Quantitative Non-Geometric Convergence Bounds for Independence Samplers

by

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

(September 2008; revised July 2009.)

1. Introduction.

Markov chain Monte Carlo (MCMC) algorithms are widely used in statistics, physics, and computer science, to sample from complicated high-dimensional probability distributions. A central question is how quickly the chain converges to the target (stationarity) distribution. In this paper, we consider this question for a particular class of MCMC algorithms, independence samplers (Hastings, 1970; Tierney, 1994).

It is well known that independence samplers are geometrically ergodic if and only if $\operatorname{ess\,sup}_{x \in \mathcal{X}} w(x) \equiv w^* < \infty$ (where the weight function $w(x)$ is defined below), in which case precise quantitative convergence bounds are available (Liu, 1996; Smith and Tierney, 1996). However, if $w^* = \infty$, then the chain is not geometrically ergodic, and no quantitative bounds were previously known. In this paper, we use the coupling method to develop general quantitative upper bounds (Theorem 6, Corollary 7) applicable even when $w^* = \infty$. Together with a corresponding lower bound (Theorem 8), they provide fairly precise information about the time to stationarity of these algorithms.

We apply our results to three examples. In one fast-converging case, we prove that the convergence time (defined as getting within 0.01 of stationarity in total variation distance) is between 24 and 50 iterations. In another, much slower-converging case, we prove that the convergence time is between 4,000,000 and 14,000,000 iterations. In still another case, we prove that the convergence time is between 5×10^{32} and 10^{34} iterations. This shows

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

**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.

the precision and flexibility of our methods. It also illustrates the variation in convergence times of MCMC algorithms, and the importance of precise analysis to understand their convergence.

Our paper extends previous work about quantitative convergence bounds on the distance to stationarity of Markov chains after n steps. For geometrically ergodic chains, this is a well studied area, see e.g. Rosenthal (1995, 2002), Roberts and Tweedie (1997), Jones and Hobert (2001, 2004), Marchev and Hobert (2004), Douc et al. (2004), and Baxendale (2005). For non-geometrically ergodic chains, convergence bounds have been studied by Meyn and Tweedie (1993), Fort and Moulines (2000, 2003), Jarner and Roberts (2002), and especially by Douc et al. (2007), who use hitting times of small sets to provide very general and useful quantitative bounds which are then applied to certain specific examples (including an independence sampler on the unit interval). Compared to the work of Douc et al. (2007), our results are less general but are better suited to the specific properties of independence samplers (as illustrated by our closely matching upper and lower bounds in the examples).

2. Preliminaries.

Let $(\mathcal{X}, \mathcal{F}, \nu)$ be a non-atomic measure space (usually a subset of $\mathbf{R}^d$ with Lebesgue measure), and let π and q be two different positive probability densities on $\mathcal{X}$ with respect to $\nu(\cdot)$. Let $\Pi(A) = \int_A \pi(x)\,\nu(dx)$ and $Q(A) = \int_A q(x)\,\nu(dx)$ be the corresponding probability measures.

The *independence sampler* Markov chain is defined as follows. Given X_{n-1}, the algorithm proposes a state $Y_n \sim Q(\cdot)$, and then accepts it with probability $\alpha(X_{n-1}, Y_n) \equiv \min(1, \frac{\pi(Y_n)\,q(X_{n-1})}{\pi(X_{n-1})\,q(Y_n)}) = \min(1, \frac{w(Y_n)}{w(X_{n-1})})$, where $w(x) = \pi(x)/q(x)$ is the weight function, otherwise it rejects it. That is, the algorithm chooses $U_n \sim \text{Uniform}[0,1]$, and then sets

$$X_n = \begin{cases} Y_n\,, & U_n \le \frac{w(Y_n)}{w(X_{n-1})} \\ X_{n-1}\,, & \text{otherwise}\,. \end{cases} \tag{1}$$

If $U_n \le \frac{w(Y_n)}{w(X_{n-1})}$ then we say the proposal Y_n was *accepted*, otherwise we say it was *rejected*. We write $P(x, A) = \mathbf{P}[X_1 \in A \,|\, X_0 = x]$ and $P^n(x, A) = \mathbf{P}[X_n \in A \,|\, X_0 = x]$, and let $\|P^n(x, \cdot) - \Pi(\cdot)\| = \sup_{A \in \mathcal{F}} |P^n(x, A) - \Pi(A)|$ be the total variation distance to stationarity after n steps.

Much is known about the theoretical properties of independence samplers. For example, they are geometrically (in fact, uniformly) ergodic if and only if $\operatorname{ess\,sup}_{x \in \mathcal{X}} w(x) \equiv w^* < \infty$ (Tierney, 1994; Mengersen and Tweedie, 1996; Roberts and Rosenthal, 1998; Rosenthal, 1997), and their complete spectral decomposition is available (Liu, 1996; Smith and Tierney, 1996). In particular, if $w^* < \infty$, then $\alpha(x,y) \geq \min(1,\ w(y)/w^*) = w(y)/w^*$, so that $P(x,dy) \geq [\pi(y)/w^*]\,dy$, a minorisation condition which implies (Liu, 1996; Smith and Tierney, 1996; Rosenthal, 1995, 2002) that

$$\|P^n(x,\cdot) - \Pi(\cdot)\| \leq \Big(1 - \frac{1}{w^*}\Big)^n,$$

a simple and useful quantitative bound. (Roughly speaking, $w^* < \infty$ when the proposal tails are at least as heavy as the target tails.) However, if $w^* = \infty$, then there is no spectral gap and the chain is not geometrically ergodic, and no quantitative bounds were previously known, though we develop some herein.

