*J. Appl. Prob.* **43,** 32–47 (2006)
*Printed in Israel*


# A CENTRAL LIMIT THEOREM FOR NON-OVERLAPPING RETURN TIMES

OLIVER JOHNSON,* *University of Cambridge*

**Abstract**

Define the non-overlapping return time of a block of a random process to be the number of blocks that pass by before the block in question reappears. We prove a central limit theorem based on these return times. This result has applications to entropy estimation, and to the problem of determining if digits have come from an independent, equidistributed sequence. In the case of an equidistributed sequence, we use an argument based on negative association to prove convergence under conditions weaker than those required in the general case.

*Keywords:* Central limit theorem; entropy estimation; match length; negative association; return time

2000 Mathematics Subject Classification: Primary 94A17
Secondary 60F05; 62H20

## 1. Introduction and main theorem

### 1.1. Statement of problem

Given a sample $(Z_1, \ldots, Z_n)$ from a random process taking values in an alphabet $\mathcal{A}$, we would like to estimate the entropy of the process. In general this is a hard problem, though if the process is assumed to be independent or stationary some progress can be made.

In particular, given a sequence of binary bits, determining whether the bits were generated by an independent equidistributed process has applications to problems in cryptography and number theory, as described in Section 1.3.

Our approach is as follows. We first partition the sample of $(Z_i)$ into blocks of size $\ell$. That is, writing $Z_a^b$ for $(Z_a, Z_{a+1}, \ldots, Z_b)$, we define block random variables $X_i = Z_{(i-1)\ell+1}^{i\ell}$ such that $X_i \in \mathcal{A}^\ell$ for each $i$. Then, given the first $k$ blocks $X_1, \ldots, X_k$, we count how long it takes for these blocks to reappear.

**Definition 1.1.** (*Non-overlapping return time.*) For a given $k$, define the random variable

$$S_j = \min\{t \geq 1 \colon X_{j+t} = X_j\}, \qquad j = 1, \ldots, k,$$

to be the return time of the $j$th block.

It appears that this definition dates back to Maurer [17]. The main result of this paper is that if the number and size of blocks grow appropriately, then the $S_j$ satisfy a central limit theorem.

**Theorem 1.1.** *Suppose that $(Z_i)$ is an independent, identically distributed finite-alphabet process with entropy $H$. Write $q_{\max} < 1$ for the maximum probability of any symbol appearing*

Received 14 June 2005; revision received 11 November 2005.
* Postal address: Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. Email address: otj1000@cam.ac.uk

*in* $(Z_i)$. *If as* $\ell \to \infty$ *the number of blocks of length* $l$, $k(\ell)$, *goes to* $\infty$ *in such a way that* $\lim_{\ell\to\infty} k(\ell)^{3/2}\ell q_{\max}^{\ell} = 0$, *then*

$$\frac{\sum_{i=1}^{k(\ell)}(\log S_i - \ell H \log 2 + \gamma)}{\sqrt{k(\ell)\pi^2/6}} \xrightarrow{\mathrm{D}} N(0,1).$$

*Here* '$\xrightarrow{\mathrm{D}}$' *denotes convergence in distribution and* $\gamma$ *is Euler's constant,* $\gamma = 0.577\,216\cdots$.

**Remark 1.1.** 1. Note that, to agree with conventions in information theory, the entropy $H$ is calculated using logarithms to the base 2. If entropy were calculated using natural logarithms, the log 2 term could be omitted.

2. For equidistributed processes, in which each symbol occurs with probability $q$, say, Proposition 3.2, below, shows that we can relax the assumption on the rate of convergence of $k(\ell)$, to only require that $k(\ell)\ell q^{\ell} \to 0$. Indeed, simulations in Section 4 suggest that a weaker condition, namely that $k(\ell)\ell 2^{-H\ell} \to 0$, may be sufficient to ensure convergence for all finite-alphabet independent processes with entropy $H$. The faster rate of convergence in the case of equidistribution is useful, since we will often test a null hypothesis of equidistribution.

3. Maurer [17] proved that $\lim_{\ell\to\infty} \log S_1 - \ell H = -\gamma$ for equidistributed binary processes. This result was extended to the case of stationary $\psi$-mixing processes by Abadi and Galves [2].

**Corollary 1.1.** *Under the conditions of Theorem 1.1, the estimator*

$$\hat{H} = \frac{\sum_{i=1}^{k(\ell)}(\log S_i + \gamma)}{\ell \log 2\sqrt{k(\ell)\pi^2/6}}$$

*satisfies* $\hat{H} \sim AN(H, 1/l^2)$ *and, so, is asymptotically normal and consistently estimates the entropy.*

If the process is independent and equidistributed on a finite alphabet, then each block occurs at each time with probability $p = |\mathcal{A}|^{-\ell}$. Hence, each of the $S_j$ is a geometric random variable, with $\mathrm{P}(S_j = r) = p(1-p)^{r-1}$ (we call this a geom($p$) variable). In general, if the process is independent and identically distributed with entropy $H$, it satisfies the asymptotic equipartition property, meaning that, asymptotically, almost exactly $2^{H\ell}$ blocks of length $\ell$ appear, each with a probability of almost exactly $2^{-H\ell}$. Hence, conditioned on the value of $X_j$, $S_j$ is a geometric random variable. However, even though the symbols $Z_i$ are independent, the return times $S_j$ are dependent, so we need to understand the dependence structure to prove Theorem 1.1.

In Section 1.2 we describe some results concerning similar return time definitions made by other authors. In Section 1.3 we describe two possible applications of these results. In Section 2 we prove Theorem 1.1, using an argument based on asymptotic independence. We transform our problem into a similar one, by eliminating the possibility of early matches. In Section 3 we give a proof, under weaker conditions, for equidistributed random variables, by using negative association. Section 4 contains the results of some simulations.

In future work, we hope to extend these results to general stationary processes, under a suitable mixing condition, and to prove similar results for other definitions of match length. Notice that, as mentioned previously, Abadi and Galves [2] proved convergence of $\log S_i - \ell H$ for stationary $\psi$-mixing processes, and that exponential bounds for individual hitting times were proved under similar conditions for random processes, by Abadi [1], and for Gibbsian random fields, by Abadi *et al.* [3].

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

### 1.2. Other, similar definitions

We briefly describe some other work concerning similar quantities. This is by no means an exhaustive list, but merely gives a flavour of some of the alternative approaches which exist.

1.2.1. *Overlapping return time.*

**Definition 1.2.** (*Overlapping return time.*) Define the random variable

$$T_k = \min\{t \geq 1 \colon Z_1^k = Z_{t+1}^{t+k}\}$$

to be the time elapsed before the block $Z_1^k$ is next seen.

Kac's lemma [13] shows that, for any stationary ergodic process,

$$\mathrm{E}[T_k \mid Z_1^k = z_1^k] = \frac{1}{\mathrm{P}(Z_1^k = z_1^k)}.$$

This was developed by Kim [14], who gave the limiting behaviour of $\mathrm{E}[T_k \mathrm{P}(Z_1^k)]$ for independent processes, and by Wyner [25], who proved the following result.

