J. Appl. Prob. **51,** 512–527 (2014)
Printed in England

ON A 'REPLICATING CHARACTER STRING' MODEL

RICHARD C. BRADLEY,* *Indiana University*

Abstract

In Chaudhuri and Dasgupta's 2006 paper a certain stochastic model for 'replicating character strings' (such as in DNA sequences) was studied. In their model, a random 'input' sequence was subjected to random mutations, insertions, and deletions, resulting in a random 'output' sequence. In this paper their model will be set up in a slightly different way, in an effort to facilitate further development of the theory for their model. In their 2006 paper, Chaudhuri and Dasgupta showed that, under certain conditions, strict stationarity of the 'input' sequence would be preserved by the 'output' sequence, and they proved a similar 'preservation' result for the property of strong mixing with exponential mixing rate. In our setup, we will in spirit slightly extend their 'preservation of stationarity' result, and also prove a 'preservation' result for the property of absolute regularity with summable mixing rate.

Keywords: Replicating character string; stationarity; absolute regularity

2010 Mathematics Subject Classification: Primary 60G10

1. Introduction

Chaudhuri and Dasgupta [8] formulated and studied a certain stochastic model for 'replicating character strings'. In that paper, they cited numerous references where other related models had been studied, and, in particular, they cited the book by Waterman [19] for the possible application of central limit theory under strong mixing conditions in the use of such models for the statistical analysis of data from biology (e.g. involving DNA sequences). In this paper we shall contribute further results and techniques to the theory for the particular model in [8], and suggest a way of setting up their model that may allow slightly easier handling of certain technical details.

Let $\mathbb{N}$ and $\mathbb{Z}$ respectively denote the set of all positive integers and the set of all integers.

The model studied by Chaudhuri and Dasgupta [8] can be briefly described as follows. It starts with an 'input' sequence $X := (X_k,\ k \in \mathbb{N})$ of random variables taking values in some finite 'alphabet'—for example, the set $\{A, C, G, T\}$ of letters that represent the nucleotides in a DNA sequence. There is another sequence $Z := (Z_k,\ k \in \mathbb{N})$ of random variables taking values in the set $\{M, I, D\}$—to indicate that at a given 'time' (or 'location') k, there should be a 'mutation' (M), 'insertion' (I), or 'deletion' (D). (This sequence Z is informally referred to below as the 'MID sequence'.) Probabilities are assigned for what letter of the alphabet is inserted when an insertion occurs, or what letter of the alphabet results from a mutation. (The—perhaps high—probability of 'no mutation' is formally represented in this scheme as the probability of 'replacing a letter by itself' when a mutation occurs.) At the end, the result is an output sequence $Y := (Y_k,\ k \in \mathbb{N})$ of random variables, with the same alphabet (e.g. $\{A, C, G, T\}$) as the 'input' sequence X.

Received 23 November 2012; revision received 16 May 2013.
* Postal address: Department of Mathematics, Indiana University, Bloomington, Indiana 47405, USA.
Email address: bradleyr@indiana.edu

In their paper, Chaudhuri and Dasgupta [8, Theorems 3.1 and 3.2] established certain conditions under which certain properties of the input sequence—specifically, strict stationarity, and strong mixing with exponential mixing rate—would be retained by the output sequence. Chaudhuri and Dasgupta [8] set up their model using ('one-sided') random sequences indexed by $\mathbb{N}$, as described above. In Section 3 we will set up their model again, but using ('two-sided') random sequences indexed by $\mathbb{Z}$. This will hopefully make it a little easier to handle various technical details, such as keeping track of relevant σ-fields when estimating mixing rates.

In the statements of their main results (though not exclusively in the initial formulation of their model), Chaudhuri and Dasgupta [8] dealt with the case where the MID sequence is an (irreducible, aperiodic) Markov chain that is independent of the 'input' sequence. We shall retain that 'independence' assumption, but allow the MID sequence itself to satisfy a somewhat more flexible dependence assumption than a 'Markov' property.

Instead of studying the mixing rates for the input and output sequences for the strong mixing condition, we shall do so for the absolute regularity condition, which is stronger than strong mixing. This will provide an opportunity to illustrate the use of a particularly handy 'coupling' property (due to Berbee [1]) that is possessed by the absolute regularity condition but not by the strong mixing condition. However, along the way, we shall also give information that may be relevant to the further development of the theory for this model under the strong mixing condition.

Instead of studying the case of exponential mixing rates (for strong mixing) as in [8], we shall focus on a certain slower ('summable') mixing rate (for absolute regularity) that is natural in central limit theory for bounded random variables (under either strong mixing or absolute regularity).

In the model in [8], one somewhat tricky facet of keeping track of relevant σ-fields was keeping track of the changes in the 'clock' resulting from deletions. We will adopt an alternative technical procedure—switching to a new probability measure based on conditioning on a certain event—in the hope of slightly simplifying that task.