For numerical concreteness in the examples, we shall say (following Cowles and Rosenthal, 1997; Jones and Hobert, 2001, 2004) that the chain's "convergence time" is the smallest number n of iterations such that $\|P^n(x,\cdot) - \Pi(\cdot)\| < 0.01$. Hence, we shall attempt to find upper and lower bounds, n^* and n_*, such that $\|P^{n^*}(x,\cdot) - \Pi(\cdot)\| < 0.01$ and $\|P^{n_*}(x,\cdot) - \Pi(\cdot)\| > 0.01$. (Since the total variation distance to stationarity is non-increasing, see e.g. Roberts and Rosenthal, 2004, this implies that the convergence time is between n_* and n^*.) If $n^* \,/\, n_*$ is not too large, this will indicate relatively tight bounds on the time to convergence.

3. Coupling Bounds.

We shall proceed using the standard method (see e.g. Roberts and Rosenthal, 2004, and the references therein) of simultaneously constructing two different copies $\{X_n\}$ and $\{X_n'\}$ of the same Markov chain, with different starting values X_0 and X_0', but identical marginal transition probabilities: $\mathbf{P}[X_1 \in A \,|\, X_0 = x] = \mathbf{P}[X_1' \in A \,|\, X_0' = x]$. (We will later let $X_0 = x$ for some specific $x \in \mathcal{X}$, but choose $X_0' \sim \Pi$, so the coupling inequality will give that $\|\mathbf{P}^n(x,\cdot) - \Pi(\cdot)\| \leq \mathbf{P}[X_n \neq X_n'].$)

For the independence sampler, we shall use a very simple joint coupling construction. We let $\{X_n\}$ and $\{X_n'\}$ be two copies of the same independence sampler, with different initial distributions $\mathcal{L}(X_0)$ and $\mathcal{L}(X_0')$, but each updated using the same random variables Y_n and U_n. That is, the first chain $\{X_n\}$ is updated according to (1), while the second chain $\{X_n'\}$ is updated according to

$$X_n' \;=\; \begin{cases} Y_n\,, & U_n \leq \frac{w(Y_n)}{w(X_{n-1}')} \\ X_{n-1}'\,, & \text{otherwise}\,. \end{cases} \tag{2}$$

using the same random variables Y_n and U_n in both (1) and (2).

Lemma 1. *With the above coupling, if $X_m = X_m'$ for some m, then $X_n = X_n'$ for all $n \geq m$.*

Proof. This follows immediately by comparing (1) and (2). ∎

Lemma 2. *With the above coupling, if $w(X_{n-1}) \leq w(X'_{n-1})$ and the n^{th} proposal Y_n is accepted by the second chain $\{X'_n\}$, then it is also accepted by the first chain $\{X_n\}$.*

Proof. $w(X_{n-1}) \leq w(X'_{n-1})$ implies $\frac{w(Y_n)}{w(X_{n-1})} \geq \frac{w(Y_n)}{w(X'_{n-1})}$, so that if $U_n \leq \frac{w(Y_n)}{w(X'_{n-1})}$, then also $U_n \leq \frac{w(Y_n)}{w(X_{n-1})}$. ∎

Lemma 3. *With the above coupling, the independence sampler is monotone with respect to the partial order induced by w. That is, if $w(X_{n-1}) \leq w(X'_{n-1})$, then $w(X_n) \leq w(X'_n)$.*

Proof. If both proposals are accepted, or both are rejected, then the conclusion is trivial. By the previous lemma, the only other possibility is that the first chain accepts the proposal but the second does not, i.e. that $X_n = Y_n$ but $X'_n = X_{n-1}$. But by (1), this can only happen if $w(Y_n)/w(X'_{n-1}) < 1$, i.e. $w(X_n)/w(X'_n) < 1$, i.e. $w(X_n) < w(X'_n)$. ∎

Corollary 4. *With the above coupling, if $w(X_0) \leq w(X'_0)$, and $X'_n \neq X'_0$, then $X_n = X'_n$.*

Proof. Since $w(X_0) \leq w(X'_0)$, therefore by Lemma 3, $w(X_{n-1}) \leq w(X'_{n-1})$ for all n. Since $X'_n \neq X'_0$, at least one proposal by time n must have been accepted by the second chain. By Lemma 2, that proposal must have also been accepted by the first chain, at which point the two chains became equal. Then, by Lemma 1, they remained equal thereafter, so $X_n = X'_n$. ∎

Theorem 5. *If $w(X_0) \leq w(X'_0)$, then $\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| \leq \mathbf{P}[X'_n = X'_0]$, i.e. the distance to stationarity is bounded by the probability that the second chain has not yet moved.*

Proof. By the coupling inequality, $\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| \leq \mathbf{P}[X'_n \neq X_n]$. But by Corollary 4, $\mathbf{P}[X'_n \neq X_0] \leq \mathbf{P}[X_n = X'_n]$, so $\mathbf{P}[X_n \neq X'_n] \leq \mathbf{P}[X'_n = X_0]$. The result follows. ∎

Remark. If $X_0 = x_0$ and $X_0' = x_0'$ are constants, with $w(x_0) < w(x_0')$, then the conclusion of Theorem 5 becomes an equality.

