PREPRINT

On the Central Limit Theorem for Geometrically Ergodic Markov Chains

OLLE HÄGGSTRÖM

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
GÖTEBORG UNIVERSITY
Göteborg Sweden 2004
On the Central Limit Theorem for Geometrically Ergodic Markov Chains

Olle H"aggstr"om
On the central limit theorem for geometrically ergodic Markov chains

Olle Haggström*

June 28, 2004

Abstract

Let \(X_0, X_1, \ldots\) be a geometrically ergodic Markov chain with state space \(X\) and stationary distribution \(\pi\). It is known that if \(h : X \to \mathbb{R}\) satisfies \(\mathbb{E}[(h(X_0))^2] < \infty\) for some \(x \in X\), then the normalized sum of the \(X_t's\) obey a central limit theorem. Here we show, by means of a counterexample, that the condition \(\mathbb{E}[(h(X_0))^2] < \infty\) cannot be weakened to only assuming a finite second moment, i.e., \(\mathbb{E}[h(X_0)^2] < \infty\).

1 Introduction

Let \(X_0, X_1, \ldots\) be a Markov chain with state space \(X\), transition kernel \(P\), and a unique stationary distribution \(\pi\), and let \(h : X \to \mathbb{R}\) be some real-valued function of the state space. This paper is concerned with under what conditions on the Markov chain \((X_t, h)\) on \(P\) and on the sum \(\sum_{t=0}^{\infty} h(X_t)\) 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 \(\mu\) and \(\nu\) on \(X\), define their total variation distance \(d_{TV}(\mu, \nu)\) as

\[
d_{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|
\]

where the supremum is taken over all measurable \(A \subseteq X\).

We write \(P^n(x, A)\) for the \(n\)-step transition law for the Markov chain, i.e., \(P^n(x, A) = \mathbb{P}(X_n \in A | X_0 = x)\). 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 \(\pi\) is said to be ergodic if for any \(x \in X\) we have

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

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

\[
d_{TV}(P^n(x, \cdot), \pi) \leq C(x)c^n
\]

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 \(c\), 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 \(\sigma^2\); 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 \(\pi\), and if for some \(c > 0\) the function \(h : X \to \mathbb{R}\) satisfies \(\mathbb{E}[(h(X_0))^2] < \infty\), then there exists a \(c'\) such that the normalized sum

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

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

It is known under certain additional assumptions that for asymptotic normality, the condition \(\mathbb{E}[(h(X_0))^2] < \infty\) can be weakened to just a finite second moment: \(\mathbb{E}[h(X_0)^2] < \infty\). In particular, this is true if geometric ergodicity is strengthen to uniform ergodicity, as shown by Glynn [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 \(\pi\), and a function \(h : X \to \mathbb{R}\) satisfying \(\mathbb{E}[h^2] < \infty\), such that the following holds. For no choice of \(c'\) does

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

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 any state space is needed; in the example \(X\) will in fact be countable.

The rest of the paper is devoted to proving Theorem 1.2. 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 \(X\) on which the Markov chain will be living. Let \(\mathbb{X}\) denote the set of all integer triples \((a,b,c)\) such that \(a \geq \frac{1}{2}, b \in \{1, \ldots, a\}\) and \(c \in \{-1,1\}\), and let \(X = [0, \infty) \cup \mathbb{X}\). For any \(x \in X\), define \(a(x), b(x)\) and \(c(x)\) as

\[
a(x) = \begin{cases} 0 & \text{if } x = 0 \\ a & \text{if } (a,b,c) \in X, \end{cases}
\]

\[
b(x) = \begin{cases} 0 & \text{if } x = 0 \\ b & \text{if } (a,b,c) \in X, \end{cases}
\]

(2)
and
\[ \gamma(x) = \begin{cases} 0 & \text{if } x = 0 \\ c & \text{if } x = (a,b,c) \in \mathcal{X}. \end{cases} \] (3)

The dynamics of the Markov chain \( X_0, X_1, \ldots \), is as follows. It is only at the times when \( X_t = 0 \) that there is any actual randomness in the choice of the next state \( X_{t+1} \). If \( X_t \) is in state \((a,b,c) \in \mathcal{X}\), then the chain moves with probability \( 1/2 \) to state
\[ \left\{ \begin{array}{ll} 0 & \text{if } a = b \\ (a,b-1,c) & \text{otherwise}. \end{array} \right. \] (4)