In their model, Chaudhuri and Dasgupta [8] assumed a finite state space (for the input and output sequences), as described above. That is the case of primary interest; but, for convenience, we shall relax that assumption and allow the input and output sequences to consist in essence of real-valued random variables. We shall actually treat those random variables as taking their values in $(0, \infty)$ (think of 'coding' a real number x by the positive number e^x), and in an intermediate stage reserve the value 0 as a temporary 'place holder' where an insertion will ultimately occur.

The model in [8] directly involved probability mass functions for what happens when a mutation or insertion occurs. As a measure-theoretic convenience, we shall handle that in a slightly different way, using independent random variables uniformly distributed on the unit interval as 'randomizers'.

In making these modifications, we will not change the actual model studied by Chaudhuri and Dasgupta [8] in any significant way. The modifications here only involve how their model is set up. Our aim is in part to facilitate further development of the theory for their model. There is of course the practical question, not addressed here, of to what extent inaccuracy may occur when, say, a 'long but finite' DNA sequence is modeled as a two-sided random sequence.

In Section 2, some preliminary information on both the strong mixing and absolute regularity conditions will be given. In Section 3, the model in [8] will be presented with the modifications in the setup described above. Then in Section 4, the main result of this paper will be stated and proved.

2. Preliminary information on two mixing conditions

In the development of the material in Sections 3 and 4, we will start with a probability space and then switch to a new probability measure (on the same measurable space) obtained by conditioning on a certain key event. Accordingly, in the notation used in the definitions below, the relevant probability measure will be specified explicitly. If only one probability measure, say $\mathbb{P}$, is specified, then the notation $\mathbb{E}(\cdot)$ will be tacitly understood to mean the expected value with respect to that particular probability measure $\mathbb{P}$.

Suppose that $(\Omega, \mathcal{F})$ is a measurable space. Suppose that $W := (W_i,\ i \in I)$ is a random variable/vector or stochastic process indexed by a nonempty set I—that is, $W\colon \Omega \to \mathbb{R}^I$ is a function which is measurable with respect to the σ-field $\mathcal{F}$ on Ω and the Borel σ-field on $\mathbb{R}^I$. The σ-field (a subset of $\mathcal{F}$) of subsets of Ω generated by W will be denoted by $\sigma(W)$ or $\sigma(W_i,\ i \in I)$.

Definition 2.1. Suppose that $(\Omega, \mathcal{F})$ is a measurable space, and that $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

For any two σ-fields $\mathcal{A}$ and $\mathcal{B}$ that are subsets of $\mathcal{F}$, define the following two measures of dependence:

$$\alpha(\mathcal{A}, \mathcal{B}; \mathbb{P}) := \sup_{A\in\mathcal{A},\, B\in\mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

$$\beta(\mathcal{A}, \mathcal{B}; \mathbb{P}) := \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|. \tag{2.1}$$

In (2.1) the supremum is taken over all pairs of partitions $\{A_1, A_2, \ldots, A_I\}$ and $\{B_1, B_2, \ldots, B_J\}$ of Ω such that $A_i \in \mathcal{A}$ for each i and $B_j \in \mathcal{B}$ for each j. It is easy to see that, for any two σ-fields $\mathcal{A}$ and $\mathcal{B}$, we have

$$\alpha(\mathcal{A}, \mathcal{B}; \mathbb{P}) \le \beta(\mathcal{A}, \mathcal{B}; \mathbb{P}). \tag{2.2}$$

Suppose that $X := (X_k,\ k \in \mathbb{Z})$ is, with respect to $\mathbb{P}$, a strictly stationary sequence of random variables. For each $n \in \mathbb{N}$, define the dependence coefficients

$$\alpha(X, n; \mathbb{P}) := \alpha(\sigma(X_k,\ k \le 0), \sigma(X_k,\ k \ge n); \mathbb{P}), \tag{2.3}$$

$$\text{and} \quad \beta(X, n; \mathbb{P}) := \beta(\sigma(X_k,\ k \le 0), \sigma(X_k,\ k \ge n); \mathbb{P}).$$

One trivially has that each of the sequences of numbers $(\alpha(X, n; \mathbb{P}),\ n \in \mathbb{N})$ and $(\beta(X, n; \mathbb{P}),\ n \in \mathbb{N})$ is nonincreasing. Also, by (2.2), $\alpha(X, n; \mathbb{P}) \le \beta(X, n; \mathbb{P})$ for every positive integer n. The sequence X is (with respect to the probability measure $\mathbb{P}$) 'strongly mixing' [16] if $\alpha(X, n; \mathbb{P}) \to 0$ as $n \to \infty$, and 'absolutely regular' [18] if $\beta(X, n; \mathbb{P}) \to 0$ as $n \to \infty$. By (2.2), absolute regularity implies strong mixing.

To motivate the results later in this paper, we will state a classic theorem of Ibragimov, from Ibragimov and Linnik [12, Theorem 18.5.4].