To make the conclusion of Theorem 5 more concrete, let $m(x) = \mathbf{P}[X_1 \neq X_0 \,|\, X_0 = x] = \mathbf{E}[\alpha(x, Y)]$ (where $Y \sim Q(\cdot)$) be the probability that the independence sampler will accept its first move when started at x. Then we have:

Theorem 6. $\|P^n(x,\cdot) - \Pi(\cdot)\| \leq \mathbf{E}[(1 - \min(m(x), m(Z)))^n]$, *where the expectation is taken with respect to* $Z \sim \Pi(\cdot)$.

Proof. We start by choosing $X_0 = x$ and $X_0' \sim \Pi$, so $\mathcal{L}(X_n) = P^n(x,\cdot)$ and $\mathcal{L}(X_n') = \Pi(\cdot)$ for all n. It follows directly from the coupling inequality that $\|\mathbf{P}^n(x,\cdot) - \Pi(\cdot)\| \leq \mathbf{P}[X_n \neq X_n']$. We then condition on the value of X_0', and break up the expectation over $X_0' \equiv Z \sim \Pi$ into the two parts $w(x) \leq w(X_0')$ and $w(x) > w(X_0')$. The first part is bounded using Theorem 5 and the observation that $\mathbf{P}[X_n' = X_0' \,|\, X_0'] = (\mathbf{P}[X_1' = X_0' \,|\, X_0'])^n$, and the second part is bounded similarly upon reversing the roles of $\{X_n\}$ and $\{X_n'\}$. This leads to the string of inequalities

$$
\begin{aligned}
\|P^n(x,\cdot) - \Pi(\cdot)\| \;&\leq\; \mathbf{P}[X_n \neq X_n'] \;=\; \mathbf{E}\Big[\mathbf{P}[X_n \neq X_n' \,|\, X_0']\Big] \\
&= \mathbf{E}\Big[\mathbf{P}[X_n \neq X_n' \,|\, X_0']\,\mathbf{1}_{w(x) \leq w(X_0')} + \mathbf{P}[X_n \neq X_n' \,|\, X_0']\,\mathbf{1}_{w(x) > w(X_0')}\Big] \\
&= \mathbf{E}\Big[\mathbf{P}[X_n' = X_0' \,|\, X_0']\,\mathbf{1}_{w(x) \leq w(X_0')} + \mathbf{P}[X_n = X_0 \,|\, X_0']\,\mathbf{1}_{w(x) > w(X_0')}\Big] \\
&= \mathbf{E}\Big[(\mathbf{P}[X_1' = X_0' \,|\, X_0'])^n\,\mathbf{1}_{w(x) \leq w(X_0')} + (\mathbf{P}[X_1 = X_0 \,|\, X_0'])^n\,\mathbf{1}_{w(x) > w(X_0')}\Big] \\
&= \mathbf{E}\Big[(1 - m(X_0'))^n\,\mathbf{1}_{w(x) \leq w(X_0')} + (1 - m(x))^n\,\mathbf{1}_{w(x) > w(X_0')}\Big] \\
&= \mathbf{E}\Big[(1 - m(Z))^n\,\mathbf{1}_{w(x) \leq w(Z)} + (1 - m(x))^n\,\mathbf{1}_{w(x) > w(Z)}\Big].
\end{aligned}
$$

Finally, it follows as in Lemma 2 that if $w(x) \leq w(Z)$ then $m(x) \geq m(Z)$ so $1 - m(Z) = 1 - \min(m(x), m(Z))$, and similarly if $w(x) > w(Z)$ then $1 - m(x) = 1 - \min(m(x), m(Z))$. We conclude that

$$
\mathbf{E}\Big[(1 - m(Z))^n\,\mathbf{1}_{w(x) \leq w(Z)} + (1 - m(x))^n\,\mathbf{1}_{w(x) > w(Z)}\Big] \;=\; \mathbf{E}\Big[(1 - \min(m(x), m(Z)))^n\Big],
$$

thus giving the result. ∎

In particular, if $w(x) = \inf_{y \in \mathcal{X}} w(y)$, then with probability 1, $w(x) \leq w(Z)$ and hence $1 - \min(m(x), m(Z)) = 1 - m(Z)$, and we conclude:

Corollary 7. *If $w(x) = \inf_{y \in \mathcal{X}} w(y)$, then $\|P^n(x, \cdot) - \Pi(\cdot)\| \le \mathbf{E}[(1 - m(Z))^n]$, where the expectation is again taken with respect to $Z \sim \Pi(\cdot)$.*

As for lower bounds, we have the following:

Theorem 8. *For any $x, z \in \mathcal{X}$, $\|P^n(x, \cdot) - \Pi(\cdot)\| \ge p_z - (1 - (1 - q_z)^n)$, where $p_z = \Pi\{y \in \mathcal{X} : w(y) > w(z)\}$ and $q_z = Q\{y \in \mathcal{X} : w(y) > w(z)\}$.*

