mathcal{Y}_{i+1}) = 1$ for all $x \in \mathcal{Y}_i$. Hence P would also be periodic (with at least as large a period).

For (b), choose a non-zero measure ν on $\mathcal{X}$, and a positive integer n_0, such that $P^{n_0}(x, A) \geq \nu(A)$, for all $x \in C$ and $A \subseteq \mathcal{X}$. Then $\widetilde{P}^{n_0}(x, A) \geq \widetilde{\nu}(A)$, where $\widetilde{\nu}(A) =$

$\nu(h^{-1}(A))$. Hence C is small for $\widetilde{P}$. ∎

Note that preservation of ϕ-irreducibility, while it will usually hold, is not automatic. Indeed, the example of Proposition 1 above shows that arbitrarily small roundoff error may result in a chain $\widetilde{P}$ which is not ϕ-irreducible.

Combining the above two lemmas, we have

Theorem 4. *Let P be a geometrically ergodic Markov chain on $\mathcal{X}$, and let V and $\beta > 0$ satisfy* (2) *for P, for some small set C and $0 < b < \infty$. Let $\widetilde{P}$ be a second Markov chain on $\mathcal{X}$, and assume that $|\widetilde{P}V - PV| < \delta V$ for some $\delta < \beta$. Assume further that $\widetilde{P}$ is ϕ-irreducible (for some non-zero σ-finite measure ϕ). Then $\widetilde{P}$ is geometrically ergodic.*

This theorem says, essentially, that the property of geometric ergodicity is robust under perturbations which are small in a certain V-related sense. This is a satisfying result, in that it suggests that approximate simulation of geometrically ergodic Markov chains will again be geometrically ergodic. However, the condition that the perturbation be small in the V-related sense is rather unnatural. It would be much preferred to have conditions saying that the roundoff error be small geometrically, in the sense that the *motion* of the function h, defined by

$$M_h = \sup_{x \in \mathcal{X}} \operatorname{dist}(x, h(x)),$$

be sufficiently small. We consider this topic in the next section.

We close this section by making a connection to certain standard norms on Markov chains. Following Meyn and Tweedie (1993), we define the *V-norm* between two probability measures μ and ν on $\mathcal{X}$, by

$$\|\mu - \nu\|_V = \sup_{\substack{f:\mathcal{X}\to\mathbf{R} \\ |f|\leq V}} \left| \int f(y)\mu(dy) - \int f(y)\nu(dy) \right|.$$

We further define the *V-norm distance* between the Markov chains P and $\widetilde{P}$ on $\mathcal{X}$, by

$$|||P - \widetilde{P}|||_V = \sup_{x \in \mathcal{X}} \frac{\|P(x,\cdot) - \widetilde{P}(x,\cdot)\|_V}{V(x)}.$$

Since $|\widetilde{P}V(x) - PV(x)| \leq |||\widetilde{P} - P|||_V \, V(x)$, we immediately obtain

Corollary 5. *Let P be a geometrically ergodic on $\mathcal{X}$, and let V and $\beta > 0$ satisfy (2) for P, for some small set C and $0 < b < \infty$. Let $\widetilde{P}$ be defined by (1), and assume that $|||\widetilde{P} - P|||_V < \beta$. Assume further that $\widetilde{P}$ is ϕ-irreducible (for some non-zero σ-finite measure ϕ). Then $\widetilde{P}$ is geometrically ergodic.*

4. Robustness of geometric ergodicity under roundoff error.

In general, there need be no connection between the topology of the state space $\mathcal{X}$ (as given by the metric $\mathrm{dist}(\cdot,\cdot)$), and the Markov chain P acting on that space. However, under certain additional continuity assumptions about P, it is possible that being close in a topological sense (e.g. requiring that $M_h < \delta$ in equation (1)) will imply being close in a probabilistic sense (e.g. that $\widetilde{P}$ is also geometrically ergodic). We consider some of these issues in this and the following sections.