**Theorem 1.2.** *If $(Z_i)$ is a Markov chain then*

$$\lim_{n\to\infty} \frac{\log_2 T_n - nH}{\sqrt{n\mathcal{V}}} \sim N(0, 1).$$

Here $\mathcal{V} = \lim_{n\to\infty} \mathrm{var}(-\log_2 \mathrm{P}(Z_1^n))/n$ is the information variance. For independent processes,

$$\mathcal{V} = \mathrm{var}(-\log_2 \mathrm{P}(Z_1)) = \sum_i \mathrm{P}(Z_1 = z_i)(-\log_2 \mathrm{P}(Z_1 = z_i) - H)^2.$$

We refer the reader also to Corollary 2 of [15], where it was shown that this result holds for general stationary processes $(Z_i)$ under explicit mixing conditions. Wyner and Ziv [24], Ornstein and Weiss [19], and Gao [8, pp. 46–57] studied similar quantities.

1.2.2. *Grassberger prefix.*

**Definition 1.3.** (*Grassberger prefix.*) Given an $n$ and an $i$, $1 \leq i \leq n$, define

$$R_{i,n}(Z_1^n) = \inf\{t \colon Z_i^{i+t-1} \neq Z_j^{j+t-1} \text{ for all } j \neq i\}.$$

In words, $R_{i,n}(Z_1^n)$ is the length of the shortest string starting at position $i$ that is different from all the others of equal length starting at position $j$, $1 \leq j \leq n$. This quantity was introduced by Grassberger [9], and studied by Kontoyiannis and Suhov [16], Quas [21] and Shields [22], [23], partly because it allows good entropy estimation for an ergodic process with a suitable degree of mixing. For example, Theorem 1 of [16] is as follows.

**Theorem 1.3.** *If the finite-alphabet process $(Z_i)$ is ergodic with entropy $H$ and satisfies a Doeblin condition, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\log n}{R_{i,n}(Z_1^n)} = H \quad \textit{almost surely.}$$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

1.2.3. *Lempel–Ziv coding.* Another problem with similar features is that of finding the asymptotic behaviour of the number of codewords in the Lempel–Ziv parsing (see Section 12.10 of [7]). Ziv [26] made a conjecture concerning the number of codewords. However, Aldous and Shields [4] were only able to resolve the problem for independent, identically equidistributed processes, and it took careful analysis by Jacquet and Szpankowski [11] to extend their results to independent, identically distributed asymmetric processes. For example, Theorem 1A of [11] is as follows.

**Theorem 1.4.** *Given a binary asymmetric* ($\mathrm{P}(Z_1 = 0) \neq \frac{1}{2}$)*, independent, identically distributed process, the total length, $L_m$, of the $m$ words in a Lempel–Ziv tree satisfies*

$$\lim_{m\to\infty} \frac{L_m - \mathrm{E}[L_m]}{\sqrt{\mathrm{var}(L_m)}} \sim N(0, 1),$$

*where*

$$\mathrm{E}[L_m] = \frac{m \log_2 m}{H} + O(m) \quad \textit{and} \quad \mathrm{var}(L_m) = \frac{\mathcal{V} m \log_2 m}{H^3} + O(m).$$

1.2.4. *Comparison of approaches.* Notice that Theorems 1.2, 1.3, and 1.4 differ in character from our Theorem 1.1. For example, Theorem 1.3 proves a law of large numbers for the Grassberger prefixes, showing that a statistic based on them acts as an entropy estimator. However, it does not tell us the rate of convergence of the estimator. Similarly, although Theorems 1.2 and 1.4 give asymptotic normality, they both refer to statistics calculated with respect to one fixed point. It is possible that this fixed point could be unrepresentative; hence, our result is stronger, in the sense that it averages over a number of different starting points.

### 1.3. Applications

We briefly describe two applications of these results.

1. *Cryptography.* The problem of deciding whether binary bits $(Z_i)$ were generated by an independent, identically equidistributed process arises in cryptography. Bits generated in this way can be used as a perfectly secure one-time pad to transmit a message $(Y_i)$. This system is secure, in the sense that the transmitted bits $Y_i \oplus Z_i$ are independent of the message, meaning that no inference about $Y_i$ can be made from them. Equivalently, Shannon's second coding theorem (see, for example, Theorem 8.7.1 of [7]) implies that the binary symmetric channel with error probability $p = \frac{1}{2}$ has capacity $C = 0$. If the $(Z_i)$ were not independent and equidistributed, then, given large enough $n$, it may be possible to infer properties of the $(Z_i)$ and perhaps read the message $(Y_i)$.

2. *Number theory.* Recall that a number is said to be normal to base $b$ if the limiting proportion of each digit in its base-$b$ expansion is $1/b$. A number which is normal to *all* bases $b$ is simply referred to as being normal. Ergodic theory shows that almost all numbers are normal, but it is hard to prove that any particular number has this property. For example, Bailey and Crandall [5] proved that a particular class of numbers (including the so-called Stoneham and Korobov numbers) has the normal property. On the other hand, in the same paper, [5], the authors discussed the fact that constants including $\pi$, e, $\ln 2$, and $\zeta(3)$ are not known to be normal. Weisstein gives a review of results concerning normal numbers that is available online at http://mathworld.wolfram.com/NormalNumber.html. An informal statement of the property of normality to base $b$ is that the digits of the number 'look as if they were generated by an independent, identically equidistributed process', which we hope to be able to test.

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

Kim [14] gave computational results concerning the speed of convergence of estimators based on overlapping matches, hoping to detect processes which are not Bernoulli. Similarly, Bradley and Suhov [6] used theoretical results concerning the Grassberger prefixes (see Definition 1.3) to consider the normality of constants such as $\pi$, e, and $\gamma$. We give some computational results in Section 4.

## 2. Proof of main theorem

### 2.1. Avoiding early matches

The difficulty in analysing the dependence structure of the random variables $S_i$ introduced in Definition 1.1 is that 'early matches' can occur at $i$. That is, it may be that $S_i \leq k - i$. The possibility of early matches leads to a complicated situation of case splitting according to where such early matches occur. To avoid this, we introduce a very similar sequence of random variables, $(R_j)$, in Definition 2.1 below. We use two main ideas to prove Theorem 1.1, the central limit theorem for the $S_i$.

First, we let $S_i = R_i + (k - i)$, unless there has been an early match. By controlling the probability of an early match, we show that a suitably scaled version of $\sum_i \log S_i - \sum_i \log R_i$ tends to 0 in probability; thus, a limit law for the $R_i$ passes over to a limit law for the $S_i$. The formal statement is given in Lemma 2.1, below. Then, in Proposition 2.1, we establish a central limit theorem for the $R_i$ using explicit bounds on conditional probabilities, which show that the variables are asymptotically independent.

**Definition 2.1.** Given a realisation of $X_1, \ldots, X_k$, we define $D$ to be the set of positions that do not see an early match, i.e. $D = \{i : X_i \neq X_j,\ j = i+1, \ldots, k\}$. For each $i \in D$, we let $b_i = X_i$. For each $i \notin D$, we let $b_i$ be a random element chosen uniformly from the set of elements not yet seen, namely $\mathcal{A}^\ell \setminus \bigcup_i b_i$.

Define the random variable $R_j$ to be the time elapsed between time $k$ and the time of first appearance of value $b_j$, i.e. $R_j = \min\{t \geq 1 : X_{k+t} = b_j\},\ j = 1, \ldots, k$.