Proof. Let $A = \{y \in \mathcal{X} : w(y) > w(z)\} \setminus \{x\}$. Then $\Pi(A) = p_z$. On the other hand, $P^n(x, A) \le \mathbf{P}(\exists\, m \le n : X_m \in A \,|\, X_0 = x) \le \mathbf{P}(\exists\, m \le n : Y_m \in A) = 1 - \mathbf{P}(Y_m \notin A\; \forall\, m \le n) = 1 - (1 - q_z)^n$. So, $\|P^n(x, \cdot) - \Pi(\cdot)\| \ge \Pi(A) - P^n(x, A) \ge p_z - (1 - (1 - q_z)^n)$. ■

Remark. We shall apply our results to several independence sampler examples below. However, it should be admitted that the results may be less applicable in genuine MCMC examples where little is known about the stationarity distribution $\Pi(\cdot)$. Indeed, to apply Corollary 7, it is necessary to know $x = \mathrm{arginf}_y w(y)$ which is often impossible. Theorem 6 does not require this, but it is still stated in terms of an expectation with respect to $\Pi(\cdot)$ which may present significant obstacles in complicated examples (though it is sometimes possible to bound such expectations using drift conditions and other techniques, see e.g. Meyn and Tweedie, 1993, Theorem 14.3.7; alternatively those integrals could themselves be estimated using auxiliary Monte Carlo simulations though at the expense of complete rigor in the resulting bounds). Even Theorem 8 requires computing or bounding certain probabilities with respect to $\Pi(\cdot)$, which may itself be challenging in some cases. In summary, while our results are applicable to certain independence samplers as we shall now see, we do not expect them to provide useful convergence bounds in all such situations.

4. Example #1: Exponential Distributions.

Let $\mathcal{X} = [0, \infty)$ with Lebesgue measure. Let $\pi(x) = e^{-x}$ be the density of a standard exponential distribution, and $q(x) = k\, e^{-k\, x}$ be that of the Exponential(k) distribution, for some fixed $k > 0$. This example was considered e.g. by Smith and Tierney (1996) and Roberts and Rosenthal (1998) and Jones and Hobert (2001); for an interactive display see Rosenthal (1997).

Here $w(x) = \pi(x)/q(x) = e^{(k-1)x}/k$. If $k \le 1$, then $w^* \equiv \sup_x w(x) = 1/k < \infty$, and the chain is geometrically ergodic, with $\|P^n(x, \cdot) - \Pi(\cdot)\| \le (1 - \frac{1}{w^*})^n = (1 - k)^n$. (In particular, the case $k = 1$ corresponds to immediate convergence, i.e. to i.i.d. sampling.) For example, if $k = 0.01$, then $\|P^n(x, \cdot) - \Pi(\cdot)\| \leq (0.99)^n$, which is less than 0.01 if $n = 459$, so the chain converges within 459 iterations.

If $k > 1$, then $w^* = \infty$, and the chain is not geometrically ergodic. In this case, the chain is known from simulations (Roberts and Rosenthal, 1998) to converge very poorly. To bound this, we compute (where $Y \sim Q(\cdot)$) that

$$
\begin{aligned}
m(x) \;&=\; \mathbf{E}[\alpha(x, Y)] \;=\; \int_0^\infty \alpha(x, y)\; k e^{-ky}\, dy \\
&= \int_0^\infty \min(1,\; e^{(k-1)(y-x)})\; k e^{-ky}\, dy \\
&= \int_0^x e^{(k-1)(y-x)}\; k e^{-ky}\, dy \;+\; \int_x^\infty k e^{-ky}\, dy \\
&= k e^{-(k-1)x}(1 - e^{-x}) \;+\; e^{-kx} \;=\; k e^{-(k-1)x} - (k-1)e^{-kx}\,.
\end{aligned}
$$

To proceed, for simplicity consider starting at the state 0, since $w(0) = \inf_{y \in \mathcal{X}} w(y)$. Then from Corollary 7, with $Z \sim \Pi(\cdot)$,

$$
\begin{aligned}
\|P^n(0, \cdot) - \Pi(\cdot)\| \;&\leq\; \mathbf{E}[(1 - m(Z))^n] \\
&= \int_0^\infty \left(1 - k e^{-(k-1)x} + (k-1)e^{-kx}\right)^n e^{-x}\, dx\,. \qquad (3)
\end{aligned}
$$

This gives a precise upper bound on the distance to stationarity after n steps. As expected, it does *not* decrease geometrically with n. The following table gives numerical upper bounds when $k = 5$, for various values of n:

Upper bounds on $\|P^n(0, \cdot) - \Pi(\cdot)\|$ from (3), with $k = 5$:

n	upper bound
10	0.3706
100	0.2008
1,000	0.1105
10,000	0.06145
100,000	0.03434
1,000,000	0.01925
14,000,000	0.009931

In particular, this shows that $\|P^n(x, \cdot) - \Pi(\cdot)\| < 0.01$ when $n = 14,000,000$. That is, the chain converges within 14 million iterations. In particular, we can take $n^* = 14,000,000$.

Now, 14,000,000 is just an *upper bound*, and it is tempting to believe that it is hugely conservative, with the actual convergence time being many orders of magnitude smaller.

