A note on geometric ergodicity and floating-point roundoff error

by

Laird Breyer*, Gareth O. Roberts*, and Jeffrey S. Rosenthal**

(August 8, 2000; revised and shortened, November 15, 2000.)

Abstract. We consider the extent to which Markov chain convergence properties are affected by the presence of computer floating-point roundoff error. This paper extends previous work of Roberts, Rosenthal, and Schwartz (1998) to the case of proportional errors.

1. Introduction.

Geometric ergodicity is an important concept in convergence of Markov chains to their stationary distributions. For example, this property is used to justify the applicability of the central limit theorem to ergodic averages along the path of the chain. When run on an actual computer, Markov chains are subject to floating-point roundoff errors. This paper considers the extent to which geometric (and other) ergodicity is affected by small roundoff errors.

A Markov chain on a state space $\mathcal{X}$, with transition probabilities $P(x, \cdot)$ and stationary distribution $\pi(\cdot)$, is said to be *geometrically ergodic* if there is $\rho < 1$ and $M : \mathcal{X} \to [0, \infty)$ such that

$$\|P^n(x, \cdot) - \pi(\cdot)\| \ \le\ M(x)\, \rho^n\,,$$

where

$$\|P^n(x, \cdot) - \pi(\cdot)\| \ \equiv\ \sup_{A \subseteq \mathcal{X}} |P^n(x, A) - \pi(A)|$$

* Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, England. Internet: `g.o.roberts@lancaster.ac.uk` and `l.breyer@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.

is the total variation distance between the law of the Markov chain after n steps (when started at the point $x \in \mathcal{X}$), and the stationary distribution $\pi(\cdot)$.

This is known (cf. Roberts and Rosenthal, 1997) to be equivalent to the existence of $\lambda < 1$, $b < \infty$, and a small set $C \subseteq \mathcal{X}$ such that

$$PV(x) \;\leq\; \lambda V(x) + b\,\mathbf{1}_C(x)\,, \qquad x \in \mathcal{X}\,, \tag{1}$$

where $PV(x) = \int V(y)\,P(x,dy)$. (Recall that a set is *small* for a Markov chain if there exists a positive integer n_0, a positive constant ϵ, and a probability measure ν on $\mathcal{X}$, such that $P^{n_0}(x,\cdot) \geq \epsilon\nu(A)$ for all $x \in C$ and $A \subseteq \mathcal{X}$.)

Roberts, Rosenthal, and Schwartz (1998) considered issues related to running such a Markov chain on a computer, and in particular the effect of various roundoff errors during the simulation. They introduced a summary roundoff function $h : \mathcal{X} \to \mathcal{X}$, with $h(x)$ "close" to x for all x. This leads to a modified Markov chain $\widetilde{P}$ given by

$$\widetilde{P}(x,A) \;=\; P(x,h^{-1}(A))\,. \tag{2}$$

In this framework, the assumption of small computer errors can be taken as

$$\|h(x) - x\| \leq \delta\,, \qquad x \in \mathcal{X}\,, \tag{3}$$

where $\|x\|$ is the norm of $x \in \mathcal{X}$. (We assume throughout that $\mathcal{X}$ is a normed vector space over $\mathbf{R}$, e.g. $\mathcal{X} \subseteq \mathbf{R}^d$.) It is shown by Roberts et al. (1998) that, if P is geometrically ergodic with drift function V such that $\log V$ is uniformly continuous, and δ is sufficiently small, then $\widetilde{P}$ will also be geometrically ergodic. That is, geometric ergodicity is preserved under small perturbations in that case.

A common case not included in the above arises when instead we have merely

$$\|h(x) - x\| \leq \delta\|x\|\,, \qquad x \in \mathcal{X}\,, \tag{4}$$

as may occur with floating-point computations (cf. Section 2 below). That is, the roundoff errors may have magnitude proportional to the magnitude of x, rather than being uniformly bounded. In this note, we shall show that this case is amenable to a technique similar to that of Roberts et al. (1998).

We note that the results we present are only the beginning of a rigorous analysis of how computer engineering realities affect the dynamics of mathematically specified Markov chains. We hope to pursue such questions more comprehensively in the future.