We begin by noting that if V is uniformly continuous on $\mathcal{X}$, then it is easily verified that given $\epsilon > 0$, there exists $\delta > 0$, such that if $M_h < \delta$, then $|\widetilde{P}V - PV| < \epsilon \leq \epsilon V$. However, the assumption of uniform continuity of V is quite strong, and will not be satisfied if V is exponential or even quadratic on an unbounded subset of $\mathbf{R}^d$. The following proposition shows that, for our purposes, it suffices to have the weaker (and much more commonly satisfied) condition that $\log V$ is uniformly continuous. (Since $|\log x - \log y| \leq |x - y|$ for $x, y \geq 1$, it follows that $\log V$ is uniformly continuous whenever V is; but the converse obviously does not hold.)

Proposition 6. *Suppose that a Markov chain P satisfies (2) for some small set $C \subseteq \mathcal{X}$, $\beta > 0$, and $b < \infty$, and some function V for which $\log V$ is uniformly continuous on $\mathcal{X}$. Then given $\epsilon > 0$, there is $\delta > 0$ such that, if $\widetilde{P}$ is given by (1), with $M_h < \delta$, then $\widetilde{P}$ also satisfies (2), for the same V and C, but with suitably modified values $\widetilde{\beta}$ and $\widetilde{b}$ satisfying $\widetilde{\beta} > 0$, $|\widetilde{\beta} - \beta| < \epsilon$ and $|\widetilde{b} - b| < \epsilon$.*

Proof. Given the value of β for P, choose $\alpha > 0$ with $\alpha < \log\left(\frac{1}{1-\beta}\right)$ and with $e^\alpha - 1 < \min\left(\frac{\epsilon}{1-\beta}, \frac{\epsilon}{b}\right)$. Then find $\delta > 0$ such that $\mathrm{dist}(x, y) < \delta$ implies that $|\log V(y) -$

$\log V(x)| < \alpha$. Then if $M_h < \delta$, we have that

$$\begin{aligned}
\widetilde{P}V(x) - PV(x) &= \int_{\mathcal{X}} (V(h(y)) - V(y))\ P(x, dy) \\
&= \int_{\mathcal{X}} \left(e^{\log V(h(y))} - e^{\log V(y)}\right)\ P(x, dy) \\
&\leq \int_{\mathcal{X}} \left(e^{(\log V(y))+\alpha} - e^{\log V(y)}\right)\ P(x, dy) \\
&= (e^{\alpha} - 1)\ PV(x)\,.
\end{aligned}$$

Hence,

$$\widetilde{P}V(x)\ \leq\ e^{\alpha}\,PV(x)\,.$$

But by assumption, $PV \leq (1-\beta)V + b\mathbf{1}_C$. Hence,

$$\widetilde{P}V\ \leq\ (1-\widetilde{\beta})V + \widetilde{b}\mathbf{1}_C\,,$$

where $\widetilde{\beta} = 1 - e^{\alpha}(1-\beta)$ and $\widetilde{b} = e^{\alpha}b$. Since $\alpha < \log\left(\frac{1}{1-\beta}\right)$, it follows that $\widetilde{\beta} > 0$. Furthermore, we compute that $|\widetilde{\beta} - \beta| = (e^{\alpha}-1)(1-\beta) < \epsilon$ and $|\widetilde{b} - b| = (e^{\alpha}-1)b < \epsilon$, completing the proof. ∎

Combining the above propositions with Theorem 4, we obtain our desired result. To state it in the strongest possible form, we make the following definition. We say that a class of Markov chain kernels $\{P_c,\ c \in \mathcal{C}\}$ is *simultaneously geometrically ergodic* if there exists a class of non-zero sigma-finite measures, $\{\phi_c,\ c \in \mathcal{C}\}$, a class of probability measures $\{\nu_c,\ c \in \mathcal{C}\}$, a measurable subset $C \subseteq \mathcal{X}$, a real-valued measurable function $V \geq 1$, a positive integer n_0, and positive constants η, β, and b, such that for each $c \in \mathcal{C}$:

(i) P_c is ϕ_c-irreducible;

(ii) C is small for P_c, with $P_c^{n_0}(x,\cdot) \geq \eta\,\nu_c(\cdot)$ for all $x \in C$;

(iii) the chain P_c satisfies the drift condition (2), with drift function V and small set C.

(The reason for this somewhat complicated definition is that by standard regeneration theory, we can give upper bounds on moments of regeneration times which can be taken to hold uniformly over all processes in the class; see Section 6.)

