On the central limit theorem for geometrically ergodic Markov chains

Olle Häggström*

November 22, 2007

Abstract

Let $X_0, X_1, \dots$ be a geometrically ergodic Markov chain with state space $\mathcal{X}$ and stationary distribution π. It is known that if $h : \mathcal{X} \to \mathbf{R}$ satisfies $\pi(|h|^{2+\varepsilon}) < \infty$ for some $\varepsilon > 0$, then the normalized sums of the $h(X_i)$'s obey a central limit theorem. Here we show, by means of a counterexample, that the condition $\pi(|h|^{2+\varepsilon}) < \infty$ cannot be weakened to only assuming a finite second moment, i.e., $\pi(h^2) < \infty$.

1 Introduction

Let $X_0, X_1, \dots$ be a Markov chain with state space $\mathcal{X}$, transition kernel P, and a unique stationary distribution π, and let $h : \mathcal{X} \to \mathbf{R}$ be some real-valued function of the state space. This paper is concerned with under what conditions on the Markov chain (i.e., on P) and on h the sum $\sum_{i=1}^n h(X_i)$ is asymptotically normal as $n \to \infty$. In other words, when does a central limit theorem hold?

To state the results, we first need some definitions. For two probability measures μ and ν on $\mathcal{X}$, define their **total variation distance** $\mathrm{d_{TV}}(\mu, \nu)$ as

$$\mathrm{d_{TV}}(\mu, \nu) \;=\; \sup_A |\mu(A) - \nu(A)|$$

where the supremum is taken over all measurable $A \subset \mathcal{X}$.

We write $P^n(x, A)$ for the n-step transition law for the Markov chain, i.e., $P^n(x, A) = \mathbf{P}(X_n \in A \,|\, X_0 = x)$. If the chain starts in state $X_0 = x$, then the distribution of X_n is $P^n(x, \cdot)$.

Definition 1.1 *The Markov chain with transition kernel P and unique stationary distribution π is said to be* **ergodic** *if for any $x \in \mathcal{X}$ we have*

$$\lim_{n \to \infty} \mathrm{d_{TV}}(P^n(x, \cdot), \pi) \;=\; 0\,.$$

If furthermore there exist $C(x)$ and a $\rho < 1$ such that

$$\mathrm{d_{TV}}(P^n(x, \cdot), \pi) \;\le\; C(x)\rho^n \tag{1}$$

for every x and every n, then the chain is said to be **geometrically ergodic**. *Finally, if in (1) we can take $C(x)$ to be a constant (i.e., independent of x), then the chain is said to be* **uniformly ergodic**.

*Research supported by the Swedish Research Council.

Write $N(0, \sigma^2)$ for the Gaussian distribution with mean 0 and variance σ^2; we allow for the possibility $\sigma^2 = 0$, in which case $N(0, \sigma^2)$ simply is a unit point mass at 0. The following result goes back to Ibragimov and Linnik [4].

Theorem 1.2 *If $X_0, X_1, \ldots$ is a geometrically ergodic Markov chain with stationary distribution π, and if for some $\varepsilon > 0$ the function $h : \mathcal{X} \to \mathbf{R}$ satisfies $\pi(|h|^{2+\varepsilon}) < \infty$, then there exists a σ such that the normalized sum*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(X_i) - \pi(h)]$$

converges in distribution to a $N(0, \sigma^2)$ distribution.

It is known under certain additional assumptions that for asymptotic normality, the condition $\pi(|h|^{2+\varepsilon}) < \infty$ can be weakened to just a finite second moment: $\pi(h^2) < \infty$. In particular, this is true if geometric ergodicity is strengthened to uniform ergodicity, as shown by Cogburn [2], and it is also true if the chain is assumed to be reversible, as shown by Roberts and Rosenthal [5]. But is it true in general? In a recent survey paper, Roberts and Rosenthal [6] emphasize the importance of this question to Markov chain Monte Carlo. Here we will show, by means of a counterexample, that the answer is no:

Theorem 1.3 *There exists a geometrically ergodic Markov chain $X_0, X_1, \ldots$ with stationary distribution π, and a function $h : \mathcal{X} \to \mathbf{R}$ satisfying $\pi(h^2) < \infty$, such that the following holds. For no choice of σ^2 does*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(X_i) - \pi(h)]$$

converge in distribution to a $N(0, \sigma^2)$ distribution.