Theorem 2.1. (Ibragimov.) *Suppose that on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, $X := (X_k,\ k \in \mathbb{Z})$ *is a strictly stationary sequence of bounded, centered random variables such that* $\sum_{n=1}^{\infty} \alpha(X, n; \mathbb{P}) < \infty$. *Then* $\sigma^2 := \mathbb{E}X_0^2 + 2\sum_{n=1}^{\infty} \mathbb{E}X_0X_n$ *exists in* $[0, \infty)$, *with this sum being absolutely convergent. If further* $\sigma^2 > 0$ *then* $(X_1 + X_2 + \cdots + X_n)/(n^{1/2}\sigma)$ *converges in distribution to the* $N(0, 1)$ *law as* $n \to \infty$.

Theorem 2.1 will not be used anywhere in what follows, but it will provide the motivation for the mathematical development in this paper. For example, in a statistical analysis of DNA

data, one might deal with indicator functions ($\{0, 1\}$-valued random variables) marking the locations of a particular pattern of nucleotides along a DNA sequence. Thus, if strong mixing is assumed as part of the statistical model then it might be natural to apply a central limit theorem for *bounded* strongly mixing sequences of random variables, such as Theorem 2.1. Now the (summable) mixing rate in Theorem 2.1 is practically sharp. This was shown to be true even under absolute regularity, by counterexamples in [9, Example 2] and [4]. The counterexample in the latter paper is a strictly stationary, three-state sequence that satisfies absolute regularity with ('not quite summable') mixing rate $\beta(X, n; \mathbb{P}) = O(1/n)$. Theorem 2.1 seems natural to use when either strong mixing or absolute regularity is assumed in the modeling of DNA sequences; it is the summable mixing rate in that theorem that we will focus on in this paper.

It is worth noting that Merlevède and Peligrad [13] proved (as a special case of the main result in their paper) a modified, refined version of Theorem 2.1 with the barely slower mixing rate $\alpha(X, n; \mathbb{P}) = o(1/n)$ and an explicit extra assumption on the rate of growth of the variances of partial sums.

As mentioned in Section 1, instead of dealing with the strong mixing condition, we will deal with absolute regularity. This will provide an opportunity to illustrate the use—in steps 5 and 6 of the proof of Lemma 4.4 in Section 4—of a handy 'coupling' property (from [1]) of the absolute regularity condition. This property does not exist, at least in as strong a form, under just strong mixing. The next three lemmas will facilitate that particular application of the coupling property.

Lemma 2.1. *Suppose that* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space,* N *is a positive integer,* $\mathcal{A}_n$ *and* $\mathcal{B}_n$, $n \in \{1, 2, \ldots, N\}$, *are* σ*-fields that are subsets of* $\mathcal{F}$*, and that the* σ*-fields* $\mathcal{A}_n \vee \mathcal{B}_n$, $n \in \{1, 2, \ldots, N\}$, *are independent (under* $\mathbb{P}$*). Then*

$$\beta\left(\bigvee_{n=1}^{N} \mathcal{A}_n, \bigvee_{n=1}^{N} \mathcal{B}_n; \mathbb{P}\right) \leq \sum_{n=1}^{N} \beta(\mathcal{A}_n, \mathcal{B}_n; \mathbb{P}).$$

In one form or another, this lemma has long been part of the folklore; see, e.g. [14, p. 73]. One reference for the particular formulation here is [5, Theorem 6.2]. (The same theorem in [5] also gives the exactly analogous inequality for the dependence coefficient $\alpha(\cdot, \cdot)$.)

The next lemma has also long been part of the folklore, but a reference for it seems hard to find. In this lemma, the random variables X and Y are not assumed to be identically distributed, and the term 'Borel space' means a measurable space $(S, \mathcal{S})$ that is bimeasurably isomorphic to the space $(\mathbb{R}, \mathcal{R})$, where $\mathcal{R}$ denotes the Borel σ-field on $\mathbb{R}$. It is well known that $\mathbb{R}^{\mathbb{N}}$ (or $\mathbb{R}^{\mathbb{Z}}$), accompanied by its Borel σ-field, is a Borel space.

Lemma 2.2. *Suppose that* $(S, \mathcal{S})$ *is a Borel space. Suppose that* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space,* X *and* Y *are random variables on this space which take values in* $(S, \mathcal{S})$*, and that* $\mathcal{A}$ *is a* σ*-field that is a subset of* $\mathcal{F}$*. Then*

$$|\beta(\mathcal{A}, \sigma(X); \mathbb{P}) - \beta(\mathcal{A}, \sigma(Y); \mathbb{P})| \leq 2\mathbb{P}(X \neq Y).$$

Proof. By symmetry, it suffices to prove that $\beta(\mathcal{A}, \sigma(X); \mathbb{P}) \leq \beta(\mathcal{A}, \sigma(Y); \mathbb{P}) + 2\mathbb{P}(X \neq Y)$. Suppose that $\{A_1, A_2, \ldots, A_I\}$ and $\{B_1, B_2, \ldots, B_J\}$ are each a partition of Ω, with $A_i \in \mathcal{A}$ for each i and $B_j \in \sigma(X)$ for each j. It suffices to show that

$$\frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| \leq \beta(\mathcal{A}, \sigma(Y); \mathbb{P}) + 2\mathbb{P}(X \neq Y). \tag{2.4}$$

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

By a well-known measure-theoretic fact (a standard elementary generalization of [2, Theorem 20.1]), there exists a partition $\{S_1, S_2, \ldots, S_J\}$ of S with $S_j \in \mathcal{S}$ for each j such that $B_j = \{X \in S_j\}$ for each j.