We then have

Theorem 7. *Let P be geometrically ergodic on $\mathcal{X}$, and let it satisfy (2) for some small set C and continuous function V, such that $\log V$ (or V) is uniformly continuous on $\mathcal{X}$. Then there is $\delta > 0$ such that, if $\widetilde{P}$ is given by (1) with $M_h < \delta$, and if $\widetilde{P}$ is ϕ-irreducible for some non-zero ϕ, then $\widetilde{P}$ is geometrically ergodic. Furthermore, the class of all such $\widetilde{P}$ is simultaneously geometrically ergodic.*

This theorem provides a useful criterion under which geometric ergodicity will be insensitive to small roundoff error. However, the theorem does require that $\log V$ be uniformly continuous, and while that condition usually holds in practice, it is not clear when this is guaranteed.

Often one can explicitly construct a function V together with a drift condition (2), such that $\log V$ is uniformly continuous; see for example Rosenthal (1994). Also, in Roberts and Tweedie (1994, Theorem 3.3), it is shown that one can sometimes use the function $V(x) = f_\pi(x)^{-1/2}$, where f_π is a density function for π; in such cases one will often have $\log V$ uniformly continuous.

In general, Meyn and Tweedie (1993, Theorem 15.2.4) show that the function

$$V(x) = \mathbf{E}_x\left(\sum_{k=0}^{\sigma_C} r^k\right)$$

will satisfy (2), where $\sigma_C = \inf\{n \geq 1\,;\, X_n \in C\}$, and where $r > 1$ is chosen to satisfy that $\sup_{x\in\mathcal{X}} \mathbf{E}_x(r^{\sigma_C}) < \infty$. Furthermore we may then take $\beta = 1 - r^{-1}$. They show (Meyn and Tweedie, 1993, Proposition 6.1.1 (ii)) that if the Markov chain is *weak Feller continuous*, then the above function V will at least be *lower semicontinuous*. However, no uniformity is provided, and it is not clear for such V when $\log V$ would be uniformly continuous on $\mathcal{X}$.

Remark. Many of the roundoff results we give in this paper are given with respect to a particular metric. For instance, the crucial log-Lipshitz property is clearly metric-dependent. At first sight this may seem unsatisfactory, since geometric convergence is a metric-free property. However if we consider a sequence of perturbations with round-off functions h_k, $k \geq 1$, then it is often the case that the metric with respect to which V is required to be log-Lipshitz is actually intrinsically defined by the sequence h_k, $k \geq 1$. Specifically, suppose that we set $M_h^\rho = \sup_{x \in \mathcal{X}} \rho(x, h(x))$ for a given metric $\rho(\cdot, \cdot)$ on $\mathcal{X}$. Let $\mathcal{M}$ be the class of all metrics ρ on $\mathcal{X}$ such that

$$\lim_{k \to \infty} M_h^\rho = 0.$$

The conclusions of Theorems 7 and 9 therefore hold if there exists $\rho \in \mathcal{M}$ with respect to which V is uniformly log-Lipshitz.

5. Rate of convergence.

A number of results (Meyn and Tweedie, 1994; Rosenthal, 1995) are available which provide bounds on the distance to stationarity of a Markov chain after k steps, using minorization and drift conditions. Such results consider the exact Markov chain P, and it is reasonable to ask if the results will apply to a slightly perturbed chain $\widetilde{P}$ as simulated by computer. Our results above provide some reassuring answers to such questions.

For example, a result of Rosenthal (1995, Theorem 12) says that for a Markov chain $P(x, \cdot)$ with stationary distribution π, and stationary distribution μ_0, if there exist $\eta > 0$, $0 < \beta < 1$, $0 < \Lambda < \infty$, $d > \frac{2\Lambda}{\beta}$, a non-negative function $f : \mathcal{X} \to \mathbf{R}$, and a probability measure $Q(\cdot)$ on $\mathcal{X}$, such that

$$Pf(x) \;\leq\; (1-\beta) f(x) + \Lambda\,, \qquad x \in \mathcal{X}$$

and

$$P(x, \cdot) \;\geq\; \eta\, Q(\cdot)\,, \qquad x \in f_d$$

where $f_d = \{x \in \mathcal{X} \,|\, f(x) \leq d\}$, then for any $0 < r < 1$, the total variation distance to stationarity after k steps satisfies