In the example we shall exhibit, we will see that no other way of normalizing sums (as opposed to the usual $\frac{1}{\sqrt{n}}$) will recover the asymptotic normality. It is also worth mentioning that no fancy state space is needed; in the example $\mathcal{X}$ will in fact be countable.

The rest of the paper is devoted to proving Theorem 1.3. In Section 2 we define the Markov chain that will be used in the counterexample, and demonstrate that it is geometrically ergodic. Then, in Section 3, we introduce the function h and show that it has the properties needed to serve as a counterexample in Theorem 1.3.

2 The Markov chain

We first define the state space $\mathcal{X}$ on which the Markov chain will be living. Let $\tilde{\mathcal{X}}$ denote the set of all integer triples (a, b, c) such that $a \geq 1$, $b \in \{1, \ldots, a\}$ and $c \in \{-1, 1\}$, and let $\mathcal{X} = \{0\} \cup \tilde{\mathcal{X}}$. For any $x \in \mathcal{X}$, define $\alpha(x)$, $\beta(x)$ and $\gamma(x)$ as

$$\alpha(x) = \begin{cases} 0 & \text{if } x = 0 \\ a & \text{if } x = (a, b, c) \in \tilde{\mathcal{X}}\,, \end{cases} \tag{2}$$

$$\beta(x) = \begin{cases} 0 & \text{if } x = 0 \\ b & \text{if } x = (a, b, c) \in \tilde{\mathcal{X}}\,, \end{cases}$$

and

$$\gamma(x) = \begin{cases} 0 & \text{if } x = 0 \\ c & \text{if } x = (a,b,c) \in \tilde{\mathcal{X}}\,. \end{cases} \tag{3}$$

The dynamics of the Markov chain $X_0, X_1, \ldots$ is as follows. It is only at the times when $X_i = 0$ that there is any actual randomness in the choice of the next state X_{i+1}. If X_i is in state $(a,b,c) \in \tilde{\mathcal{X}}$, then the chain moves with probability 1 to state

$$\begin{cases} 0 & \text{if } b = 1 \\ (a, b-1, c) & \text{otherwise.} \end{cases} \tag{4}$$

If, on the other hand, $X_i = 0$, then the next state is chosen from $\mathcal{X}$ according to

$$X_{i+1} = \begin{cases} 0 & \text{with probability } \frac{1}{2} \\ (a,b,c) & \text{with probability } \begin{cases} 2^{-(a+2)} & \text{if } a = b \\ 0 & \text{otherwise.} \end{cases} \end{cases} \tag{5}$$

The easiest way to think of this Markov chain is as follows. Let $\ldots, Y_{-1}, Y_0, Y_1, \ldots$ be a sequence of i.i.d. random variables such that $\mathbf{P}(Y_i = 0) = \mathbf{P}(Y_i = 1) = 1/2$. Then construct $\ldots, X_{-1}, X_0, X_1, \ldots$ by

- if $Y_i = 0$, then let $X_i = 0$,
- otherwise, let $X_i = (a,b,c)$, where
  - a is the length of the consecutive sequence (run) of 1's in $(\ldots, Y_{-1}, Y_0, Y_1, \ldots)$ that contains Y_i,
  - b is the number of 1's in this run remaining at time i (including Y_i itself),
  - for each run of 1's in $(\ldots, Y_{-1}, Y_0, Y_1, \ldots)$, c is taken to be identical in all corresponding X_i's, taking value -1 or 1 with probability $1/2$ each, independently for separate runs.