Remark. Even if a modified Markov chain $\widetilde{P}$ is proven to converge as quickly as the original chain, there is still the question of what the new target distribution is. Roberts et al. (1998) investigate this issue in total variation distance and in the weak topology. The methods in our paper could also be extended to consider this issue, but we do not pursue that here.

2. Floating point representations in computers.

IEEE standard 754 (IEEE, 1985) is a specification commonly adhered to for the representation of floating point numbers in computers, using a fixed number B of bits (e.g. $B = 32$ with single precision numbers or $B = 64$ with double precision numbers). Mathematically, numbers x are encoded using $B = M + N + 1$ bits to a finite precision, in the following way (called normalized floating point representation):

$$x := \sigma \cdot (1 + k/2^N) \cdot 2^e,$$

where $\sigma = \pm 1$ is the sign (1 bit), $k \in \{0, \ldots, 2^N - 1\}$ is the fractional part (N bits), and $e \in \{-2^{M-1} + 2, \ldots, 2^{M-1} - 1\}$ is the exponent (M bits). Single precision numbers use $M = 8$ and $N = 23$, giving an effective range (excluding the sign) of $2^{-126} \approx 10^{-44.85}$ to $(2 - 2^{-23}) \cdot 2^{127} \approx 10^{38.53}$, while double precision is represented by $M = 11$, $N = 52$, with an effective range (excluding sign) of $2^{-1022} \approx 10^{-323.3}$ to $(2 - 2^{-52}) \cdot 2^{1023} \approx 10^{308.3}$. Numbers larger than this are represented by the special symbol +**Infinity**, which is often encoded as $\sigma = 0$, $k = 0$, $e = 2^{M-1}$. Moreover, the number zero is nonunique; more precisely there exist two distinct values $+0$ and -0 which are only equal when compared directly.

Clearly, not all real numbers x can be represented with a fixed number B of bits in this way. Indeed, given a real number x, setting $e = \lfloor \log_2 |x| \rfloor$ and $\sigma = \mathrm{sign}(x)$, the closest the computer can come to approximating x is as

$$h(x) = \sigma \lfloor \frac{1}{2} + |x| 2^{N-1-e} \rfloor 2^{-(N-1-e)} .$$

It follows that

$$|h(x) - x| \leq 2^{-(N-1-e)} \leq 2^{-(N-1)} |x| .$$

We see from the above that the error $|h(x) - x|$ is proportional to $|x|$, thus violating (3). (Strictly speaking, (3) holds for a sufficiently large δ since $|x|$ is bounded by the finite range of the computer. However, in the present paper, we ignore issues related to finite range, and concentrate solely on issues related to finite precision, i.e. to roundoff errors. It is in that sense that (3) is violated.) However, the assumption (4) does hold here with $\delta = 2^{-(N-1)}$.

With regard to Markov chain algorithms and their implementations on computer systems, we shall therefore assume that the final error for each update behaves as (4). Of course, this is meant as a convenient summary of the cumulative effect of various complicated roundoff errors introduced at each stage of the update calculation. (For example, a side effect of using floating point representations is that the corresponding arithmetic becomes inexact and non-commutative, e.g. perhaps $(x \cdot y)/y \neq x$ or $x + y \neq y + x$.)

3. Geometric ergodicity under perturbations satisfying (4).

Suppose a Markov chain P is geometrically ergodic, thus satisfying (1) for some function V and small set C. Suppose further that $\widetilde{P}$ is obtained via (2), for some roundoff function h satisfying (4) for some $\delta > 0$. Assume also that V satisfies

$$V(y+u) - V(y) \leq \delta K V(y), \qquad \|u\| \leq \delta \|y\|, \quad y \in \mathcal{X}\,, \tag{5}$$

for some $K < \infty$. (Of course, we could subsume the product of δ and K into a single constant, but our notation better emphasizes the dependence upon δ.) For example, condition (5) holds if $\mathcal{X} = \mathbf{R}$ and $V(x) = C_1|x|^n + C_2$. with $C_1, C_2 \geq 0$.