**Lemma 2.1.** *Suppose that $(Z_i)$ is an independent, identically distributed finite-alphabet process with entropy $H$. If, as $\ell \to \infty$, $k(\ell) \to \infty$ in such a way that $\lim_{\ell\to\infty} k(\ell)^{3/2} \ell q_{\max}^\ell = 0$, then the difference term*

$$\frac{\sum_{i=1}^k (\log R_i - \log S_i)}{\sqrt{k}}$$

*tends to 0 in probability.*

*Proof.* The key observation is that $S_i = R_i + (k - i)$ unless $i \in D^c$. Furthermore, $1 \leq S_i \leq R_i + (k - i)$. This means that we can decompose as follows, since $R \sim \text{geom}(p)$ with $\mathrm{E}[1/R] = -p \log p/(1-p) = O(q_{\max}^\ell \ell)$:

$$\begin{aligned}
\mathrm{P}\left( \frac{\sum_{i=1}^k (\log S_i - \log R_i)}{\sqrt{k}} \geq \delta \right) &\leq \mathrm{P}\left( \frac{\sum_{i=1}^k \log(1 + (k-i)/R_i)}{\sqrt{k}} \geq \delta \right) \\
&\leq \mathrm{P}\left( \sum_{i=1}^k \sqrt{k}\, \frac{1}{R_i} \geq \delta \right) \\
&\leq \frac{\sqrt{k}}{\delta} \sum_{i=1}^k \mathrm{E}[1/R_i] = O(k(\ell)^{3/2} \ell q_{\max}^\ell).
\end{aligned}$$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

Similarly, we know that

$$\begin{aligned}\mathrm{P}\left(\frac{\sum_{i=1}^{k}(\log R_i - \log S_i)}{\sqrt{k}} \geq \delta\right) &\leq \mathrm{P}\left(\frac{\sum_{i=1}^{k}\log R_i\, 1(i \in D^{\mathrm{c}})}{\sqrt{k}} \geq \delta\right)\\ &\leq \frac{1}{\delta\sqrt{k}}\left(\sum_{i=1}^{k}\mathrm{E}[\log R_i]\,\mathrm{P}(i \in D^{\mathrm{c}})\right)\\ &= O(k(\ell)^{3/2}\ell q_{\max}^{\ell}),\end{aligned}$$

since

$$\mathrm{P}(i \in D^{\mathrm{c}}) = 1 - \mathrm{P}(i \in D) \leq (1 - (1 - q_{\max}^{\ell})^{k}) \leq k q_{\max}^{\ell},$$

independently of $R_i$, and $\mathrm{E}[\log R_i] = O(\ell)$.

### 2.2. Moments of $\log R_i$

We first find the leading-order terms in $\mathrm{E}[\log R_i]$ and $\mathrm{var}(\log R_i)$ for all $i$. The values $\gamma$ and $\pi^2/6$, which appear in Lemma 2.3, were first recognised by Maurer [17]; in addition to these, we need to know the order of the error term as well. We use the following lemma, the simplest form of the Euler–Maclaurin sum formula.

**Lemma 2.2.** *For any differentiable function $f$ such that $f(x) \to 0$ as $x \to \infty$, we have*

$$\left|\sum_{i=1}^{\infty} f(i) - \int_{1}^{\infty} f(x)\,\mathrm{d}x\right| \leq \frac{1}{2}|f(1)| + \frac{1}{2}\int_{1}^{\infty}|f'(x)|\,\mathrm{d}x. \tag{2.1}$$

*Proof.* Note that (by integrating by parts), for all differentiable functions $f$ and all $a$,

$$\int_{a}^{a+1} f(x)\,\mathrm{d}x = \frac{1}{2}(f(a) + f(a+1)) - \int_{a}^{a+1} f'(x)\left(x - a - \frac{1}{2}\right)\mathrm{d}x,$$

which implies that

$$\int_{a}^{a+1} f(x)\,\mathrm{d}x - \frac{1}{2}(f(a) + f(a+1)) = R,$$

where $|R| \leq (\int_{a}^{a+1}|f'(x)|\,\mathrm{d}x)/2$. By summing such results from $a = 1$ to $a = \infty$, we deduce that (2.1) holds.

**Lemma 2.3.** *For $R \sim \mathrm{geom}(p)$,*

$$\mu(p) := \mathrm{E}[\log R] = -\gamma - \log p + O(p), \tag{2.2}$$

$$\sigma^2(p) := \mathrm{var}(\log R) = \frac{\pi^2}{6} + O(p\log p), \tag{2.3}$$

$$\mathrm{E}[|\log R - \mu(p)|^3] \leq K, \tag{2.4}$$

*where $\gamma$ is Euler's constant as before, and $K$ is a finite constant.*

*Proof.* Let $c = -\log(1-p)$ and $f(x) = \mathrm{e}^{-cx}\log x$, and use the fact that

$$|f'(x)| = \left|\frac{\mathrm{e}^{-cx}}{x} - c\mathrm{e}^{-cx}\log x\right| \leq \left|\frac{\mathrm{e}^{-cx}}{x}\right| + c|\mathrm{e}^{-cx}\log x|.$$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

We deduce, by (2.1), that the difference between the following integral $I$ and sum $S$ is

$$\begin{aligned}|S-I| &:= \left|\sum_{i=1}^{\infty} \mathrm{e}^{-ci}\log i - \int_1^{\infty} \mathrm{e}^{-cx}\log x\,\mathrm{d}x\right| \\ &\le \frac{1}{2}\int_1^{\infty}\frac{\mathrm{e}^{-cx}}{x}\,\mathrm{d}x + \frac{c}{2}\int_1^{\infty}\mathrm{e}^{-cx}\log x\,\mathrm{d}x \\ &= \left[\frac{1}{2}\mathrm{e}^{-cx}\log x\right]_1^{\infty} + c\int_1^{\infty}\mathrm{e}^{-cx}\log x\,\mathrm{d}x \\ &= cI.\end{aligned}$$

Using the fact that

$$I = \frac{\Gamma(0,c)}{c} = -\gamma - \log c + c + O(c^2),$$

where we write $\Gamma(0,\cdot)$ for the incomplete gamma function, we deduce that

$$\mathrm{E}[\log R] = \sum_{x=1}^{\infty} p(1-p)^{x-1}\log x = \frac{p}{1-p}\frac{\Gamma(0,c)}{c}(1+\varepsilon) = -\gamma - \log p + O(p),$$

where $|\varepsilon| < c$. For $f(x) = \mathrm{e}^{-cx}(\log x)^2$, we have

$$|f'(x)| = \left|\frac{2\mathrm{e}^{-cx}\log x}{x} - c\mathrm{e}^{-cx}(\log x)^2\right| \le \left|\frac{2\mathrm{e}^{-cx}\log x}{x}\right| + c|\mathrm{e}^{-cx}(\log x)^2|.$$

As above, we find that

$$|S-I| := \left|\sum_{i=1}^{\infty}\mathrm{e}^{-ci}\log i - \int_1^{\infty}\mathrm{e}^{-cx}\log x\,\mathrm{d}x\right| \le cI.$$