That this indeed produces a Markov chain with the desired transition kernel is immediate from the construction. It is also clear the the chain is irreducible and aperiodic, and has a stationary distribution π given by

$$\pi(0) \;=\; \tfrac{1}{2}$$

and, for any $(a,b,c) \in \tilde{\mathcal{X}}$,

$$\pi((a,b,c)) \;=\; 2^{-(a+3)}\,.$$

In order for this construction to be useful as a counterexample in Theorem 1.3, we need to prove the following.

Proposition 2.1 *The Markov chain with state space $\mathcal{X}$ and transition kernel given by* (4) *and* (5) *is geometrically ergodic.*

Proof: Pick any state $x \in \mathcal{X}$, and let $X_0, X_1, \ldots$ be a Markov chain with the prescribed transition kernel starting in $X_0 = x$. We will construct this chain together with another Markov chain $X'_0, X'_1, \ldots$ with the same transition kernel, but with X'_0 chosen according to π. Then X'_i will have distribution π for any i, and it follows by the usual coupling inequality that for any n we have

$$\mathrm{d_{TV}}(P^n(x,\cdot),\pi) \;\le\; \mathbf{P}(X_n \neq X'_n)\,. \tag{6}$$

So in order to prove rapid decay of $\mathrm{d_{TV}}(P^n(x,\cdot),\pi)$, the challenge is to produce a coupling where the two chains coalesce (and stay together) as early as possible.

For any fixed $x \in \mathcal{X}$, there exists a deterministic number $k \geq 0$ such that if $X_0 = x$, then we know for certain that X_k will equal 0. Indeed, if $x = 0$, then we can take $k = 0$, while if $x = (a, b, c)$, then we can take $k = b$. In both cases, $k = \beta(x)$; hence $\beta(X_i)$ may be interpreted as the waiting time from time i until the chain will hit the state 0.

To produce the coupling, we begin by generating $X_0, X_1, \ldots, X_k$, which is a deterministic sequence. We know that $X_k = 0$, and by integrating β with respect to $P(0,\cdot)$ (i.e., the transition probabilities indicated in (5)), we get that

$$\mathbf{P}(\beta(X_{k+1}) = i) \; = \; 2^{-(i+1)} \quad \text{for } i = 0, 1, 2, \ldots \tag{7}$$

Furthermore, X'_{k+1} has distribution π, and integrating β with respect to π yields that $\beta(X'_{k+1})$ has the same distribution (7) as $\beta(X_{k+1})$. We are therefore free to pick X_{k+1} and X'_{k+1} in such a way that $\mathbf{P}(\beta(X_{k+1}) = \beta(X'_{k+1})) = 1$; let us do that. (For completeness, we also fill in $X'(k), X'(k-1), \ldots, X'_0$ backwards in time using the time-reversal of the transition kernel P.) Then the two chains will continue deterministically until and including time $k + 1 + \beta(X_{k+1})$ when they are both forced to take value 0. From that time and on, we can generate the X_n chain and the X'_n chain by letting them make identical moves according to P. This defines the coupling, which for any n has the property that

$$\begin{aligned}
\mathbf{P}(X_n \neq X'_n) &\leq \mathbf{P}(n < k + 1 + \beta(X_{k+1})) \\
&= \mathbf{P}(\beta(X_{k+1}) > n - k - 1) \\
&= \begin{cases} 1 & \text{for } n \leq k \\ \left(\frac{1}{2}\right)^{n-k} & \text{for } n > k \end{cases}
\end{aligned}$$

which for any n is bounded by $\left(\frac{1}{2}\right)^{n-k} = 2^k \left(\frac{1}{2}\right)^n$. Hence, using (6), we get

$$\begin{aligned}
\mathrm{d_{TV}}(P^n(x,\cdot),\pi) &\leq \mathbf{P}(X_n \neq X'_n) \; \leq \; 2^k \left(\tfrac{1}{2}\right)^n \\
&= 2^{\beta(x)} \left(\tfrac{1}{2}\right)^n ,
\end{aligned}$$

which means that the chain is geometrically ergodic with $\rho = \frac{1}{2}$ and $C(x) = 2^{\beta(x)}$. □