$$\|\mu_0 P^k - \pi\|_{\text{var}} \;\leq\; (1-\eta)^{rk} + \left(\alpha^{-(1-r)} \gamma^r\right)^k \left(1 + \frac{\Lambda}{\beta} + \mathbf{E}_{\mu_0}(f(X_0))\right),$$

where

$$\alpha^{-1} \;=\; \frac{1 + 2\Lambda + (1-\beta)d}{1+d} \;<\; 1; \quad \gamma \;=\; 1 + 2((1-\beta)d + \Lambda)\,.$$

Now, if $\widetilde{P}$ is defined by (1), then (as in the proof of Lemma 3 (b) above) we will have $\widetilde{P}(x, \cdot) \geq \eta\, \widetilde{Q}(\cdot)$ for $x \in f_d$, where $\widetilde{Q}(A) = Q(h^{-1}(A))$. Hence, using Proposition 6 above (with $C = \mathcal{X}$), we obtain

Theorem 8. *Let $P(x,\cdot)$ be the transition probabilities for a Markov chain on a state space $\mathcal{X}$, with stationary distribution π, such that there exist $\eta > 0$, $0 < \beta < 1$, $0 < \Lambda < \infty$, $d > \frac{2\Lambda}{\beta}$, a non-negative function $f : \mathcal{X} \to \mathbf{R}$, and a probability measure $Q(\cdot)$ on $\mathcal{X}$, with $Pf(x) \leq (1-\beta)f(x) + \Lambda$ for $x \in f_d$, and $P(x,\cdot) \geq \eta\, Q(\cdot)$ for $x \in \mathcal{X}$. Assume that $\log f$ is uniformly continuous on $\mathcal{X}$. Then for any $\epsilon > 0$ such that $d > \frac{2(\Lambda+\epsilon)}{\beta-\epsilon}$, there is $\delta > 0$, such that if $\widetilde{P}$ is defined by (1), and if $M_h < \delta$, then*

$$\|\mu_0 \widetilde{P}^k - \pi\|_{\mathrm{var}} \;\leq\; (1-\eta+\epsilon)^{rk} + \left(\widetilde{\alpha}^{-(1-r)}\widetilde{\gamma}^{r}\right)^k \left(1 + \frac{\Lambda+\epsilon}{\beta-\epsilon} + \mathbf{E}_{\mu_0}(f(X_0))\right),$$

where

$$\widetilde{\alpha}^{-1} \;=\; \frac{1+2(\Lambda+\epsilon)+(1-\beta+\epsilon)d}{1+d}\,; \quad \widetilde{\gamma} \;=\; 1+2\Big((1-\beta+\epsilon)d+\Lambda+\epsilon\Big).$$

Other similar convergence rate bounds (e.g. Meyn and Tweedie, 1994) can be similarly modified. The main point is that, if the logarithm of the drift function is uniformly continuous, then rate bounds will be robust under small roundoff errors or other perturbations.

6. Robustness of the stationary distributions

In this section we consider the issue of whether the stationary distribution $\widetilde{\pi}$ of the perturbed chain $\widetilde{P}$ will be close to the stationary distribution π of the original chain P. For sufficiently small perturbations in total variation distance or in the roundoff metric M_h, we shall show that $\widetilde{\pi}$ and π may be made arbitrarily close in an appropriate metric. In the roundoff error case, since the range of the roundoff function h (and hence the support of $\widetilde{\pi}$) might, say, be discrete, it will *not* be true in general that π and $\widetilde{\pi}$ are close in total variation distance. Thus, for this case we shall instead consider weak convergence, written $\Rightarrow$, and metrized by the Prohorov metric (Ethier and Kurtz, 1986, Section 3.1), defined by

$$d(\mu,\nu) = \inf\{\epsilon\,;\, \mathbf{P}\left(\mathrm{dist}(X,Y) \geq \epsilon\right) \leq \epsilon \text{ for some } (X,Y) \in \mathcal{D}_{\mu,\nu}\},$$

where $\mathcal{D}_{\mu,\nu}$ is the collection of all random variable pairs (X,Y) taking values in $\mathcal{X}$ with laws given by $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$. Recall also that $\mu_k \Rightarrow \mu$ is equivalent to saying that $\int f d\mu_k \to \int f d\mu$, for each uniformly continuous bounded function $f : \mathcal{X} \to \mathbf{R}$ (Ethier and Kurtz, 1986, p. 108). Finally, it is easily seen that

$$\|\mu - \nu\|_{\mathbf{var}} \le d(\mu, \nu), \tag{3}$$

where $\|\mu - \nu\|_{\mathbf{var}} = \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|$ is total variation distance.