However, that is not the case. To see this, consider lower bounds from Theorem 8. Let $z = 4$. Then $p_z = \Pi(4, \infty) = e^{-z} = e^{-4} \doteq 0.0183$, and $q_z = Q(4, \infty) = e^{-kz} = e^{-20} \doteq 2.06 \times 10^{-9}$. Hence, by Theorem 8, for any $x \in \mathcal{X}$,

$$\|P^n(x, \cdot) - \Pi(\cdot)\| \ \geq \ p_z - (1 - (1 - q_z)^n) \ = \ e^{-4} - (1 - (1 - e^{-20})^n) \,. \tag{4}$$

If $n = 4,000,000$, then this lower bound equals $0.01010492 > 0.01$. So, we conclude that, starting from any $x \in \mathcal{X}$, this chain has still *not* converged after four million iterations, and we can take $n_* = 4,000,000$. That is, it really does take millions of iterations for this (rather simple) Markov chain to converge. The ratio of upper to lower bound for this example is

$$n^* / n_* \ = \ 14,000,000 \,/\, 4,000,000 \ = \ 3.5 \,,$$

a fairly small number, indicating fairly tight upper and lower bounds on convergence.

In addition to numerical bounds, we can also consider the functional form by which the distance to stationarity decreases as a function of n. While we know that the decrease cannot be geometric in this case, it can still correspond to *polynomial ergodicity* (Fort and Moulines, 2000, 2003; Jarner and Roberts, 2002). Looking at the above numerical table, it appears that the upper bounds are decreasing at approximately the rate $O(n^{-1/4})$.

For another approach to polynomial rates, combining Theorems 3.6 and 5.3 of Jarner and Roberts (2002) yields the following result (note that they write $q(x)$ for our $1/w(x)$):

Proposition 9. *For an independence sampler, for any* $x \in \mathcal{X}$*,*

$$\lim_{n \to \infty} (n+1)^{\beta - 1} \, \|P^n(x, \cdot) - \Pi(\cdot)\| \ = \ 0 \,,$$

for any $1 \leq \beta \leq 1/(1 - \alpha)$*, with* $\alpha = 1 - \frac{r}{s}$*, and* $r < s < r + 1$*, and* $\Pi(A_\epsilon) = O(\epsilon^{1/r})$ *as* $\epsilon \to 0^+$*, where* $A_\epsilon = \{x \in \mathcal{X} : (1/w(x)) \leq \epsilon\}$*.*

Now, in the above example, $w(x) = e^{(k-1)x}$, whence

$$\{x \in \mathcal{X} : (1/w(x)) \leq \epsilon\} \ = \ [-\frac{\log(\epsilon)}{k-1}, \infty) \,,$$

so for $0 < \epsilon < 1$,

$$\Pi(A_\epsilon) \ = \ \int_{-\frac{\log(\epsilon)}{k-1}}^{\infty} e^{-y} \, dy \ = \ e^{-(-\log(\epsilon)/(k-1))} \ = \ \epsilon^{1/(k-1)} \,.$$

Hence, we can take $r = k - 1$, whence α has upper bound $1 - \frac{r}{r+1} = 1/k$. Then β has upper bound $1/(1 - (1/k)) = k/(k-1)$, and the polynomial rate approaches $O(n^{-(k/(k-1)-1)}) = O(n^{-1/(k-1)})$. In the case $k = 5$, this gives a polynomial rate approaching $O(n^{-1/4})$, which exactly agrees with the numerically imputed rate $O(n^{-1/4})$ from the bound (3). This suggests that the polynomial rate from Proposition 9 is indeed sharp in this case.

Remark. Another approach to assessing the polynomial decay rate of the integral (3) is to make the transformation $z = e^{-x}$, to get that as $n \to \infty$,

$$(3) \;=\; \int_0^1 \left(1 - kz^{k-1} + (k-1)z^k\right)^n dz \;=\; \int_0^1 e^{-nkz^{k-1}} \, (1 + o(1)) \, dz \,.$$

If we now make the transformation $u = n^{1/(k-1)}z$, then this becomes

$$n^{-1/(k-1)} \int_0^{n^{1/(k-1)}} e^{-ku^{k-1}} \, (1 + o(1)) \, du \;\sim\; n^{-1/(k-1)} \int_0^\infty e^{-ku^{k-1}} \, du \,.$$

Since $\int_0^\infty e^{-ku^{k-1}} \, du < \infty$ for $k > 1$, this again suggests that the upper bound (3) is decreasing at the rate $O(n^{-1/(k-1)})$, thus again confirming the same polynomial rate as before.

Remark. It may also be possible to derive polynomial rates from the formula (3) by means of *Tauberian theorems*, see e.g. Bingham et al. (1987), but we do not pursue that here.

Remark. If we instead take $k = 2$ in this example, i.e. let $q(x) = 2\,e^{-2\,x}$, then it follows from (3) that the chain converges (to within 0.01) after 50 iterations, i.e. we can take $n^* = 50$. Furthermore, we would now have $q_z = e^{-8}$ in (4), so $\|P^n(x,\cdot) - \Pi(\cdot)\| \geq e^{-4} - (1 - (1 - e^{-8})^n)$, which is > 0.01 if $n = 24$. So, we conclude that the chain converges in between $n_* \equiv 24$ and $n^* \equiv 50$ iterations. More interestingly, the upper bound values appear numerically to decrease at a rate of $O(n^{-1})$, and this is equal to the rate $O(n^{-1/(k-1)})$ from Proposition 9, again showing agreement between the two rates. Furthermore, the rate $O(n^{-1})$ is the cut-off for establishing a Markov chain $\sqrt{n}$-CLT (see e.g. Tierney, 1994), and it is proved by Roberts (1999) that the above independence sampler fails to have a $\sqrt{n}$-CLT for $k \geq 2$, thus showing precise agreement between all three different perspectives for convergence of this particular independence sampler.