Using the fact that $I = \pi^2/(6c) + (-\gamma + \log c)^2/c + 1 - O(c)$, we deduce that

$$\mathrm{E}[\log R^2] - \mu(p)^2 = \frac{\pi^2}{6} + O(p\log p).$$

Finally, we bound the centred absolute third moment, $\mathrm{E}[|\log R - \mu(p)|^3]$. We partition the real line into three intervals: the sets $A_1 = \{x\colon |\log x - \mu(p)| \le 1\}$, $A_2 = \{x\colon \log x - \mu(p) \ge 1\}$, and $A_3 = \{x\colon \log x - \mu(p) \le -1\}$. We also define the integrals

$$K_i = \mathrm{E}[|\log R - \mu(p)|^3 1(R \in A_i)], \qquad i = 1,2,3.$$

Clearly $K_1 \le 1$. By Chernoff's bound, for $t \ge 1$, we have

$$\begin{aligned}\mathrm{P}(\log R - \mu(p) \ge t) &\le \frac{\mathrm{E}[R^s]}{\exp(s(t+\mu(p)))} \\ &\le \mathrm{e}^{-2t}\frac{(2-p)/p^2}{\exp(-2\gamma + O(p))/p^2} \\ &\le 2\mathrm{e}^{2\gamma}\mathrm{e}^{-2t},\end{aligned}$$

taking $s = 2$. Hence, $K_2 = \int_1^{\infty} 3t^2\,\mathrm{P}(\log R - \mu(p) \ge t)\,\mathrm{d}t \le 4$. Similarly, we have

$$\mathrm{P}(\log R - \mu(p) \le -t) = \mathrm{P}(R \le \mathrm{e}^{-\gamma - t}) \le 1 - \exp(-\mathrm{e}^{-\gamma - t}),$$

meaning that $K_3 \le 4$. Thus, we can take $K = 9$.

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

### 2.3. Asymptotic independence

Next we prove a lemma that shows that the $R_i$ are approximately independent, giving explicit bounds on the difference between the joint probability distribution and the product of the marginals.

**Lemma 2.4.** *Suppose that $(Z_i)$ is an independent, identically distributed finite-alphabet process with entropy $H$. For $(R_i)$ as defined in Definition 2.1, and for any $s, m$, and $a = (a_1, \dots, a_{m-1})$, with $R = (R_1, \dots, R_{m-1})$ we have*

$$\left(1 - \frac{p_m}{1 - W^*}\right)^{s-1} \le \mathrm{P}(R_m \ge s \mid R = a) \le (1 - p_m)^{s-m},$$

*where $W^* = p_1 + \cdots + p_{m-1}$ and $p_i = \mathrm{P}(b_i)$.*

*Proof.* Given the values of $R_1, \dots, R_{m-1}$, we can write down an explicit expression for the distribution of $R_m$:

$$\mathrm{P}(R_m \ge s \mid R = a) = \prod_{i=1}^{s-1} \mathrm{P}(X_{k+i} \ne b_m \mid R = a). \tag{2.5}$$

We consider this product term by term, for each value of $i$. If $a_j = i$ for some $j \le m - 1$, then $X_{k+i} = b_j$ and, automatically, $X_{k+i} \ne b_m$; hence, the contribution to (2.5) from that $i$-value is 1. Otherwise, if $a_j \ne i$ for all $j \le m - 1$, then

$$\mathrm{P}(X_{k+i} \ne b_m \mid R = a) = 1 - \mathrm{P}(X_{k+i} = b_m \mid R = a) = 1 - \frac{p_m}{1 - W_i},$$

where $W_i = \sum_{j=1}^{m-1} p_j 1(a_j > i)$. This is a decreasing function of $W_i$.

It is clear that the product in (2.5) is maximised when the first $(m - 1)$ values of $a_i$ occur in the first $m - 1$ places, that is, when $\{a_1, \dots, a_{m-1}\} = \{1, \dots, m - 1\}$. In this case, the product in (2.5) equals $(1 - p_m)^{s-m}$. Similarly, the product in (2.5) is minimised when $W_i$ is maximised for each $i$, that is, when $a_j \ge s$ for each $j$. In this case, $W_i = \sum_{j=1}^{m-1} p_j = W^*$ for each $i$, and the product equals $(1 - p_m/(1 - W^*))^{s-1}$.

**Proposition 2.1.** *Suppose that $(Z_i)$ is an independent, identically distributed finite-alphabet process with entropy $H$. For $(R_i)$ as defined in Definition 2.1, we find that, for any $\theta$,*

$$\left|\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}} \sum_{j=1}^{m} \log R_j\right)\right] - \mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}} \sum_{j=1}^{m-1} \log R_j\right)\right] \mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}} \log R_m\right)\right]\right|$$

*is $O(\ell k(\ell)^{1/2} q_{\max}^{\ell})$.*

*Proof.* By adapting Equation (22) of [18], for any complex, continuously differentiable functions $f$ and $g$, and for random variables $U$ and $V$, we have

$$\mathrm{cov}(f(U), g(V)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f'(u) g'(v) H_{U,V}(u, v)\, \mathrm{d}u\, \mathrm{d}v, \tag{2.6}$$

where

$$\begin{aligned} H_{U,V}(u, v) &= \mathrm{P}(U \ge u, V \ge v) - \mathrm{P}(U \ge u)\, \mathrm{P}(V \ge v) \\ &= -\mathrm{P}(U \ge u, V < v) + \mathrm{P}(U \ge u)\, \mathrm{P}(V < v). \end{aligned}$$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

Let $U = \log R_m$ and $V = \sum_{i=1}^{m-1} \log R_i$. We will find nonnegative functions $h_-$ and $h_+$ such that $-h_-(u,v) \le H_{U,V}(u,v) \le h_+(u,v)$ for all $u$ and $v$. Since $U$ and $V$ take nonnegative values only, if $u < 0$ or $v < 0$ then $H_{U,V}(u,v) = 0$. Hence, with $f(t) = g(t) = \exp(\mathrm{i}\theta t/\sqrt{k})$, (2.6) simplifies to

$$\begin{aligned}
&\left|\operatorname{cov}\left(\exp\left(\frac{\mathrm{i}\theta U}{\sqrt{k}}\right), \exp\left(\frac{\mathrm{i}\theta V}{\sqrt{k}}\right)\right)\right| \\
&\quad = \left|-\frac{\theta^2}{k}\int_0^\infty\int_0^\infty \exp\left(\frac{\mathrm{i}\theta u}{\sqrt{k}}\right)\exp\left(\frac{\mathrm{i}\theta v}{\sqrt{k}}\right) H_{U,V}(u,v)\,\mathrm{d}u\,\mathrm{d}v\right| \\
&\quad \le \left|\frac{\theta^2}{k}\right|\int_0^\infty\int_0^\infty |H_{U,V}(u,v)|\,\mathrm{d}v\,\mathrm{d}u \\
&\quad \le \left|\frac{\theta^2}{k}\right|\left(\int_0^\infty\int_0^\infty h_-(u,v)\,\mathrm{d}v\,\mathrm{d}u + \int_0^\infty\int_0^\infty h_+(u,v)\,\mathrm{d}v\,\mathrm{d}u\right). \qquad (2.7)
\end{aligned}$$