Note that the condition (5) is implied, if $\mathcal{X}$ is finite-dimensional and $V(x)$ is continuously differentiable, by

$$\|\nabla \log V(y)\| \;\leq\; K' \,/\, \|y\|\,, \tag{6}$$

where $K' = \delta^{-1} \log(1 + K\delta) \approx K$.

Proposition 1. If (1) and (5) hold, and if $\widetilde{P}$ is derived from P via (2), where h satisfies (4), then

$$\widetilde{P}V(x) \;\leq\; (1 + \delta K)(\lambda V(x) + b\,\mathbf{1}_C(x))\,.$$

Proof. We have that

$$\begin{aligned}
\widetilde{P}V(x) &= PV(x) + (\widetilde{P} - P)V(x) \\
&= PV(x) + \int (V(h(y)) - V(y))\, P(x, dy) \\
&\leq \lambda V(x) + b\mathbf{1}_C(x) + \int \delta\, K\, V(y)\, P(x, dy) \\
&\leq \lambda V(x) + b\,\mathbf{1}_C(x) + \delta K(\lambda V(x) + b\,\mathbf{1}_C(x))\,,
\end{aligned}$$

which gives the result. ■

This ensures geometric ergodicity provided that $(1+\delta K)\lambda < 1$, or equivalently

$$\delta < K^{-1}(\lambda^{-1} - 1). \tag{7}$$

We thus obtain

Theorem 2. Suppose a Markov chain P is geometrically ergodic, satisfying (1) for some V and C. Assume that V satisfies (5) for some $K < \infty$. Suppose further that $\widetilde{P}$ is obtained via (2), for some perturbation function h satisfying (4) and (7). Then $\widetilde{P}$ is also geometrically ergodic.

This theorem thus proves that geometric ergodicity is preserved, under sufficiently small floating-point-type perturbations, provided that the drift function V satisfies the smoothness condition (5) (or (6)).

Remark. If $\{X_t\}$ satisfies (1) for some drift function V and small set C, and $\varphi : \mathcal{X} \to \mathcal{X}'$ is a bi-measurable bijection, we can define a Markov chain $\{X_t^\varphi\}$ on $\mathcal{X}'$ by $X_t^\varphi = \varphi(X_t)$, with corresponding transition kernel P^φ. Then $\varphi(C)$ is a small set for $\{X_t^\varphi\}$, and furthermore

$$P^\varphi V^\varphi(x') \ \leq \ \lambda V^\varphi(x') + b\,\mathbf{1}_{\varphi(C)}(x')\,, \qquad x' \in \mathcal{X}'\,, \tag{8}$$

where $V^\varphi(x') = V(\varphi^{-1}(x'))$. Hence, $\{X_t^\varphi\}$ is also geometrically ergodic with the same constants, and indeed $\|(P^\varphi)^n(\varphi(x),\cdot) - \pi^\varphi\| = \|P^n(x,\cdot) - \pi\|$ for all n, where $\pi^\varphi(dx') = \pi(\varphi^{-1}(dx'))$ is stationary for $\{X_t^\varphi\}$. Furthermore, if $\widetilde{X}$ is a perturbation of X with approximation function h, then $\varphi(\widetilde{X})$ is a perturbation of X^φ with approximation function $h^\varphi \equiv \varphi h \varphi^{-1}$. In the special case that $\mathcal{X} = \mathbf{R}$, $\mathcal{X}' = \mathbf{R}^+$, and $\varphi(x) = \exp(x)$, then h satisfies (3) if and only if $h^{\exp}$ satisfies (4). In that case, if $P^{\exp}$ is geometrically ergodic, then by Proposition 6 of Roberts et al. (1998), $P^{\exp}$ is robust to perturbations of the kind satisfying (4) provided that $\log V(\cdot) = \log V^{\exp}(\exp(\cdot))$ is uniformly continuous.