5. Example #2: Normal Distributions.

Let $\mathcal{X} = \mathbf{R}$ with Lebesgue measure. Let $\pi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$ be the density of a standard normal distribution, and $q(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/2\sigma^2}$ that of a $N(0, \sigma^2)$ distribution for some fixed $\sigma > 0$. Then $w(x) = \pi(x)/q(x) = \sigma\, e^{-(x^2/2)(1-\sigma^{-2})} = \sigma\, e^{(x^2/2)(\sigma^{-2}-1)}$.

If $\sigma \geq 1$, then $w^* \equiv \sup_{x\in\mathcal{X}} w(x) = \sigma < \infty$, and the chain is uniformly ergodic with $\|P^n(x,\cdot) - \Pi(\cdot)\| \leq (1 - \frac{1}{\sigma})^n$. (In particular, if $\sigma = 1$, then we again have i.i.d. sampling.)

However, if $0 < \sigma < 1$, then $w^* = \infty$, and again the chain is not uniformly or geometrically ergodic. In that case, we compute (where $Y \sim Q(\cdot)$) that

$$m(x) \;=\; \mathbf{E}[\alpha(x, Y)] \;=\; \int_{-\infty}^{\infty} \alpha(x, y)\; q(y)\, dy$$

$$= \int_{-\infty}^{\infty} \min(1,\ e^{(y^2-x^2)(\sigma^{-2}-1)/2})\ \frac{1}{\sigma\sqrt{2\pi}}\, e^{-y^2/2\sigma^2}\, dy$$

$$= 2 \int_{-\infty}^{-|x|} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-y^2/2\sigma^2}\, dy \ + \ \int_{-|x|}^{|x|} e^{(y^2-x^2)(\sigma^{-2}-1)/2}\ \frac{1}{\sigma\sqrt{2\pi}}\, e^{-y^2/2\sigma^2}\, dy$$

$$= 2\,\Phi(-|x|/\sigma) \ + \ \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2(\sigma^{-2}-1)/2} \int_{-|x|}^{|x|} e^{-y^2/2}\, dy$$

$$= 2\,\Phi(-|x|/\sigma) \ + \ \sigma^{-1}\, e^{-x^2(\sigma^{-2}-1)/2}\, [1 - 2\,\Phi(-|x|)]\,. \tag{5}$$

To proceed, again start at 0 since $w(0) = \inf_{y\in\mathcal{X}} w(y)$. Then from Corollary 7, with $Z \sim \Pi(\cdot)$,

$$\|P^n(0,\cdot) - \Pi(\cdot)\| \ \leq\ \mathbf{E}[(1-m(Z))^n] \ = \ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (1-m(x))^n\, e^{-x^2/2}\, dx\,, \tag{6}$$

with $m(x)$ as in (5).

In the case $\sigma = 0.5$, the bound (6) gives:

Upper bounds on $\|P^n(0,\cdot) - \Pi(\cdot)\|$ from (6), with $\sigma = 0.5$:

σ	n	upper bound
0.5	10	0.145
0.5	100	0.0546
0.5	1000	0.0219
0.5	7000	0.0104
0.5	8000	0.00989

This indicates that for the case $\sigma = 0.5$, the chain converges (to within 0.01 of stationarity) within 8,000 iterations, which is not too quick but not overly slow. From the numbers in the table, the polynomial order appears to be approximately $O(n^{-0.4})$.

As for lower bounds, we take $z = 2.5$ in Theorem 8. Then, writing $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx$ for the cumulative distribution function of a standard normal distribution, we have that

$$p_z \ = \ \Pi\{y \in \mathcal{X} : w(y) > w(z)\} \ = \ \Pi\{y \in \mathcal{X} : y^2 > z^2\} \ = \ 2\,\Phi(-z) \ = \ 2\,\Phi(-2.5)\,,$$

while

$$q_z \ = \ Q\{y \in \mathcal{X} : w(y) > w(z)\} \ = \ Q\{y \in \mathcal{X} : y^2 > z^2\} \ = \ 2\,\Phi(-z/\sigma) \ = \ 2\,\Phi(-2.5/\sigma)\,.$$

So, by Theorem 8,

$$\|P^n(x,\cdot) - \Pi(\cdot)\| \ \geq\ p_z - (1-(1-q_z)^n) \ \doteq\ 2\,\Phi(-2.5) - (1-(1-\Phi(-2.5/\sigma))^n)\,. \tag{7}$$

If $\sigma = 0.5$, then this bound equals 0.0101 when $n = 4000$. This shows that when $\sigma = 0.5$, the chain converges in between $n_* = 4000$ and $n^* = 8000$ iterations, giving a ratio $n^*/n_* = 2$, again showing fairly tight bounds, specifically that we have identified the precise convergence time within a factor of 2.