We prove the two results using different approaches. While the approach for the roundoff case (Theorem 11) is quite direct, for the V-norm perturbation (Theorem 9), we shall use the representation of π in terms of *tours* between *renewal times* of the Markov chain (Athreya and Ney, 1978; Asmussen, 1987; Meyn and Tweedie, 1993).

Theorem 9. *Suppose that there exists a sequence of Markov chain kernels $P_1, P_2, \ldots$ and P_∞ on a state space $\mathcal{X}$, satisfying the following conditions.*

(i) The kernels $\{P_k\}$ are simultaneously geometrically ergodic, as defined just before Theorem 7, with small set C and drift function V satisfying that $\sup_{x \in C} V(x) < \infty$.

(ii) For all $x \in \mathcal{X}$,

$$\lim_{k \to \infty} \|P_k(x, \cdot) - P_\infty(x, \cdot)\|_{\mathbf{var}} = 0.$$

Then the stationary distributions of the P_k, $\pi_k(\cdot)$ say, satisfy that

$$\lim_{k \to \infty} \|\pi_k(\cdot) - \pi_\infty(\cdot)\|_{\mathbf{var}} = 0.$$

Remark. (i) and (ii) are implied by convergence of the kernels in the V-norm discussed in Section 4 under the log-Lipshitz condition on V.

Proof. Our approach shall be to consider regeneration times constructed for the processes by a standard splitting argument (see Nummelin, 1984; Meyn and Tweedie, 1993). Accordingly consider the usual splitting. Let $\mathcal{X}' = \mathcal{X} \cup C'$ where C' is a copy of C. For each $k \in \mathbf{N} \cup \infty$, set $P'_k(x, A) = P_k(x, A)$ for $x \in C^c$, $A \subseteq C^c$; $P'_k(x, A) = (1 - \eta) P_k(x, A)$ for $x \in C^c$, $A \subseteq C$; $P'_k(x, A) = \eta P_k(x, A)$ for $x \in C^c$, $A \subseteq C'$; $P'_k(x, A) = Q(A)$ for $x \in C'$; $P'_k(x, A) = (1 - \eta)^{-1}(P_k(x, A) - \eta Q(A))$ for $x \in C$.

In this setup, C' is an atom, and therefore the return times to C' are regenerations. By (ii), we can construct a probability space, simultaneously supporting processes $X^k =$

$(X_0^k, X_1^k, \ldots)$, $k \in \mathbf{N} \cup \{\infty\}$, with respective transition kernels P_k', such that $X_0^k \in C'$, and such that if τ_k denotes the first non-zero return time of the process X^k to the set C', then

$$\lim_{k \to \infty} \mathbf{P}(X_s^k = X_s^\infty, 0 \le s \le \tau_k = \tau_\infty) \ = \ 1 \, .$$

(For details see Jacka and Roberts, 1995.)

Let f be a function on $\mathcal{X}$ (extended to C' by copying its values from C), such that $|f| \le 1$. Then standard regeneration theory (Athreya and Ney, 1978; Asmussen, 1987) allows us to write

$$\pi_k(f) = \frac{\mathbf{E}(\sum_{i=1}^{\tau_k} f(X_i^k))}{\mathbf{E}(\tau_k)} .$$

Now, as $k \to \infty$, we have almost sure convergence of the integrands in both the numerator and the denominator to the corresponding values for $k = \infty$. Thus, to establish our result, it suffices to prove uniform integrability of each integrand.