Example. Consider the Gaussian Random Walk Metropolis algorithm of the form

$$X_{t+1} = \begin{cases} Y_t\,, & \pi(Y_t) \,/\, \pi(X_t) > \xi_t, \\ X_t\,, & \text{otherwise}\,, \end{cases}$$

where $Y_t \sim N(X_t, 1)$, and where the ξ_t are independently chosen as i.i.d. Uniform$[0,1]$. This chain (assuming no roundoff errors) is geometrically ergodic with drift function $V(x) = \exp(|x|)$ (see e.g. Mengersen and Tweedie, 1996). Furthermore, it follows from Roberts et al. (1998) that, if this chain is perturbed by a roundoff function h satisfying (3), geometric ergodicity is preserved. However, this Markov chain is *not* robust to perturbations of type (4). For example, let

$$h(x) = \text{sign}(x) \cdot 2^{\lfloor \log_2 |x| \rfloor} \cdot (1 + 2^{-52} \lfloor 2^{52} (|x| 2^{-\lfloor \log_2 |x| \rfloor} - 1) \rfloor)$$

(which is an idealisation of the IEEE discretisation described in Section 2, without any truncation). Then for $z_i \in \text{Image}(h)$ with $|z_i| \to \infty$, we have $\lim_{i\to\infty} \inf_{y \in (h^{-1}(z_i))^C} |y - z_i| = \infty$, which implies that $\lim_{i\to\infty} \mathbf{P}(z_i, \{z_i\}^C) = 0$. Therefore the algorithm is not geometrically ergodic (see Roberts and Tweedie, 1996). On the other hand, the Markov chain $X_t^{\exp} \equiv \exp(X_t)$ is also geometrically ergodic, with $V^{\exp}(x) = V(\exp^{-1}(x)) = V(\log x) = \exp(|\log x|) = \max\{x, x^{-1}\}$. Furthermore, by the previous remark, the geometric ergodicity of $\{X_t^{\exp}\}$ is robust to perturbations satisfying (4). Furthermore, the target distribution π can be "recovered" from $\{X_t^{\exp}\}$ by inversion: $\log X_t^{\exp}$ converges in distribution to the desired stationary distribution π.

Remark. Recall that a Markov chain with transition kernel P converges at the polynomial rate α if

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq C(x) n^{-(\alpha/1-\alpha)}, \qquad n \in \mathbf{N},$$

for some $0 < \alpha < 1$. This is implied (Roberts and Jarner, 2000) by the existence of a function $V \geq 1$, a small set C, and constant $a > 0$ such that

$$PV \leq V - aV^{\alpha} + b\,\mathbf{1}_C \tag{9}$$

Suppose that P is polynomial ergodic with polynomial rate α, and satisfies (9) for some small set C and some drift function V which satisfies

$$V(y+u) - V(y) \leq \delta K (V(y))^{\gamma}, \qquad \|u\| \leq \delta \|y\|^{\epsilon}, \qquad y \in \mathcal{X},$$

for some constants $\gamma \leq \min(1, \alpha)$ and $\epsilon > 0$. Define $\widetilde{P}$ by (2), and assume that $\|h(x) - x\| \leq c\|x\|^{\beta}$ for some $\beta \leq \epsilon$ and $c \leq \delta$. Then it is straightforward to show that

$$\widetilde{P}V \leq V - aV^{\alpha} + b'\,\mathbf{1}_C + cKV^{\alpha},$$

for some $b' < \infty$. In particular, if $c < a/K$, then the chain defined by $\widetilde{P}$ is also polynomially ergodic, with the same polynomial rate α.

REFERENCES

IEEE (1985), ANSI/IEEE Standard 754-1985. In *IEEE Standard for Binary Floating-Point Arithmetic*, IEEE, New York.

K.L. Mengersen and Tweedie, R.L. (1996), Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics* **24**, 101–12.

G.O. Roberts (1999), A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms. *Journal of Applied Probability* **36**, 1210–1217.

G.O. Roberts and J.S. Rosenthal (1997), Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2**, Paper no. 2, 13–25.

G.O. Roberts, J.S. Rosenthal, and P.O. Schwartz (1998), Convergence properties of perturbed Markov chains. *Journal of Applied Probability* **35**, 1–11.

G.O. Roberts and R.L. Tweedie (1996), Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, no. 1, 95–110.