Remark. Readers interested in the subtleties of coupling of Markov chains may note the following feature of the above coupling. Even though the conditional distribution of X_{k+1} given $(X_0, X_1, \ldots X_k)$ is given by (5) as it ought to (otherwise $X_0, X_1, \ldots$ would have the wrong distribution and the coupling would not be correct), we get a different distribution of X_{k+1} if we condition on the past of *both* chains, i.e., on $(X_0, X_1, \ldots X_k)$ and on $(X'_0, X'_1, \ldots X'_k)$. Indeed, if $\beta(X'_k) > 0$, then X_{k+1} is *forced* to take a value such that $\beta(X_{k+1}) = \beta(X'_k) - 1$, which is clearly not in agreement with (5). In the language of Rosenthal [7], this means that we are dealing with a *non-faithful* coupling. Non-faithful couplings are unusual in applications; see also Häggström [3] for an example of the kind of counterintuitive behavior they may exhibit. □

3 The function h

The choice of the function $h : \mathcal{X} \to \mathbf{R}$ will be made with the specific target of making the partial sums $\sum_{i=1}^n X_i$ fit the following lemma, which deals with a situation reminiscent of *Twin Peaks.*

Lemma 3.1 *Let $Z_1, Z_2, \ldots$ be a sequence of real-valued random variables with the property that there exist arbitrarily large n such that for some normalizing constants s_n we have*

$$\mathbf{P}\left(-1.001 \leq \frac{Z_n}{s_n} \leq -0.999\right) \geq 0.1$$

and

$$\mathbf{P}\left(0.999 \leq \frac{Z_n}{s_n} \leq 1.001\right) \geq 0.1\,.$$

Then, for no choice of $\mu_1, \mu_2, \ldots$ and $\sigma_1, \sigma_2, \ldots$, does $\frac{Z_n - \mu_n}{\sigma_n}$ converge in distribution to $N(0,1)$.

Proof: Obvious. $\square$

In the construction of h, we will let $\{A_k\}_{k=1}^{\infty}$ and $\{B_k\}_{k=1}^{\infty}$ be two strictly and rapidly increasing sequences of positive integers – precisely how rapidly will soon be specified. Recall from (2) and (3) the definitions of $\alpha(x)$ and $\gamma(x)$, and let

$$h(x) = \begin{cases} \frac{B_k}{A_k} 2^{\frac{A_k+2}{2}} \gamma(x) & \text{if } \alpha(x) = A_k \text{ for some } k \\ 0 & \text{otherwise.} \end{cases}$$

We also define a kind of truncation of h by setting

$$h_m(x) = \begin{cases} \frac{B_k}{A_k} 2^{\frac{A_k+2}{2}} \gamma(x) & \text{if } \alpha(x) = A_k \text{ for some } k \leq m \\ 0 & \text{otherwise.} \end{cases}$$

Note that under π, $\gamma(x)$ equals -1 and $+1$ with equal conditional probabilities given $\alpha(x)$. Hence, by symmetry, and the fact that h_m is bounded, we get $\pi(h_m) = 0$. Furthermore, by Theorem 1.2, there exists a σ_m such that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_m(X_i) \tag{8}$$

converges in distribution to $N(0, \sigma_m^2)$.

We now go on to specify the sequences $\{A_k\}_{k=1}^{\infty}$ and $\{B_k\}_{k=1}^{\infty}$. First set, somewhat arbitrarily, $A_1 = B_1 = 1$. This is enough to define the truncated function h_1. To define $A_2, A_3, \ldots$ and $B_2, B_3, \ldots$, we go on inductively as follows.

Suppose that $A_1, \ldots, A_{k-1}$ as well as $B_1, \ldots, B_{k-1}$ are specified; then we also know the truncated function h_{k-1}, and the variance σ_{k-1}^2 arising in the asymptotic distribution of (8) with $m = k-1$. We are then free to choose first B_k and then A_k large enough so that the following conditions hold.