However, by (i), and by Lemma 4 of Rosenthal (1995), there exists a constant $R > 1$ (which may depend on β, b, η, and n_0), such that

$$\mathbf{E}(R^{\tau_k}) \le \sup_{x \in C} V(x),$$

for all $k \in \mathbf{N}$. This implies the uniform integrability of the sequence $\{\tau_k\}$.

Now, using that $\mathbf{E}(\tau_k) \ge 1$ and $\mathbf{E}(\sum_{i=1}^{\tau_\infty} f(X_i^\infty)) \le \mathbf{E}(\tau_\infty)$, we obtain that

$$\begin{aligned} \left|\pi_k(f) - \pi_\infty(f)\right| &\le \left|\mathbf{E}(\sum_{i=1}^{\tau_k} f(X_i^k)) - \mathbf{E}(\sum_{i=1}^{\tau_\infty} f(X_i^\infty))\right| + \left|\mathbf{E}(\tau_k) - \mathbf{E}(\tau_\infty)\right| , \\ &\le \mathbf{E}\big((\tau_k + \tau_\infty)\mathbf{1}_{X^\infty|_{\{0,\ldots,\tau_\infty\}} \ne X^k|_{\{0,\ldots,\tau_k\}}}\big) + \left|\mathbf{E}(\tau_k - \tau_\infty)\right| , \end{aligned}$$

which goes to zero uniformly for $|f| \le 1$ by the uniform integrability of the τ_k. The result follows. ∎

In order to consider the roundoff case, it is necessary to impose some continuity structure on the Markov chains under consideration. Therefore recall the Feller properties. $P(\cdot, \cdot)$ is *weak Feller continuous* if Pf is continuous for all continuous bounded functions f. It is *strong Feller continuous* if Pf is continuous for all bounded functions f.

Lemma 10. *Suppose that $P_k(\cdot,\cdot), k \in \mathbf{N} \cup \{\infty\}$ are weak Feller continuous transition functions such that for all $x \in \mathcal{X}$,*

$$P_k(x,\cdot) \Rightarrow P_\infty(x,\cdot). \tag{4}$$

Suppose also that convergence in (4) holds uniformly over compacts, in the sense that for each uniformly continuous bounded function $f : \mathcal{X} \to \mathbf{R}$, and for each compact subset $A \subseteq \mathcal{X}$,

$$\lim_{k\to\infty} \sup_{x\in A} |P_k f(x) - P_\infty f(x)| = 0\,. \tag{5}$$

Then, for all $x \in \mathcal{X}$ and $n \in \mathbf{N}$, we have $P_k^n(x,\cdot) \Rightarrow P_\infty^n(x,\cdot)$, i.e. the higher-order iterates converge weakly.

Proof. We shall proceed by induction on n. By hypothesis the conclusion holds for $n = 1$. Suppose that it holds for $n-1$. Let $f : \mathcal{X} \to \mathbf{R}$ be a uniformly continuous bounded function, and let $x \in \mathcal{X}$. We can write

$$P_k^n f - P_\infty^n f = P_k^{n-1}(P_k f - P_\infty f) + (P_k^{n-1} - P_\infty^{n-1})P_\infty f.$$

By assumption, $P_k f - P_\infty f \to 0$ uniformly on compacts. Also, by the inductive hypothesis, the measures $\{P_k^{n-1}(x,\cdot),\ k \in \mathbf{N}\}$ are tight. It therefore follows that

$$P_k^{n-1}(P_k f - P_\infty f)(x) \longrightarrow 0\ ,$$

as $k \to \infty$. Also, by the weak Feller property $P_\infty f$ is continuous and bounded, so that by the induction hypothesis,

$$(P_k^{n-1} - P_\infty^{n-1})P_\infty f(x) \longrightarrow 0\ ,$$

as $k \to \infty$. It follows that $P_k^n f(x) \to P_\infty^n f(x)$ as $k \to \infty$, for any uniformly continuous bounded f. Therefore $P_k^n(x,\cdot) \Rightarrow P_\infty^n(x,\cdot)$, for each $x \in \mathcal{X}$, as required. ∎

Theorem 11. *Suppose that h_k, $k \in \mathbf{N}$ is a sequence of roundoff functions with $\lim_{k\to\infty} M_{h_k} = 0$. Let P be strong Feller continuous, and let P_k denote the successive perturbations $P_k(x,\cdot) = P(x, h_k^{-1}(\cdot))$. Suppose P is geometrically ergodic and satisfies (2), for a log-Lipschitz function V and small set C. Suppose further that P_k is ϕ_k-irreducible for each k. Then for all large enough k, P_k are simultaneously geometrically ergodic and the corresponding sequence of invariant measures, π_k satisfy*

$$\pi_k \Rightarrow \pi$$

where π is the invariant distribution of P.