When $\sigma = 0.2$, the values from (6) become more extreme:

Upper bounds on $\|P^n(0,\cdot) - \Pi(\cdot)\|$ from (6), with $\sigma = 0.2$:

σ	n	upper bound
0.2	1000	0.399
0.2	10,000	0.340
0.2	10^{10}	0.149
0.2	10^{20}	0.0452
0.2	10^{30}	0.0148
0.2	10^{33}	0.010725
0.2	10^{34}	0.00963
0.2	10^{40}	0.00507

This indicates that 10^{34} iterations (a huge number!) are sufficient for convergence, i.e. we can take $n^* = 10^{34}$. Furthermore, the polynomial order appears to be approximately $O(n^{-1/20})$.

One can again wonder if this bound is hugely conservative. However, with $\sigma = 0.2$, the lower bound (7) is > 0.01 when $n = 5 \times 10^{32}$. So, the chain still has not converged after 5×10^{32} iterations. That is, the convergence time is between $n_* \equiv 5 \times 10^{32}$ and $n^* \equiv 10^{34}$. This gives a ratio of $n^*/n_* = 20$, which is still fairly small, especially considering the huge values of n^* and n_* involved. Thus, once again we have identified the convergence time fairly precisely, even though the chain converges extremely slowly.

Remark. In the case $\sigma = 0.2$, numerical computation of the integral (6) required special care regarding numerical precision. This is because $m(x)$ is often extremely close to zero, requiring lots of computations like $(1-10^{-21})^{10^{34}}$, which should be very close to zero, but which would be rounded to $1^{10^{34}} = 1$ using ordinary double-precision floating-point arithmetic. To avoid such problems, we performed these computations using Mathematica (Wolfram, 1988) with the parameter `WorkingPrecision` set to a large number (e.g. 55), instead of the default value (15.95). This produced accurate numerical integration even with such small values of $m(x)$ and such large values of n.

6. Example #3: Unit Interval.

Finally, we consider the independence sampler example of Douc et al. (2007), where $\mathcal{X} = (0,1]$, $\pi(x) \equiv 1$, and $q(x) = (r+1)x^r$ for some $r > 0$. (We exclude the point 0 from the state space simply to avoid the problem that $q(0) = 0$; alternatively we could just re-define the single value of $q(0)$ arbitrarily.) Douc et al. (2007) use hitting times of small sets to prove that when $r = 2$, then $\|P^n(x,\cdot) - \Pi(\cdot)\| < 0.1$ for $n = 500$, while when $r = 1/2$, then $\|P^n(x,\cdot) - \Pi(\cdot)\| < 0.1$ for $n = 50$.

Here $w(x) = \pi(x)/q(x) = x^{-r}/(r+1)$, so $\alpha(x,y) = \min(1, x^r/y^r)$. We then compute (with $Y \sim Q(\cdot)$) that

$$\begin{aligned} m(x) \;&=\; \mathbf{E}[\alpha(x,Y)] \;=\; \int_0^1 \min(1,\, x^r/y^r)\,(r+1)\,y^r\,dy \\ &=\; \int_0^x 1\,(r+1)\,y^r\,dy \;+\; \int_x^1 (x^r/y^r)\,(r+1)\,y^r\,dy \\ &=\; x^{r+1} + x^r(1-x)(r+1) \;=\; x^r(r+1) - rx^{r+1}\,. \end{aligned}$$

If we start at the state 1 for simplicity, since $w(1) = \inf_{y \in \mathcal{X}} w(y)$, then we have from Corollary 7 (with $Z \sim \Pi(\cdot)$) that

$$\|P^n(1,\cdot) - \Pi(\cdot)\| \;\le\; \mathbf{E}[(1 - m(Z))^n] \;=\; \int_0^1 (1 - x^r(r+1) + rx^{r+1})^n\,dx\,. \tag{8}$$

If $r = 2$, then (8) is < 0.1 when $n = 28$. (Thus, we have improved the bound 500 of Douc et al., 2007, by a factor of nearly 18. On the other hand, the bounds of Douc et al. are based on very general results, while ours are specifically designed for the independence sampler.) By contrast, to make (8) be < 0.01 requires $n = 2640$, i.e. we have that $n^* = 2640$. (This again shows the slow nature of the convergence; it is much faster to get within 0.1 of stationarity than to get within 0.01 of stationarity.)

As for lower bounds, for $z \in \mathcal{X}$ we compute that $p_z = \Pi(0,z) = z$, and $q_z = Q(0,z) = z^{r+1}$, so Theorem 8 gives that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\ge\; p_z - (1 - (1-q_z))^n \;=\; z - (1 - (1 - z^{r+1})^n)\,. \tag{9}$$

If $r = 2$, then choosing $z = 0.16$ gives that for any $x \in \mathcal{X}$, $\|P^n(x,\cdot) - \pi(\cdot)\| > 0.1$ for $n = 15$, so the time to get within 0.1 of stationarity is between 15 and 28. Also, choosing $z = 0.016$ gives that for any $x \in \mathcal{X}$, $\|P^n(x,\cdot) - \pi(\cdot)\| > 0.01$ for $n = 1450$, so the time to get within 0.01 of stationarity is between $n_* = 1450$ and $n^* = 2640$, a factor of $n^*/n_* \doteq 1.8$.