(i) $B_k > 3000\sigma_{k-1}$

(ii) A_k is large enough so that the approach to normality in (8) with $m = k-1$ guarantees that

$$\mathbf{P}\left(\frac{1}{\sqrt{2^{A_k+2}}} \sum_{i=1}^{2^{A_k+2}} h_{k-1}(X_i) \in (-3\sigma_{k-1}, 3\sigma_{k-1})\right) \geq 0.99$$

(iii) $A_k \geq 2^k B_k^2$

(iv) $A_k \geq A_{k-1} + 10$

(That **(ii)** can be ensured by picking A_k large is, of course, due to the fact that $\frac{1}{\sqrt{2\pi}} \int_{-3}^{3} e^{-x^2/2} dx > 0.99$.) Thus, A_k and B_k are specified, and the induction can continue.

This defines the function h. To use h as a counterexample for Theorem 1.3, we first need to establish that it has a finite second moment under the stationary distribution π.

Lemma 3.2 *With h defined as above, we get $\pi(h^2) < \infty$.*

Proof: For $k = 1$ we have that

$$\pi(\{x \in \mathcal{X} : \alpha(x) = A_k\}) \left(\frac{B_k 2^{\frac{A_k+2}{2}}}{A_k} \right)^2 = \frac{1}{8}(2^{3/2})^2 = 1$$

and a further direct calculation gives

$$\begin{aligned}
\pi(h^2) &= \sum_{k=1}^{\infty} \pi(\{x \in \mathcal{X} : \alpha(x) = A_k\}) \left(\frac{B_k 2^{\frac{A_k+2}{2}}}{A_k} \right)^2 \\
&= 1 + \sum_{k=2}^{\infty} \pi(\{x \in \mathcal{X} : \alpha(x) = A_k\}) \left(\frac{B_k 2^{\frac{A_k+2}{2}}}{A_k} \right)^2 \\
&= 1 + \sum_{k=2}^{\infty} A_k 2^{-(A_k+2)} \left(\frac{B_k 2^{\frac{A_k+2}{2}}}{A_k} \right)^2 \\
&= 1 + \sum_{k=2}^{\infty} \frac{B_k^2}{A_k} \\
&\leq 1 + \sum_{k=2}^{\infty} 2^{-k} = \frac{3}{2}
\end{aligned}$$

where the inequality is due to condition **(iii)**. $\square$

For the next lemma, we introduce for simplicity the notation $Z_n = \sum_{i=1}^{n} X_i$ and $C_k = 2^{A_k+2}$.

Lemma 3.3 *Let the chain $X_0, X_1, \ldots$ start according to the stationary distribution π. Then, for all sufficiently large k, we have*

$$\mathbf{P}\left(-1.001 \leq \frac{Z_{C_k}}{B_k \sqrt{C_k}} \leq -0.999 \right) \geq 0.1 \tag{9}$$

and

$$\mathbf{P}\left(0.999 \leq \frac{Z_{C_k}}{B_k \sqrt{C_k}} \leq 1.001 \right) \geq 0.1\,. \tag{10}$$

Proof: Without loss of generality, we may assume that the chain $X_0, X_1, \ldots$ is obtained from the bi-infinite i.i.d. sequence $\ldots, Y_{-1}, Y_0, Y_1, \ldots$ as in Section 2. Define events E_k^1, E_k^2, E_k^3 and E_k^4 as follows.

- Let E_k^1 be the event that the sequence $(Y_1, \ldots, Y_{C_k})$ is not intersected by any run of 1's of length A_{k+1} or more. By condition **(iv)**, E_k^1 has probability at least $1 - 2 \cdot 2^{-10} = 1 - \frac{1}{512}$.

- Let E_k^2 be the event that the sequence $(Y_1, \ldots, Y_{C_k})$ contains exactly one run of 1's (from the bi-infinite sequence) of length exactly A_k. By a standard Poisson approximation argument (see, e.g., Barbour et al [1]), the distribution of the number of such runs converges in total variation to a Poisson distribution with mean 1, so that $\mathbf{P}(E_k^2) \to e^{-1} \approx 0.368$ as $k \to \infty$.