For any event A,

$$\begin{aligned}
&\sum_{j=1}^{J} |\mathbb{P}(A \cap \{X \in S_j\}) - \mathbb{P}(A \cap \{Y \in S_j\})| \\
&\quad \le \sum_{j=1}^{J} |\mathbb{P}(A \cap \{X \in S_j\} \cap \{X = Y\}) + \mathbb{P}(A \cap \{X \in S_j\} \cap \{X \ne Y\}) \\
&\qquad\qquad - \mathbb{P}(A \cap \{Y \in S_j\} \cap \{X = Y\}) - \mathbb{P}(A \cap \{Y \in S_j\} \cap \{X \ne Y\})| \\
&\quad = \sum_{j=1}^{J} |\mathbb{P}(A \cap \{X \in S_j\} \cap \{X \ne Y\}) - \mathbb{P}(A \cap \{Y \in S_j\} \cap \{X \ne Y\})| \\
&\quad \le 2\mathbb{P}(A \cap \{X \ne Y\}).
\end{aligned}$$

Applying that with $A = A_i$ and then also with $A = \Omega$, we have

$$\begin{aligned}
&\sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| \\
&\quad \le \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(A_i \cap \{X \in S_j\}) - \mathbb{P}(A_i \cap \{Y \in S_j\})| \\
&\qquad + \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(A_i \cap \{Y \in S_j\}) - \mathbb{P}(A_i)\mathbb{P}(Y \in S_j)| \\
&\qquad + \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(A_i)\mathbb{P}(Y \in S_j) - \mathbb{P}(A_i)\mathbb{P}(X \in S_j)| \\
&\quad \le \left[\sum_{i=1}^{I} 2\mathbb{P}(A_i \cap \{X \ne Y\})\right] + 2\beta(\mathcal{A}, \sigma(Y); \mathbb{P}) \\
&\qquad + \sum_{i=1}^{I}\left[\mathbb{P}(A_i) \sum_{j=1}^{J} |\mathbb{P}(Y \in S_j) - \mathbb{P}(X \in S_j)|\right] \\
&\quad \le 2\mathbb{P}(X \ne Y) + 2\beta(\mathcal{A}, \sigma(Y); \mathbb{P}) + \sum_{i=1}^{I} [\mathbb{P}(A_i) 2\mathbb{P}(X \ne Y)] \\
&\quad = 4\mathbb{P}(X \ne Y) + 2\beta(\mathcal{A}, \sigma(Y); \mathbb{P}).
\end{aligned}$$

Thus, (2.4) holds. This completes the proof.

The following lemma, the final item of Section 2 here, will play a key role (in Section 4) in the comparison of the mixing rates for the input and output sequences in the model in [8]. It will be applied for absolute regularity, but it holds under strong mixing (as stated and proved here) as well.

Lemma 2.3. *Suppose that* $(H(0), H(1), H(2), H(3), \ldots)$ *is a nonincreasing sequence of nonnegative numbers such that* $\sum_{n=0}^{\infty} H(n) < \infty$. *Suppose that on some probability space*

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

$(\Omega, \mathcal{F}, \mathbb{P})$, $X := (X_k,\ k \in \mathbb{Z})$ is a nondegenerate, strictly stationary sequence of random variables taking only the values 0 *and* 1 *such that* $\sum_{n=1}^{\infty} \alpha(X, n; \mathbb{P}) < \infty$. *Then*

$$\sum_{n=1}^{\infty} \mathbb{E}H(X_1 + X_2 + \cdots + X_n) < \infty. \tag{2.5}$$

Proof. Referring to the hypothesis (of Lemma 2.3), define the number

$$p := \mathbb{P}(X_0 = 1) = \mathbb{E}X_0 > 0. \tag{2.6}$$

Define the constant random variable $S_0 := 0$, and, for each positive integer n, define the partial sum $S_n := X_1 + X_2 + \cdots + X_n$. By the hypothesis (of Lemma 2.3), the sequence X is strongly mixing and, hence, ergodic. Hence, from (2.6), $S_n \to \infty$ (monotonically) almost surely as $n \to \infty$. For technical convenience, without loss of generality (redefining the random variables X_k on a $\mathbb{P}$-null set if necessary), we assume that this happens at literally every $\omega \in \Omega$.