If, on the other hand, \( X_t = 0 \), then the next state is chosen from \( X \) according to
\[ X_{t+1} = \begin{cases} 0 \text{ with probability } \frac{1}{2} \\ (a,b,c) \text{ with probability } \frac{1}{4} (a,b-1) \text{ if } a = b \\ 0 \text{ otherwise.} \end{cases} \] (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 \( P(Y_t = 0) = P(Y_t = 1) = 1/2 \). 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 \( X_t \)’s, taking value 1 or \( 0 \) 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 that the chain is irreducible and aperiodic, and has a stationary distribution \( \pi \) given by
\[ \pi(0) = \frac{1}{2} \]
and, for any \((a,b,c) \in \mathcal{X}\),
\[ \pi((a,b,c)) = 2 \ \text{(as)} \]
In order for this construction to be used as a counterexample in Theorem 1.3, we need to prove the following,

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

**Proof:** Pick any state \( x \in X \), 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 \( X_0^{(b)} \), \( X_1^{(b)} \), \ldots \), with the same transition kernel, but with \( X_0 \) chosen according to \( x \). Then \( X_t \) will have distribution \( \pi \) for any \( t \), and it follows by the usual coupling inequality that for any \( n \) we have
\[ \text{d}_{TV}(P^n(x_i), \pi) \leq P(X_n \neq X_n^{(b)}) \leq \text{d}_{TV}(P^n(x_i), \pi), \] (6)

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

For any fixed \( x \in X \), there exists a determinate 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 \( k = \beta(x) \). Hence \( \beta(X_0) \) may be interpreted as the waiting time from time \( i \) until the chain hits 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 \( X_0 = 0 \), and by choosing \( \beta \) with respect to \( P(0, \cdot) \) \( \text{(i.e., the transition probabilities indicated in (5))} \), we get
\[ \beta(X_{k+1}) = \beta = 2 \ \text{for } i = 0, 1, 2, \ldots. \] (7)

Furthermore, \( X_{k+1} \) has distribution \( \pi \), and choosing \( \beta \) with respect to \( \pi \) yields that \( \beta(X_{k+1}) \) has the same distribution \( \pi \) as \( \beta(X_{k+1}) \). We are therefore free to pick \( X_{k+1} \), and \( X_{k+1} \) is such a way that \( P(\beta(X_{k+1}) = \beta(X_{k+1})) = 1 \), and using this, (7) formally implies that the two chains will coalesce deterministically until and including time \( k + 1 + \beta(X_{k+1}) \) if they are both forced to take value 0 from that time and on we can generate the \( X_n \) chain and the \( X_n^{(b)} \) chain by letting them make identical moves according to \( P \). This defines the coupling, which for any \( n \) has the property that
\[ P(X_n \neq X_n^{(b)}) \leq P(n + 1 + \beta(X_{k+1})) = \beta(X_{k+1}) \]
\[ = \beta(X_{k+1}) \leq n \leq 1 \]
\[ \leq k \]
\[ \leq \left( \frac{1}{2} \right)^n \text{ for } n < k \]
\[ \leq \left( \frac{1}{2} \right) \text{ for } n = k \]

which means that the chain is geometrically ergodic with \( \rho = \frac{1}{2} \) and \( C(z) = \rho \), \( z = z \).

**Remark.** Readers interested in another type 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 as indicated in (5) in that it is taken to be identical in all corresponding \( X_t \)’s, taking value 1 or 0 with probability \( 1/2 \) each, independently for separate runs.

3 The function \( h \)

The choice of the function \( h : X \to \mathbb{R} \) will be made with the spectrum of making the partial means \( \sum_{i=0}^{n} X_i \) 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 normalized random variables with the property that there exist arbitrarily large \( n \) such that for some normalizing constants \( a_n \), we have

\[
P \left( 1.001 \leq \frac{Z_n}{a_n} \leq 0.999 \right) \geq 0.1
\]

and

\[
P \left( 0.999 \leq \frac{Z_n}{a_n} \leq 1.001 \right) \geq 0.1.
\]

The \( a_n \) for no choice of \( \mu_1, \mu_2, \ldots \), and \( \sigma_1, \sigma_2, \ldots \), does \( \frac{Z_n}{a_n} \) 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 \( \gamma(x) \) and \( \gamma(x) \), and let

\[
h(x) = \left\{ \begin{array}{ll}
\frac{B_k 2^{-k} \gamma(x)}{A_k} & \text{if } \alpha(x) = A_k \text{ for some } k \\
0 & \text{otherwise},
\end{array} \right.
\]

We also define a kind of truncation of \( h \) by setting

\[
h_m(x) = \left\{ \begin{array}{ll}
\frac{B_k 2^{-k} \gamma(x)}{A_k} & \text{if } \alpha(x) = A_k \text{ for some } k \leq m \\
0 & \text{otherwise},
\end{array} \right.
\]

Note that under \( \gamma(x) \) equals \(-1+1\) 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 \( \sigma_m \) such that

\[
\frac{1}{\sqrt{2\pi}} \sum_{i=1}^m h_m(X_i) \tag{8}
\]

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 \( A_1, \ldots, A_k \); as well as \( B_1, \ldots, B_k \); are specified; then we also know the truncated function \( h_{k-1} \) and the variance \( \sigma_k^2 \), 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 to that the following conditions hold.