We know that $\int_0^\infty \mathrm{P}(U \ge u)\,\mathrm{d}u = \mathrm{E}[U]$, $\int_0^\infty \mathrm{P}(V \ge v)\,\mathrm{d}v = \mathrm{E}[V]$, and

$$\int_0^\mu \mathrm{P}(V < v)\,\mathrm{d}v + \int_\mu^\infty \mathrm{P}(V \ge v)\,\mathrm{d}v = \mathrm{E}[|V-\mu|].$$

Furthermore, we can evaluate

$$f(p) = \int_0^\infty (1-p)^{\mathrm{e}^u - 1}\,\mathrm{d}u = \frac{\Gamma(0, -\log(1-p))}{1-p}.$$

Since $f(p)$ has an increasing, but negative, gradient, with

$$-f'(p) \le \frac{1}{-(1-p)\log(1-p)} \le \frac{1}{p(1-p)},$$

we know that $f(p) - f(q) \le (q-p)/(p(1-p))$ for $p \le q$. This means that

$$\int_0^\infty (1-p_m)^{\mathrm{e}^u-1} - \left(1 - \frac{p_m}{1-W^*}\right)^{\mathrm{e}^u-1}\mathrm{d}u = O(k(\ell)q_{\max}^\ell). \qquad (2.8)$$

Let us rearrange the result of Lemma 2.4 and sum over values of $a$ such that $\sum_j \log a_j \ge v$ or $\sum_j \log a_j < v$. For $v \ge \mathrm{E}[V]$, we find that

$$h_-(u,v) \le \left((1-p_m)^{\mathrm{e}^u-1} - \left(1 - \frac{p_m}{1-W^*}\right)^{\mathrm{e}^u-1}\right)\mathrm{P}(V \ge v),$$

and similarly can take $h_+(u,v) \le (1-p_m)^{1-m}\,\mathrm{P}(U \ge u)\,\mathrm{P}(V \ge v)$. Thus, by (2.8), over this region,

$$\begin{aligned}
\int h_-(u,v)\,\mathrm{d}u\,\mathrm{d}v &\le O(k(\ell)q_{\max}^\ell)\,\mathrm{E}[|V-\mu|] = O(k(\ell)^{3/2}q_{\max}^\ell), \\
\int h_+(u,v)\,\mathrm{d}u\,\mathrm{d}v &\le (1-p_m)^{1-m}\,\mathrm{E}[U]\,\mathrm{E}[|V-\mu|] = O(\ell k(\ell)^{3/2}q_{\max}^\ell).
\end{aligned}$$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

For $v \le \mathrm{E}[V]$, we find that

$$h_+(u,v) \le \left((1-p_m)^{\mathrm{e}^u-1} - \left(1-\frac{p_m}{1-S}\right)^{\mathrm{e}^u-1}\right)\mathrm{P}(V \ge v),$$

and similarly can take $h_-(u,v) \le (1-p_m)^{1-m}\,\mathrm{P}(U \ge u)\,\mathrm{P}(V \ge v)$. As before, the integrals satisfy

$$\int h_+(u,v)\,\mathrm{d}u\,\mathrm{d}v = O(k(\ell)^{3/2}q_{\max}^{\ell}) \quad \text{and} \quad \int h_-(u,v)\,\mathrm{d}u\,\mathrm{d}v = O(\ell k(\ell)^{3/2}q_{\max}^{\ell}).$$

Substitution of these expressions into (2.7) yields the result.

### 2.4. Completing the proof of Theorem 1.1

The Lyapunov central limit theorem (see, for example, Theorem 4.9 of [20]) implies that, for independent random variables $Y_1, \dots, Y_k$, where $Y_i$ has mean $\mu_i$, variance $\sigma_i^2$, and finite centred absolute third moment $m_i = \mathrm{E}[|Y_i - \mu_i|^3]$, if

$$\frac{\sum_{i=1}^k m_i}{(\sum_{i=1}^k \sigma^2)^{3/2}} \to 0, \tag{2.9}$$

then

$$\frac{\sum_{i=1}^k (Y_i - \mu_i)}{\sqrt{\mathrm{var}(\sum_{i=1}^k Y_i)}} \xrightarrow{\mathrm{D}} N(0,1). \tag{2.10}$$

*Proof of Theorem 1.1.* Define a sequence of independent random variables $(T_i)$ with $T_i \sim R_i$. Then Lemma 2.3 (in particular (2.3) and (2.4)) shows that the Lyapunov condition (2.9) holds if $Y_i = \log T_i$. Thus, given values $p_1, \dots, p_k$, we have

$$\frac{\sum_{i=1}^k (\log T_j - \mu(p_j))}{\sqrt{k}} \xrightarrow{\mathrm{D}} N(0,v), \tag{2.11}$$

where

$$v = \frac{1}{k}\left(\sum_{i=1}^k \mathrm{var}(\log R_i)^2\right) = \frac{\pi^2}{6} + O(p\log p),$$

and (by the law of large numbers) $\sum_{i=1}^k \mu(p_i) \to k(-\gamma + H\ell\log 2)$.

By repeated use of Proposition 2.1, we then find that

$$\left|\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}}\sum_{j=1}^k(\log R_j - \mu(p_j))\right)\right] - \prod_{j=1}^k \mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}}(\log R_j - \mu(p_j))\right)\right]\right|$$

is $O(\ell k(\ell)^{3/2}q_{\max}^{\ell})$, so if this quantity tends to 0 then the central limit theorem for $\log T_i$, (2.11), carries over to give a central limit theorem for $\log R_i$.

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

## 3. Equidistribution and negative association

In the case of equidistributed random variables, we can establish a central limit theorem under conditions on $k(\ell)$ weaker than those required to prove Theorem 1.1, using negative association. This property captures the sense of dependence whereby one random variable being large forces the others to be smaller. Formally, we make the following definition.

**Definition 3.1.** A collection of real-valued random variables $(U_k)$ is negatively associated if

$$\operatorname{cov}(f_1(U_i, i \in A_1),\, f_2(U_j, j \in A_2)) \leq 0 \tag{3.1}$$

for all increasing functions $f_1$ and $f_2$ that take arguments in disjoint sets of indices $A_1$ and $A_2$.

The negative association property proves useful in many situations, not least since Newman [18] showed that the central limit theorem holds for negatively associated sequences of random variables. Furthermore, if $(U_k)$ forms a negatively associated sequence then, for any increasing function $f$, $(f(U_k))$ also forms a negatively associated sequence.

**Proposition 3.1.** *Suppose that $(Z_i)$ is an independent, equidistributed process with finite alphabet $\mathcal{A}$. The $(R_i)$ introduced in Definition 2.1 are then negatively associated.*

*Proof.* Given the ordering $R_{\tau(1)} < R_{\tau(2)} < \cdots < R_{\tau(k)}$ (for an appropriate permutation, $\tau$), the actual values satisfy $R_{\tau(1)} \sim \text{geom}(kp)$, $R_{\tau(2)} - R_{\tau(1)} \sim \text{geom}((k-1)p)$ (independently), and so on. That is, if we define independent random variables $W_i, i = 1, \ldots, k$, with $W_i \sim \text{geom}((k+1-i)p)$, and define $U_j = \sum_{i=1}^{j} W_i$, then $R_{\tau(i)} = U_i$ or, equivalently, $R_i = U_{\tau^{-1}(i)}$.