If $r = 1/2$, then (8) is < 0.1 when $n = 2$ (thus improving on Douc et al.'s bound of 50), and is < 0.01 when $n = 9$, so we can take $n^* = 9$ in this case, indicating very fast convergence. On the other hand, if $r = 5$, then $n^* = 1.1 \times 10^9$ iterations are required to make (8) be < 0.01, and after $n_* = 3.5 \times 10^8$ iterations (9) is still > 0.01, thus bounding the time to convergence within a factor of $1.1 \times 10^9 \,/\, 3.5 \times 10^8 \doteq 3.1$, and again showing how small changes in parameter values (e.g. changing r from 2 to 5) can have a tremendous effect on convergence times.

7. Discussion.

This paper has provided precise quantitative upper and lower bounds for independence samplers which are not geometrically ergodic. We provided both general results (Theorem 6, Corollary 7, Theorem 8), and applications to specific examples. For the exponential example, we proved that with $k = 2$ the convergence time is between 24 and 50 iterations, while for $k = 5$ it is between 4,000,000 and 14,000,000. For the normal example, we proved that with $\sigma = 0.5$ the convergence time is between 4,000 and 8,000, while for $\sigma = 0.2$ it is between 5×10^{32} and 10^{34}.

We believe this analysis to be useful for several reasons:

- It provides clear examples of precise quantitative convergence bounds for specific examples of non-geometrically ergodic MCMC algorithms, thus adding to the previous results of e.g. Douc et al. (2007).
- It complements the previous analysis of Liu (1996) and Smith and Tierney (1996), who studied convergence rates for geometrically ergodic independence samplers.
- It shows the usefulness of the coupling method for bounding convergence of MCMC, without the use of minorisation conditions as in e.g. Rosenthal (1995) and Douc et al. (2007).
- It shows that slight changes to a parameter value (e.g., changing k from 2 to 5) can have an enormous effect on convergence times (e.g., from 50 to 14,000,000).
- It illustrates that even simple-seeming Markov chains can often converge extremely slowly, requiring millions of iterations or more, so users of MCMC should not be confident of convergence without careful analysis. (For related discussion see e.g. Jones and Hobert, 2001.)

We hope that in the future the coupling method can be applied to other Markov chains and other examples of MCMC, in a continuing effort to better understand the nature and speed of the convergence of Markov chains to their stationarity distributions.

Acknowledgements. We thank Kerrie Mengersen for asking a question that inspired this paper, and Gersende Fort and Randal Douc and Eric Moulines for organising a workshop in Fleurance, France where this research was initiated, and Radford Neal for a useful conversation about floating-point precision in numerical integration, and the anonymous referee for a number of helpful comments.

REFERENCES

P.H. Baxendale (2005), Renewal theory and computable convergence rates for geometrically ergodic Markov chains. Ann. Appl. Prob. **15**, 700–738.

N.H. Bingham, C.M. Goldie, and J.L. Teugels (1987), Regular Variation. Cambridge University Press, Cambridge.

R. Douc, E. Moulines, and J.S. Rosenthal (2004), Quantitative bounds on convergence of time-inhomogeneous Markov chains. Ann. Appl. Prob. **14**, 1643–1665.

R. Douc, E. Moulines, and P. Soulier (2007), Computable convergence rates for sub-geometric ergodic Markov chains. Bernoulli **13**, 831–848.

G. Fort and E. Moulines (2000), Computable Bounds For Subgeometrical And Geometrical Ergodicity. Available at: http://citeseer.ist.psu.edu/fort00computable.html

G. Fort and E. Moulines (2003), Polynomial ergodicity of Markov transition kernels. Stoch. Proc. Appl. **103**, 57–99.

C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci., Vol. **7**, No. **4**, 473–483.

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

S.F. Jarner and G.O. Roberts (2002), Polynomial convergence rates of Markov chains. Ann. Appl. Prob., 224–247, 2002.

G.L. Jones and J.P. Hobert (2001), Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Stat. Sci. **16**, 312–334.

G.L. Jones and J.P. Hobert (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. Ann. Stat. **32**, 784–817.

J. Liu (1996), Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Stat. and Comput. **6**, 113–119.

D. Marchev and J.P. Hobert (2004), Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate Student's t model. J. Amer. Stat. Assoc. **99**, 228–238.

K. L. Mengersen and R. L. Tweedie (1996), Rates of convergence of the Hastings and Metropolis algorithms. Ann. Stat. **24**, 101–121.

S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London. Available at: http://probability.ca/MT/

G.O. Roberts (1999), A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms. J. Appl. Prob. **36**, 1210–1217.

G.O. Roberts and J.S. Rosenthal (1998), Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). Can. J. Stat. **26**, 5–31.

G.O. Roberts and J.S. Rosenthal (2004), General state space Markov chains and MCMC algorithms. Prob. Surv. **1**, 20–71.

G.O. Roberts and R.L. Tweedie (1996), Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms. Biometrika **83**, 95–110.

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

J.S. Rosenthal (1997), Independence sampler Java applet. Available at: http://probability.ca/jeff/java/exp.html

J.S. Rosenthal (2002), Quantitative convergence rates of Markov chains: A simple account. Electronic Comm. Prob. **7**, 123–128.

R.L. Smith and L. Tierney (1996), Exact transition probabilities for the independence Metropolis sampler. Preprint.

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

S. Wolfram (1988), Mathematica: A system for doing mathematics by computer. Addison-Wesley, New York.