(i) \( B_k \geq 3\sigma_k \).

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

\[
P \left( \sqrt{2\pi} \sum_{i=1}^k \frac{1}{2^{-k/2}} h_k \mid (X_i) \in \{ 3A_k \} \right) \leq 0.99
\]

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

(iv) \( A_k \geq A_k + 10 \).

(That (iv) can be ensured by picking \( A_k \) large in\( \epsilon \), of course, due to the fact that \( \frac{2^k}{2^{k/2}} e^{-2^k/2} > 0.99 \). 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 \( \pi \).

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

**Proof:** For \( k = 1 \) we have that

\[
\pi \left( \{ x \in X : \alpha(x) = A_k \} \right) \left( \frac{B_k 2^{-k + 1}}{A_k} \right)^2 = \frac{1}{8} (2^k)^2 = 1
\]

and a further direct calculation gives

\[
\pi(h^2) = \sum_{k=1}^\infty \pi \left( \{ x \in X : \alpha(x) = A_k \} \right) \left( \frac{B_k 2^{-k + 1}}{A_k} \right)^2
\]

\[
= 1 + \sum_{k=2}^\infty \pi \left( \{ x \in X : \alpha(x) = A_k \} \right) \left( \frac{B_k 2^{-k + 1}}{A_k} \right)^2
\]

\[
= 1 + \sum_{k=2}^\infty A_k 2^{-k (A_k + 2)} \left( \frac{B_k 2^{-k + 1}}{A_k} \right)^2
\]

\[
\leq 1 + \sum_{k=2}^\infty 2^{-k k}
\]

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 \( C_n = 2^{-k/2} \).

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

\[
P \left( \frac{Z_n}{B_k \sqrt{C_k}} \leq -0.99 \right) \geq \pi \geq 0.1
\]

and

\[
P \left( 0.99 \leq \frac{Z_n}{B_k \sqrt{C_k}} \leq 1.001 \right) \geq \pi \geq 0.1
\]

**Proof:** Without loss of generality, we may assume that the chain \( X_0, X_1, \ldots \) is obtained from the base index \( i \in B \), sequence \( Y \), \( Y_0, Y_1, \ldots \), as in Section 2. Define events \( E_k, E_k^*, E_k^* \) and \( E_k^* \) as follows.

...
Let $E_k^1$ be the event that the sequence $(Y_1, \ldots, Y_n)$ is not intersected by any run of 1's of length $A_k$ or more. By condition (iv), $P(E_k^1)$ has probability at least $1 - 2 \cdot 2^{-k} = 1 - 1/2^k$.

Let $E_k^2$ be the event that the sequence $(Y_1, \ldots, Y_n)$ contains exactly one run of 1's (from the 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 $P(E_k^2) \to e^{-1} = 0.368$ as $k \to \infty$.

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

Let $E_k^4$ be the event that

$$-0.001 \leq \frac{1}{D_k \sqrt{C_k}} \sum_{i=1}^{C_k} h_k(X_i) \leq 0.001,$$

By condition (ii), we have that

$$P(-3 \leq \frac{1}{\sigma_k} \sum_{i=1}^{C_k} h_k(X_i) \leq 3) \geq 0.99,$$

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

Finally, define the event $E_k = E_k^1 \cap E_k^2 \cap E_k^3 \cap E_k^4$. Bonferroni's inequality gives that

$$\lim_{k \to \infty} P(E_k) \geq e^{-1} \cdot 0.99 \cdot 0.368 > 0.2,$$

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

$$0.999 \leq \left| \frac{Z_k}{D_k \sqrt{C_k}} \right| \leq 1.001,$$

Conditional on $E_k$, we have by symmetry that $Z_k$ 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 3.4 implies that the sums $\sum_{i=1}^{k} h(X_i)$ are asymptotically normal. Hence the theorem is established.

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

References


Dept of Mathematics
Chalmers University of Technology
412 96 Gothenburg
Sweden

olle@math.chalmers.se

http://www.math.chalmers.se/~olle/