As in the proof of Theorem 3.4 of [10], it suffices to show that (3.1) holds for symmetric functions $f_1$ and $f_2$ with $A_1 = \{1, \ldots, p\}$ and $A_2 = \{p+1, \ldots, k\}$. Given such functions, define

$$f_1^*(i_1, \ldots, i_p) = \mathrm{E}[f_1(U_{i_1}, \ldots, U_{i_p})],$$
$$f_2^*(i_{p+1}, \ldots, i_k) = \mathrm{E}[f_2(U_{i_{p+1}}, \ldots, U_{i_k})].$$

If any index $i_l$, $l \in \{1, \ldots, p\}$, increases then (since $U_1 < U_2 < \cdots < U_k$) so does $U_{i_l}$, and, hence (since $f_1$ is increasing), so does $f_1^*$. That is, $f_1^*$ is an increasing function of $\{i_1, \ldots, i_p\}$. Similarly, $f_2^*$ is increasing.

For any permutation $\tau$, we define the increasing functions

$$g_1^*(\tau) = f_1^*(\tau^{-1}(1), \ldots, \tau^{-1}(p)),$$
$$g_2^*(\tau) = f_2^*(\tau^{-1}(p+1), \ldots, \tau^{-1}(k)).$$

Theorem 2.11 of [12] implies that the uniform distribution on the set of permutations is negatively associated, whence

$$\mathrm{E}[g_1^*(\tau) g_2^*(\tau)] \leq \mathrm{E}[g_1^*(\tau)]\,\mathrm{E}[g_2^*(\tau)]. \tag{3.2}$$

Now,

$$\begin{aligned}\mathrm{E}[g_1^*(\tau)] &= \mathrm{E}[f_1^*(\tau^{-1}(1), \ldots, \tau^{-1}(p))] = \mathrm{E}[f_1(U_{\tau^{-1}(1)}, \ldots, U_{\tau^{-1}(p)})] \\ &= \mathrm{E}[f_1(R_1, \ldots, R_p)].\end{aligned}$$

Similarly, $\mathrm{E}[g_2^*(\tau)] = \mathrm{E}[f_2(R_{p+1}, \ldots, R_k)]$ and

$$\mathrm{E}[g_1^*(\tau) g_2^*(\tau)] = \mathrm{E}[f_1(R_1, \ldots, R_p) f_2(R_{p+1}, \ldots, R_k)];$$

thus, (3.2) implies (3.1), as required.

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

**Lemma 3.1.** *Suppose that $(Z_i)$ is an independent, equidistributed process on a finite alphabet $\mathcal{A}$. If, as $\ell \to \infty$, $k(\ell) \to \infty$ in such a way that $k(\ell)\ell|\mathcal{A}|^{-\ell} \to 0$, then the difference term*

$$\frac{\sum_{i=1}^{k}(\log R_i - \log S_i)}{\sqrt{k}}$$

*tends to* 0 *in probability.*

*Proof.* As before, $S_i = R_i + (k - i)$ unless $i \in D^{\mathrm{c}}$. Furthermore, $1 \le S_i \le R_i + (k - i)$. This means that we can decompose as follows:

$$\begin{aligned}|\log R_i - \log S_i| &\le |\log R_i - \log(R_i + (k - i))| + |\log(R_i + (k - i)) - \log S_i| \\ &\le \frac{k}{R_i} + \log(R_i + k)1(i \in D^{\mathrm{c}}).\end{aligned}$$

Since the $R_i$ are negatively associated, so are the $-k/R_i$. Thus, by the Cauchy–Schwarz inequality, recalling that $p = |\mathcal{A}|^{-\ell}$, we have

$$\begin{aligned}\mathrm{E}\left[\left(\sum_{i=1}^{k}(\log R_i - \log S_i)\right)^2\right] &\le 2\sum_{i=1}^{k} k^2\,\mathrm{E}[1/R_i^2] + 2\sum_{i\in D^{\mathrm{c}}}(\mathrm{E}[\log(R_i)^2] + k^2\,\mathrm{E}[1/R_i^2]) \\ &\le 2k^3(p^2 + O(p^3)) + 2kp(O((-\log p)^2) \\ &\quad + k^2(p^2 + O(p^3))). \qquad (3.3)\end{aligned}$$

This follows both because $\mathrm{E}[1/R_i^2] = p\mathrm{Li}_2(p)/(1 - p) = p^2 + O(p^3)$, where $\mathrm{Li}_2(\cdot)$ is the dilogarithm function, and because, for any $i$,

$$\mathrm{P}(i \in D^{\mathrm{c}}) \le \mathrm{P}(1 \in D^{\mathrm{c}}) = 1 - \mathrm{P}(1 \in D) = 1 - (1 - p)^k \le pk,$$

independently of $R_i$. The lemma follows on dividing (3.3) by $k$.

**Lemma 3.2.** *Suppose that $(Z_i)$ is an independent, equidistributed process on a finite alphabet $\mathcal{A}$. For any $i$ and $j$, $i \ne j$, the $R_i$ defined in Definition 2.1 satisfy*

$$|\mathrm{cov}(R_i, R_j)| = O(\ell|\mathcal{A}|^{-\ell}).$$

*Proof.* From the negative association proved in Proposition 3.1, we know that the covariance is negative, so we need only bound it from below. For any $x$, we know that

$$\mathrm{P}(R_j = y \mid R_i = x) = \begin{cases} \dfrac{p}{1-p}\left(\dfrac{1-2p}{1-p}\right)^{y-1} & \text{for } y < x, \\ \left(\dfrac{1-2p}{1-p}\right)^{x-1}(1-p)^{y-x-1}p & \text{for } y > x. \end{cases}$$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

This means that

$$\begin{aligned}
&\mathrm{E}[\log R_j \mid R_i = x] \\
&\quad = \sum_{y=1}^{x-1} \frac{p}{1-p}\left(\frac{1-2p}{1-p}\right)^{y-1} \log y + \sum_{y=x+1}^{\infty} p\left(\frac{1-2p}{1-p}\right)^{x-1} (1-p)^{y-x-1} \log y \\
&\quad = \sum_{y=1}^{\infty} \frac{p}{1-p}\left(\frac{1-2p}{1-p}\right)^{y-1} \log y - \frac{p}{1-p}\left(\frac{1-2p}{1-p}\right)^{x-1} \log x \\
&\qquad + \sum_{y=x+1}^{\infty} \left( p\left(\frac{1-2p}{1-p}\right)^{x-1} (1-p)^{y-x-1} - \frac{p}{1-p}\left(\frac{1-2p}{1-p}\right)^{y-1} \right) \log y. \qquad (3.4)
\end{aligned}$$