For each nonnegative integer j, define the random variable $\eta_j := \text{card}\{n \in \mathbb{N}\colon S_n = j\}$. Then, for every integer $J \geq 0$,

$$\sum_{j=0}^{J} \eta_j = \max\{n \geq 0\colon S_n = J\}. \tag{2.7}$$

In what follows, for any real number x, let $[x]$ denote the greatest integer less than or equal to x. Also, in the calculations below, by the hypothesis (of Lemma 2.3), all sums and summands ('numerical' or random) take their values in $[0, \infty] := [0, \infty) \cup \{\infty\}$, and, hence, we can change the orders of summations arbitrarily.

Recall that, for any nonnegative integer-valued random variable W, $\mathbb{E}W = \sum_{n=1}^{\infty} \mathbb{P}(W \geq n)$. For each integer $J \geq 0$, by (2.7) and the trivial inequality $\mathbb{P}(S_n \leq J) \leq 1$, we have

$$\mathbb{E}\left(\sum_{j=0}^{J} \eta_j\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(\sum_{j=0}^{J} \eta_j \geq n\right) = \sum_{n=1}^{\infty} \mathbb{P}(S_n \leq J) \leq \frac{2J}{p} + \sum_{n=[2J/p]+1}^{\infty} \mathbb{P}(S_n \leq J). \tag{2.8}$$

Let us examine the last sum in (2.8). For each integer $J \geq 0$ and each integer $n > 2J/p$, we have $J < np/2$ and, hence, $J - np < -np/2$. Hence, for each integer $J \geq 0$,

$$\begin{aligned}
\sum_{n=[2J/p]+1}^{\infty} \mathbb{P}(S_n \leq J) &\leq \sum_{n=[2J/p]+1}^{\infty} \mathbb{P}\left(S_n - np \leq -\frac{np}{2}\right) \\
&\leq \sum_{n=[2J/p]+1}^{\infty} \mathbb{P}\left(|S_n - np| \geq \frac{np}{2}\right) \\
&\leq \sum_{n=[2J/p]+1}^{\infty} \left(\frac{np}{2}\right)^{-4} \mathbb{E}(S_n - np)^4 \\
&\leq \frac{16}{p^4} \sum_{n=1}^{\infty} n^{-4} \mathbb{E}(S_n - np)^4.
\end{aligned} \tag{2.9}$$

Extend definition (2.3) to include $n = 0$. Then, for each positive integer n,

$$\mathbb{E}(S_n - np)^4 \leq (20\,000)n \sum_{m=0}^{n-1} (m+1)^2 \alpha(X, m; \mathbb{P}) + 24n^2 \left[\sum_{m=0}^{n-1} \alpha(X, m; \mathbb{P})\right]^2. \tag{2.10}$$

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

This is a simple direct application of [6, Theorem 14.63], which in turn is a convenient but crude version of a much sharper and more general inequality due to Rio [15, Theorem 2.1]. Keep in mind that, since the random variables X_k take only the values 0 and 1, the ('upper-tail') quantile functions in those particular statements in both references take only the values 0 and 1. Equation (2.10) can also be obtained directly, in a sharper form, from a careful calculation from Ibragimov's proof in [12, Theorem 18.5.4] of Theorem 2.1 (examine carefully the argument for [12, Lemma 18.5.2]).

Our next task is to use (2.10) to show that the last sum in (2.9) is finite. From simple calculus, let C_1 be a positive number such that $\sum_{n=q}^{\infty} n^{-3} \le C_1 q^{-2}$ for every positive integer q. By the hypothesis (of Lemma 2.3),

$$\begin{aligned}\sum_{n=1}^{\infty}\Bigg[n^{-4}n\sum_{m=0}^{n-1}(m+1)^2\alpha(X,m;\mathbb{P})\Bigg] &= \sum_{m=0}^{\infty}\sum_{n=m+1}^{\infty} n^{-3}(m+1)^2\alpha(X,m;\mathbb{P}) \\ &\le \sum_{m=0}^{\infty} C_1\alpha(X,m;\mathbb{P}) \\ &< \infty.\end{aligned}$$

Also, trivially by the hypothesis, $\sum_{n=1}^{\infty}[n^{-4}n^2[\sum_{m=0}^{n-1}\alpha(X,m;\mathbb{P})]^2] < \infty$. Applying those two inequalities to (2.10), we find that the last sum in (2.9) is finite.