- Let E_k^3 be the event that $(Y_1, \ldots, Y_{C_k})$ is intersected by no other runs of length A_k than those which it contains. Obviously, $\mathbf{P}(E_k^3) \to 1$ as $k \to \infty$.

- Let E_k^4 be the event that

$$-0.001 \leq \frac{1}{B_k \sqrt{C_k}} \sum_{i=1}^{C_k} h_{k-1}(X_i) \leq 0.001 .$$

By condition **(ii)**, we have that

$$\mathbf{P}\left(-3 \leq \frac{1}{\sigma_{k-1}\sqrt{C_k}} \sum_{i=1}^{C_k} h_{k-1}(X_i) \leq 3 \right) \geq 0.99 ,$$

and the choice **(i)** of B_k therefore ensures that $\liminf_{k\to\infty} \mathbf{P}(E_k^4) \geq 0.99 = 1-0.01$.

Finally, define the event $E_k = E_k^1 \cap E_k^3 \cap E_k^3 \cap E_k^4$. Bonferroni's inequality gives that

$$\liminf_{k\to\infty} \mathbf{P}(E_k) \geq e^{-1} - \frac{1}{512} - 0.01 > 0.2 . \tag{11}$$

On the event E_k, the (unique) run of 1's of length A_k in $(Y_1, \ldots, Y_{C_k})$ contributes a term $+1$ or -1 (depending on $\gamma(X_i)$ for the X_i's corresponding to the run) to $\frac{Z_{C_k}}{B_k\sqrt{C_k}}$, while $\frac{1}{B_k\sqrt{C_k}} \sum_{i=1}^{C_k} h_{k-1}(X_i)$ contributes between -0.001 and 0.001. Hence we have, still on the event E_k, that

$$0.999 \leq \left| \frac{Z_{C_k}}{B_k\sqrt{C_k}} \right| \leq 1.001 .$$

Conditional on E_k, we have by symmetry that Z_{C_k} is positive or negative with probability $\frac{1}{2}$ each. In combination with (11), this implies (9) and (10), and we are done. □

Proof of Theorem 1.3: Choose the Markov chain $X_0, X_1, \ldots$ and the function h as above. By Lemma 3.2, we have $\pi(h^2) < \infty$, while a combination of Lemmas 3.3 and 3.1 implies that the sums $\sum_{i=1}^{n} h(X_i)$ are not asymptotically normal. Hence the theorem is established. □

Remark. Since $B_k \to \infty$ as $k \to \infty$, we can deduce from Lemma 3.3 that the $1/\sqrt{n}$-normalized sums $\frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} h(X_i)$ fail to define a tight sequence of probability distributions. □

References

[1] Barbour, A.D., Holst, L. and Janson, S. (1992) *Poisson Approximation*, Oxford University Press.

[2] Cogburn, R. (1972) The central limit theorem for Markov processes, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol II* (Le Cam, L, Neyman, J. and Scott, E., eds), 485–512.

[3] Häggström, O. (2001) A note on disagreement percolation, *Random Structures Algorithms* **18**, 267–278.

[4] Ibragimov, I.A., and Linnik, Y.V. (1971) *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff, Groningen.

[5] Roberts, G.O. and Rosenthal, J.S. (1997) Geometric ergodicity and hybrid Markov chains, *Electr. Comm. Probab.* **2**, 13–25.

[6] Roberts, G.O. and Rosenthal, J.S. (2004) General state space Markov chains and MCMC algorithms, preprint, `http://xxx.lanl.gov/abs/math.PR/0404033` .

[7] Rosenthal, J.S. (1997) Faithful coupling of Markov chains: now equals forever, *Adv. Appl. Math.* **18**, 372–381.

Dept of Mathematics
Chalmers University of Technology
412 96 Göteborg
Sweden
`olleh@math.chalmers.se`
`http://www.math.chalmers.se/~olleh/`