In (3.4) each summand is positive, so we can replace $\log y$ by $\log(y-x)$, which is smaller, and write $z = y - x$, to find that (3.4) is greater than

$$\begin{aligned}
&\sum_{z=1}^{\infty} \left( p\left(\frac{1-2p}{1-p}\right)^{x-1} (1-p)^{z-1} - \frac{p}{1-p}\left(\frac{1-2p}{1-p}\right)^{z+x-1} \right) \log z \\
&\qquad = \left(\frac{1-2p}{1-p}\right)^{x-1} \left( \sum_{z=1}^{\infty} \left( p(1-p)^{z-1} - \frac{p}{1-p}\left(\frac{1-2p}{1-p}\right)^{z} \right) \log z \right) \\
&\qquad = \left(\frac{1-2p}{1-p}\right)^{x-1} \left( \mu(p) - \frac{1-2p}{1-p} \mu \frac{p}{1-p} \right).
\end{aligned}$$

Overall then, using the notation of Lemma 2.3, we have

$$\begin{aligned}
&\mathrm{E}[\log R_i \log R_j] \\
&\quad \geq \sum_{x=1}^{\infty} p(1-p)^{x-1} \mu\left(\frac{p}{1-p}\right) \log x - \frac{p}{2(1-p)} \sum_{x=1}^{\infty} 2p(1-2p)^{x-1} (\log x)^2 \\
&\qquad + \frac{1}{2}\left( \sum_{x=1}^{\infty} 2p(1-2p)^{x-1} \log x \right) \left( \mu(p) - \frac{1-2p}{1-p} \mu\left(\frac{p}{1-p}\right) \right) \\
&\quad = \mu(p)\mu \frac{p}{1-p} - \frac{p}{2(1-p)} (\sigma^2(2p) + \mu(2p)^2) \\
&\qquad + \frac{\mu(2p)}{2} \left( \mu(p) - \frac{1-2p}{1-p} \mu \frac{p}{1-p} \right).
\end{aligned}$$

By expanding this using Lemma 2.3, we deduce that $\mathrm{cov}(\log R_i, \log R_j) \geq -p \log p$. Indeed, asymptotically, $\mathrm{cov}(\log R_i, \log R_j) \geq p((-\log p)/2 + \gamma)$.

We can now deduce the central limit theorem for $\log S_i$.

**Proposition 3.2.** *Suppose that $(Z_i)$ is an independent, equidistributed finite-alphabet process with entropy $H$. If, as $\ell \to \infty$, $k(\ell) \to \infty$ in such a way that $k(\ell)\ell |\mathcal{A}|^{-\ell} \to 0$, then*

$$\frac{\sum_{i=1}^{k(\ell)} (\log S_i - \ell H \log 2 + \gamma)}{\sqrt{k(\ell)\pi^2/6}} \xrightarrow{\mathrm{D}} N(0,1).$$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

*Proof.* By Lemma 3.1, we need only prove the corresponding result for $\log R_i$. By Proposition 3.1, the $R_i$ are negatively associated and, hence, so are the $\log R_i$. Since $H(u, v) \le 0$ for all $u$ and $v$, by adapting (2.7) as in [18, cf. Equation (22)], we obtain the following result: if $U_1, \dots, U_k$ are negatively associated then

$$\left|\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}}\sum_{j=1}^{k} U_j\right)\right] - \prod_{j=1}^{k}\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta U_j}{\sqrt{k}}\right)\right]\right| \le \frac{\theta^2}{k}\sum_{i\neq j}|\mathrm{cov}(U_i, U_j)|. \tag{3.5}$$

This means that, with $\mu = \ell H \log 2 - \gamma$, and taking $\varphi$ to be the characteristic function of the distribution $N(0, v)$, we have

$$\begin{aligned}
&\left|\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}}\sum_{i=1}^{k}(\log R_i - \mu)\right)\right] - \varphi(\theta)\right| \\
&\quad\le \left|\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}}\sum_{j=1}^{k}(\log R_j - \mu)\right)\right] - \prod_{j=1}^{k}\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}}(\log R_j - \mu)\right)\right]\right| \\
&\qquad + \left|\prod_{j=1}^{k}\mathrm{E}\left[\exp\left(\frac{\mathrm{i}\theta}{\sqrt{k}}(\log R_j - \mu)\right)\right] - \varphi(\theta)\right|.
\end{aligned}$$

Equation (3.5) bounds the first term by $k\theta^2|\mathrm{cov}(R_i, R_j)| = O(k(\ell)\ell\mathcal{A}^{-\ell})$, so we can control that term. We control the second term by using the Lyapunov central limit theorem, (2.10).

## 4. Computational results

We now present the results of some calculations based both on simulations with random number generators and on the decimal digits of well-known constants. In each case, we calculate the value of the statistic

$$\frac{\sum_{i=1}^{k(\ell)}(\log S_i - \ell H \log 2 + \gamma)}{\sqrt{k(\ell)\pi^2/6}}.$$

We present the results as quantile–quantile plots produced using the statistical computing environment R. In each plot, the upper and lower quartiles are connected by a straight line. If the distribution of the statistic were exactly $N(0, 1)$, we would see the majority of the points lying very close to the line $y = x$.

To produce Figures 1 and 2, we performed 500 trials on simulated data. Figure 3 is based on breaking the first 20 million decimal digits of $\pi$ and e into 50 blocks of 400 000 digits. We used the program PiFast, version 4.3, by X. Gourdon (freely available online at http://numbers.computation.free.fr/Constants/PiProgram/pifast.html), which can easily calculate tens of million of digits of constants such as $\pi$ and e. In each case, the points do appear to lie on a straight line, although the sample variance is slightly smaller than expected. This could be remedied by dividing by the square root of the true variance,

$$\mathrm{var}\left(\sum_{i=1}^{k(\ell)}\log S_i\right) = \frac{k(\ell)\pi^2}{6} + k(\ell)(k(\ell) - 1)\,\mathrm{cov}(S_i, S_j) \le \frac{k(\ell)\pi^2}{6}.$$

In order to do this we would require an expansion, rather than simply an approximation, for the covariance in Lemma 3.2 (since the proof of Lemma 3.1 shows that the sums of $\log R_i$

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

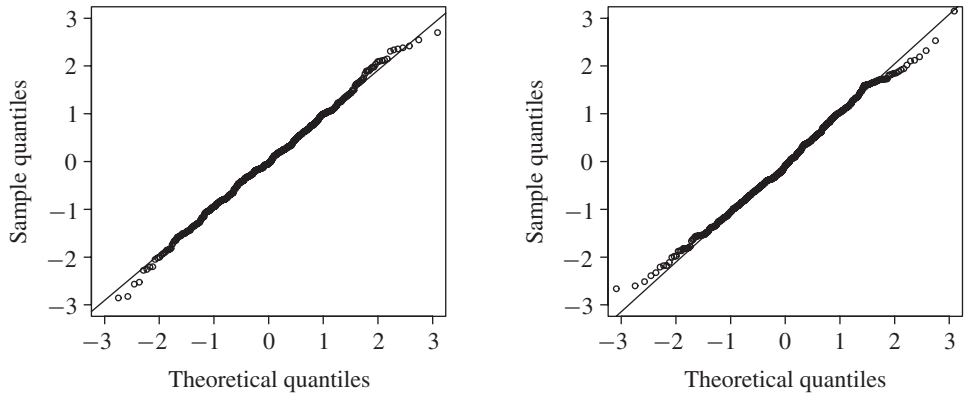


FIGURE 1: Quantile–quantile plots of equidistributed binary data: $k = 250$ and $\ell = 10$ (*left*); $k = 1000$ and $\ell = 13$ (*right*).