Accordingly, defining the finite numbers $C_2 := (16/p^4)\sum_{n=1}^{\infty} n^{-4}\mathbb{E}(S_n - np)^4$ and $C_3 := 2/p + C_2$, we have, by (2.8) and (2.9), for every integer $J \ge 0$,

$$\mathbb{E}\left(\sum_{j=0}^{J}\eta_j\right) \le \frac{2J}{p} + C_2 \le C_3(J+1). \tag{2.11}$$

Now refer to the function H in the statement of Lemma 2.3. By the hypothesis, $H(n) \downarrow 0$ as $n \to \infty$. Using the notation $S(n)$ for S_n in subscripts, we have the equality of nonnegative random variables (possibly taking the value ∞)

$$\begin{aligned}\sum_{n=1}^{\infty} H(S_n) &= \sum_{j=0}^{\infty}\sum_{\{n\in\mathbb{N}:\ S(n)=j\}} H(j) \\ &= \sum_{j=0}^{\infty} H(j)\eta_j \\ &= \sum_{j=0}^{\infty}\sum_{i=j}^{\infty}\eta_j[H(i)-H(i+1)] \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{i}[H(i)-H(i+1)]\eta_j.\end{aligned}$$

Hence, by (2.11),

$$\begin{aligned}\mathbb{E}\sum_{n=1}^{\infty} H(S_n) &\le \sum_{i=0}^{\infty}[[H(i)-H(i+1)]C_3(i+1)] \\ &= C_3\sum_{i=0}^{\infty}\sum_{j=0}^{i}[H(i)-H(i+1)]\end{aligned}$$

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

$$
\begin{aligned}
&= C_3 \sum_{j=0}^{\infty} \sum_{i=j}^{\infty} [H(i) - H(i+1)] \\
&= C_3 \sum_{j=0}^{\infty} H(j) \\
&< \infty.
\end{aligned}
$$

Thus, (2.5) holds. This completes the proof of Lemma 2.3.

3. The model of Chaudhuri and Dasgupta, in two-sided form

In this section we will describe, step by step, the 'replicating character string' model studied by Chaudhuri and Dasgupta [8]. As explained in Section 1, essentially the only changes here will be in the 'style': (i) the use of (two-sided) random sequences indexed by $\mathbb{Z}$, rather than (one-sided) random sequences indexed by $\mathbb{N}$, and (ii) the trivial allowing of the alphabet or 'state space' to be $(0, \infty)$ instead of just a finite set. In the presentation here, the essential mathematical substance of their model will not be changed at all. Much of the notation listed below will be taken directly from their paper. For convenient reference, the stages in this construction will be referred to as paragraphs (P1), (P2), etc.

(P1) Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. All the random variables defined below will be understood to be defined on this space.

(P2) Suppose that $X := (X_k,\ k \in \mathbb{Z})$ is (under $\mathbb{P}$) a strictly stationary sequence of random variables taking their values in the open half-line $(0, \infty)$. (This is the input sequence, as in the model in [8].)

(P3) Suppose that $Z := (Z_k, k \in \mathbb{Z})$ is (under $\mathbb{P}$) a strictly stationary, ergodic sequence of random variables taking values in the set $\{M, I, D\}$, with this sequence Z being independent of the sequence X. Assume further that $\mathbb{P}(Z_0 = s) > 0$ for all three elements $s \in \{M, I, D\}$. For technical convenience, without loss of generality (by ergodicity), assume that, for *every* $\omega \in \Omega$ and all three elements $s \in \{M, I, D\}$, $Z_k(\omega) = s$ for infinitely many negative integers k and infinitely many positive integers k. (Again, the letters M, I, and D stand for mutation, insertion, and deletion; Z is the MID sequence, as in [8].)

(P4) Define the strictly increasing sequence $\zeta := (\zeta_j,\ j \in \mathbb{Z})$ of integer-valued random variables as follows. For every $\omega \in \Omega$,

$$
\begin{aligned}
&\cdots < \zeta_{-2}(\omega) < \zeta_{-1}(\omega) < \zeta_0(\omega) \le 0 < 1 \le \zeta_1(\omega) < \zeta_2(\omega) < \zeta_3(\omega) < \cdots \\
&\qquad \text{and} \quad \{j \in \mathbb{Z}\colon Z_j(\omega) \in \{M, D\}\} = \{\zeta_k(\omega)\colon k \in \mathbb{Z}\}.
\end{aligned} \tag{3.1}
$$

The random variables ζ_k will sometimes be written $\zeta(k)$ for typographical convenience.

(P5) Define the sequence $\bar{X} := (\bar{X}_k,\ k \in \mathbb{Z})$ of random variables (taking their values in the closed half-line $[0, \infty)$) as follows. For every $\omega \in \Omega$,

$$
\bar{X}_{\zeta(k)(\omega)}(\omega) := X_k(\omega) \quad \text{for all } k \in \mathbb{Z} \quad \text{and} \quad \bar{X}_j(\omega) := 0 \quad \text{for all } j \notin \{\zeta_k(\omega)\colon k \in \mathbb{Z}\}. \tag{3.2}
$$

The state 0 is used here only as a 'temporary placeholder' for a spot where an insertion will eventually occur (in (3.5) below for the case where $\bar{Y}_\ell(\omega) = 0$ there). This was the sole motivation for choosing a state space for the original sequence X, namely $(0, \infty)$, that does not include 0.