Proof. The simultaneous geometric ergodicity follows directly from Theorem 7. This implies (cf. Athreya and Ney, 1978; Asmussen, 1987) that, choosing an arbitrary $x_0 \in \mathcal{X}$, if X^k is a process with transition kernel $P_k(x,\cdot)$ and with $X_0^k = x_0$, then we can construct a single sequence of (aperiodic, finite-mean) regeneration times which are common to all of the processes X^k simultaneously, for all sufficiently large $k \in \mathbf{N} \cup \infty$. It follows that the total variation distances $\|P_k^n(x_0,\cdot) - \pi_k(\cdot)\|$ can be bounded independently of k. That is, for some positive integer K, for any $\epsilon > 0$ there exists a positive integer n_0 such that

$$\|P_k^{n_0}(x_0,\cdot) - \pi_k(\cdot)\|_{\text{var}} \le \epsilon$$

for all $k \ge K$ or $k = \infty$.

On the other hand, it is easily checked directly that the strong Feller continuity of P implies strong Feller continuity of P_k for each k. Hence, each P_k is weak Feller continuous. Set

$$r_k(x) \;=\; |P_k f(x) - P f(x)| \;=\; \left| \int (f(h_k(y)) - f(y))\, P(x,dy) \right| .$$

By uniform continuity of f, we have $\lim_{k\to\infty} \sup_{x\in\mathcal{X}} r_k(x) = 0$; hence, the uniform convergence property (5) of Lemma 10 is satisfied (without even requiring that A be compact). We may thus apply Lemma 10. Hence, for sufficiently large k we will have $d(P_k^{n_0}(x_0,\cdot), P_\infty^{n_0}(x_0,\cdot)) \le \epsilon$.

Therefore, by the triangle inequality and by (3), for all sufficiently large k, the quantity $\|\pi_k - \pi_\infty\|_{\text{var}}$ is bounded above by

$$\|\pi_k(\cdot) - P_k^{n_0}(x_0,\cdot)\|_{\text{var}} + d\left(P_k^{n_0}(x_0,\cdot), P_\infty^{n_0}(x_0,\cdot)\right) + \|P_\infty^{n_0}(x_0,\cdot) - \pi_\infty(\cdot)\|_{\text{var}} \;\le\; 3\epsilon .$$

The result follows. ∎

Remark. It is interesting that Theorems 10 and 12 have analogues for non-geometrically ergodic chains. Specifically, we require that the perturbations P_h all simultaneously satisfy the drift condition

$$P_h V \leq V - \delta + b\mathbf{1}_C \tag{6}$$

for suitable drift function V and small set C . We omit the details here except to remark that for the perturbations to preserve the drift condition, we will require a uniformly Lipshitz condition on the drift function V.

REFERENCES

S. Asmussen (1987), Applied Probability and Queues. John Wiley & Sons, New York.

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.

S.N. Ethier and T.G. Kurtz (1986), Markov processes, characterization and convergence. Wiley, New York.

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

S.D. Jacka and G.O. Roberts (1995), Strong forms of weak convergence. Preprint.

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.

R.M. Neal (1993), Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

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

G.O. Roberts and R.L. Tweedie (1994), Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Research Rep. **94-9**, Statistical Laboratory, University of Cambridge.

J.S. Rosenthal (1994), Convergence of Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput., to appear.

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

A. Sinclair (1992), Improved bounds for mixing rates of Markov chains and multicommodity flow. Combinatorics, Prob., Comput. **1**, 351–370.

A.F.M. Smith and G.O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). J. Roy. Stat. Soc. Ser. B **55**, 3-24.

A.D. Sokal (1989), Monte Carlo methods in statistical mechanics: foundations and new algorithms. Dept. of Physics, New York University. Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne, Switzerland.