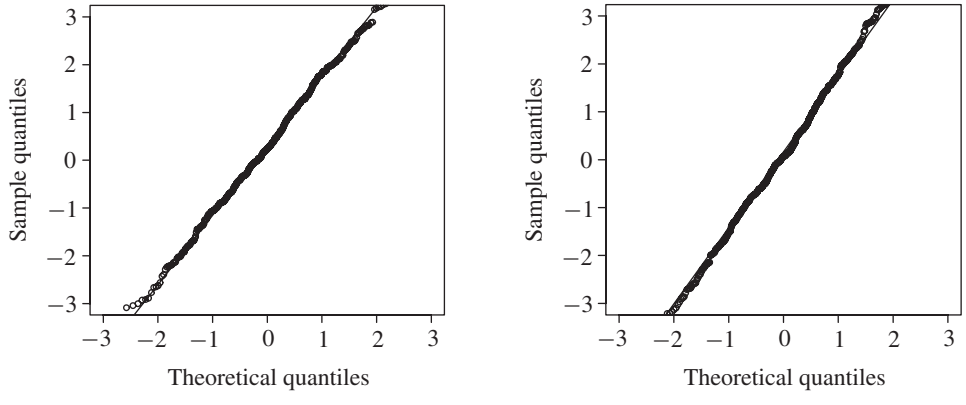


FIGURE 2: Quantile–quantile plots of asymmetric ($P(Z_1 = 0) = 0.75$) binary data: $k = 250$ and $\ell = 10$ (*left*); $k = 1000$ and $\ell = 13$ (*right*).

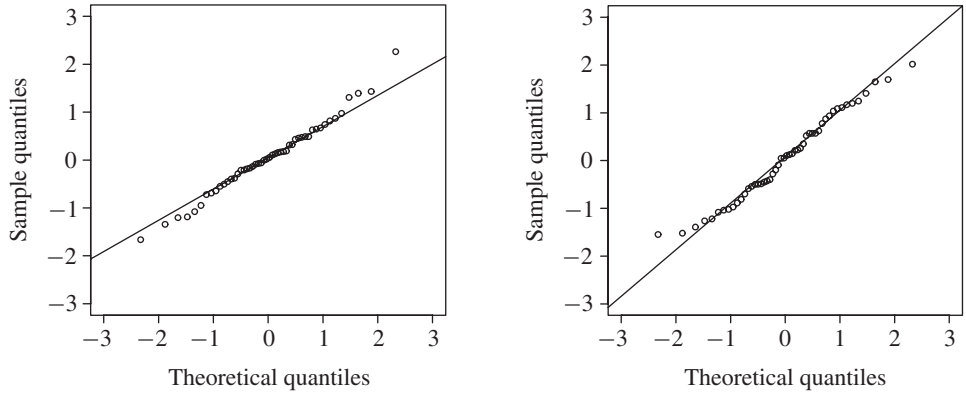


FIGURE 3: Quantile–quantile plots of decimal data with $k = 1000$ and $\ell = 4$: digits of $\pi$ (*left*); digits of e (*right*).

https://doi.org/10.1239/jap/1143936241 Published online by Cambridge University Press

and $\log S_i$ each have the same variance, asymptotically). Numerical calculation suggests that $\operatorname{cov}(R_i, R_j) \sim p \log p/4$. Of course, Lemma 3.2 only holds for equidistributed processes. However, in general, the asymptotic equipartition property suggests that we can assume that $\operatorname{cov}(R_i, R_j) \sim 2^{-H\ell} H\ell \log 2/4$.

## References


[1] Abadi, M. (2004). Sharp error terms and necessary conditions for exponential hitting times in mixing processes. *Ann. Prob.* **32,** 243–264.

[2] Abadi, M. and Galves, A. (2004). A version of Maurer's conjecture for stationary $\psi$-mixing processes. *Nonlinearity* **17,** 1357–1366.

[3] Abadi, M., Chazottes, J.-R., Redig, F. and Verbitskiy, E. (2004). Exponential distribution for the occurrence of rare patterns in Gibbsian random fields. *Commun. Math. Phys.* **246,** 269–294.

[4] Aldous, D. and Shields, P. (1988). A diffusion limit for a class of randomly-growing binary trees. *Prob. Theory Relat. Fields* **79,** 509–542.

[5] Bailey, D. H. and Crandall, R. E. (2002). Random generators and normal numbers. *Experiment. Math.* **11,** 527–546.

[6] Bradley, W. F. and Suhov, Y. M. (1997). The entropy of famous reals: some empirical results. *Random Comput. Dynam.* **5,** 349–359.

[7] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley, New York.

[8] Gao, Y. (2004). *Statistical Models in Neural Information Processing*. Doctoral Thesis, Brown University.

[9] Grassberger, P. (1989). Estimating the information content of symbol sequences and efficient codes. *IEEE Trans. Inf. Theory* **35,** 669–675.

[10] Hu, T. and Yang, J. (2004). Further developments on sufficient conditions for negative dependence of random variables. *Statist. Prob. Lett.* **66,** 369–381.

[11] Jacquet, P. and Szpankowski, W. (1995). Asymptotic behavior of the Lempel–Ziv parsing scheme and digital search trees. *Theoret. Comput. Sci.* **144,** 161–197.

[12] Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables, with applications. *Ann. Statist.* **11,** 286–295.

[13] Kac, M. (1947). On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.* **53,** 1002–1010.

[14] Kim, D. H. (2003). The recurrence of blocks for Bernoulli processes. *Osaka J. Math.* **40,** 171–186.

[15] Kontoyiannis, I. (1998). Asymptotic recurrence and waiting times for stationary processes. *J. Theoret. Prob.* **11,** 795–811.

[16] Kontoyiannis, I. and Suhov, Yu. (1994). Prefixes and the entropy rate for long-range sources. In *Probability, Statistics and Optimisation*, ed. F. P. Kelly, John Wiley, Chichester, pp. 89–98.

[17] Maurer, U. M. (1992). A universal statistical test for random bit generators. *J. Cryptology* **5,** 89–105.

[18] Newman, C. M. (1980). Normal fluctuations and the FKG inequalities. *Commun. Math. Phys.* **74,** 119–128.

[19] Ornstein, D. S. and Weiss, B. (1993). Entropy and data compression schemes. *IEEE Trans. Inf. Theory* **39,** 78–83.

[20] Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford University Press.

[21] Quas, A. N. (1998). An entropy estimator for a class of infinite alphabet processes. *Teor. Veroyat. Primen.* **43,** 610–621 (summary in Russian). English translation: *Theory Prob. Appl.* **43,** 496–507.

[22] Shields, P. C. (1992). Entropy and prefixes. *Ann. Prob.* **20,** 403–409.

[23] Shields, P. C. (1997). String matching bounds via coding. *Ann. Prob.* **25,** 329–336.

[24] Wyner, A. D. and Ziv, J. (1989). Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inf. Theory* **35,** 1250–1258.

[25] Wyner, A. J. (1999). More on recurrence and waiting times. *Ann. Appl. Prob.* **9,** 780–796.

[26] Ziv, J. (1978). Coding theorems for individual sequences. *IEEE Trans. Inf. Theory* **24,** 405–412.