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

(P6) Define the strictly increasing sequence $\xi := (\xi_j,\ j \in \mathbb{Z})$ of integer-valued random variables as follows. For every $\omega \in \Omega$,

$$\cdots < \xi_{-2}(\omega) < \xi_{-1}(\omega) < \xi_0(\omega) \le 0 < 1 \le \xi_1(\omega) < \xi_2(\omega) < \xi_3(\omega) < \cdots$$
$$\text{and} \quad \{j \in \mathbb{Z}\colon Z_j(\omega) \in \{M, I\}\} = \{\xi_k(\omega)\colon k \in \mathbb{Z}\}. \tag{3.3}$$

These random variables ξ_k will sometimes be written $\xi(k)$.

(P7) Define the sequence $\bar{Y} := (\bar{Y}_\ell,\ \ell \in \mathbb{Z})$ of random variables (taking values in $[0, \infty)$) as follows. For every $\omega \in \Omega$,

$$\bar{Y}_\ell(\omega) := \bar{X}_{\xi(\ell)(\omega)}(\omega) \quad \text{for all } \ell \in \mathbb{Z}. \tag{3.4}$$

(P8) (i) Let $U := (U_k,\ k \in \mathbb{Z})$ be (under $\mathbb{P}$) a sequence of independent, identically distributed random variables, each uniformly distributed on the interval $[0, 1]$, with this sequence U being independent of the pair of sequences (X, Z).

(ii) Let $g\colon (0, \infty) \times [0, 1] \to (0, \infty)$ be a Borel function.

(iii) Let $h\colon [0, 1] \to (0, \infty)$ be a Borel function.

(iv) Define the sequence $Y := (Y_\ell,\ \ell \in \mathbb{Z})$ of random variables (taking values in $(0, \infty)$) as follows. For every $\omega \in \Omega$,

$$Y_\ell(\omega) := \begin{cases} g(\bar{Y}_\ell(\omega), U_\ell(\omega)) & \text{if } \bar{Y}_\ell(\omega) \in (0, \infty), \\ h(U_\ell(\omega)) & \text{if } \bar{Y}_\ell(\omega) = 0, \end{cases} \quad \text{for all } \ell \in \mathbb{Z}. \tag{3.5}$$

(The sequence Y is the output sequence, as in [8]. As described in more detail in Remark 3.1(d) below, the functions g and h deal with mutations and insertions, respectively.)

The final two 'paragraphs' below give a few more items that were not needed in the formulation of the output sequence Y, but will be needed in the formulation of the main result (Theorem 4.1 below).

(P9) Let $\mathbb{P}_0$ denote the probability measure on $(\Omega, \mathcal{F})$ defined as follows:

$$\mathbb{P}_0(F) := \mathbb{P}(F \mid Z_0 \in \{M, I\}) = \mathbb{P}(F \mid \xi_0 = 0) \quad \text{for all } F \in \mathcal{F}. \tag{3.6}$$

(The second equality follows from (P6).)

(P10) Define the sequence $V := (V_k,\ k \in \mathbb{Z})$ of $(\{M, I\} \times \mathbb{N})$-valued random variables as

$$V_k := (Z_{\xi(k)}, \xi_k - \xi_{k-1}) \quad \text{for all } k \in \mathbb{Z}. \tag{3.7}$$

Also, define the sequence $\Upsilon := (\Upsilon_k,\ k \in \mathbb{Z})$ of $(\{M, I\} \times \mathbb{N} \times (0, \infty))$-valued random variables as

$$\Upsilon_k := (V_k, Y_k) = (Z_{\xi(k)}, \xi_k - \xi_{k-1}, Y_k) \quad \text{for all } k \in \mathbb{Z}. \tag{3.8}$$

This completes the (two-sided) presentation of the model in [8].

Remark 3.1. We present several comments pertaining to the model from [8] as spelled out in (P1)–(P10).

(a) It is well known from renewal theory that even though the sequence Z is (under the original probability measure $\mathbb{P}$) strictly stationary, the sequence $(Z_{\xi(k)}, k \in \mathbb{Z})$ is in general *not* strictly stationary under $\mathbb{P}$. As a consequence, under $\mathbb{P}$, the output sequence Y will in general not be strictly stationary. To obtain the stationarity of Y, Chaudhuri and Dasgupta [8, Theorem 3.1] directly assumed that the sequence $(Z_{\xi(k)},\ k \in \mathbb{Z})$ (though not

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

necessarily the entire sequence Z) is strictly stationary, with the original MID sequence Z itself being a Markov chain with certain properties. Theorem 4.1 below (the main result of this paper) will employ the procedure, common in renewal theory, of formally switching to the new probability measure $\mathbb{P}_0$ in (3.6), which under our assumptions will yield the stationarity of Y (and of the entire sequence Υ in (3.8)), with no 'Markov' assumption. This is the role here of the probability measure $\mathbb{P}_0$.

(b) By (P4), (P6), and (P10), we have $\sigma(\zeta, \xi, V) \subset \sigma(Z)$. Recall from (P3) and (P8) that, under $\mathbb{P}$, the random sequences X, Z, and U are independent of each other. By (3.6) and a trivial argument, these three sequences are independent of each other under $\mathbb{P}_0$ as well. Also, by (3.6), the random sequences X and U (but, in general, not Z or even V) each have the same distribution under $\mathbb{P}_0$ as they do under $\mathbb{P}$. In particular,

$$\beta(X, n; \mathbb{P}_0) = \beta(X, n; \mathbb{P}) \quad \text{for all } n \in \mathbb{N}. \tag{3.9}$$

(The analogous equality holds for $\alpha(\cdot, \cdot)$.)

(c) In (P3), it was implicitly understood in the phrase 'without loss of generality' that on a certain 'bad' event F with $\mathbb{P}(F) = 0$, one might need to redefine certain random variables Z_k, $k \in \mathbb{Z}$. By (3.6), $\mathbb{P}_0(F) = 0$ as well, and, hence, the phrase 'without loss of generality' applies under $\mathbb{P}_0$ as well as under $\mathbb{P}$.

(d) In [8] (with a finite alphabet), the probabilities involving mutations and insertions were specified directly. In [8] it was also pointed out how the context of mutation could include, as part of the model, 'high probability of no mutation'. Paragraph (P8) just gives an alternative way to set all that up, using the independent random variables U_k uniformly distributed on the interval [0, 1] as 'randomizers', and using appropriate choices of the Borel functions g (to determine mutations) and h (to determine insertions).

For example, suppose (again in the spirit of [8]) that the input sequence X is a DNA sequence, with the four nucleotides represented by the letters A, C, G, and T. To 'fit' (P2) and the subsequent paragraphs, one can respectively 'code' those letters as the numbers 1, 2, 3, and 4 in $(0, \infty)$. In (3.5), suppose that whenever an insertion is to occur, the respective probabilities of inserting the nucleotides A, C, G, and T (the numbers 1, 2, 3, and 4) are to be 0.15, 0.5, 0.25, and 0.1. One can set that up by using in (3.5) the function h on [0, 1] defined by

$$h(u) := \begin{cases} 1 & \text{if } 0 \le u < 0.15, \\ 2 & \text{if } 0.15 \le u < 0.65, \\ 3 & \text{if } 0.65 \le u < 0.9, \\ 4 & \text{if } 0.9 \le u \le 1. \end{cases}$$

Similarly, suppose that whenever a nucleotide A (the state 1) occurs, it is to be 'left alone' with probability 0.95 ('high probability of no mutation') or to mutate to C, G, or T (2, 3, or 4) with probability 0.01, 0.03, or 0.01, respectively. One can set that up by using in (3.5) a function g that satisfies

$$g(1, u) := \begin{cases} 1 & \text{if } 0 \le u < 0.95, \\ 2 & \text{if } 0.95 \le u < 0.96, \\ 3 & \text{if } 0.96 \le u < 0.99, \\ 4 & \text{if } 0.99 \le u \le 1. \end{cases}$$

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

One can similarly define $g(2, u)$, $g(3, u)$, and $g(4, u)$, $0 \le u \le 1$, in an appropriate way to model particular probabilities of specific mutations—or 'no mutation'—whenever the nucleotide C, G, or T (the state 2, 3, or 4) occurs.

4. The main result and its proof

This section is devoted to the proof of the following theorem, the main result of this paper.

Theorem 4.1. *Assume that the entire context of paragraphs (P1)–(P10) holds, with all assumptions there satisfied.*

(i) *Under the probability measure $\mathbb{P}_0$ in (3.6), the sequence Υ (in (P10)) is strictly stationary (and, hence, under $\mathbb{P}_0$, the sequences V and Y are each strictly stationary).*

(ii) *If also $\sum_{n=1}^{\infty} \beta(X, n; \mathbb{P}) < \infty$ (see also (3.9)) and $\sum_{n=1}^{\infty} \beta(V, n; \mathbb{P}_0) < \infty$, then $\sum_{n=1}^{\infty} \beta(Y, n; \mathbb{P}_0) \le \sum_{n=1}^{\infty} \beta(\Upsilon, n; \mathbb{P}_0) < \infty$.*

Theorem 4.1(i) is in spirit a slight extension of [8, Theorem 3.1], which in their setup was a corresponding 'preservation of strict stationarity' result involving a 'Markov' assumption on the MID sequence. Theorem 4.1(ii) was inspired by [8, Theorem 3.2], which in their setup was a corresponding 'preservation of mixing rate' result involving strong mixing with exponential mixing rate. It seems clear that the setup here in (P1)–(P10), involving two-sided random sequences, can facilitate the proofs of such 'preservation of mixing rate' results involving absolute regularity, such as Theorem 4.1(ii); but it is yet to be determined to what extent the setup here might facilitate such results involving strong mixing. The emphasis on summable mixing rates in Theorem 4.1(ii) is motivated by Theorem 2.1, the very sharp central limit theorem of Ibragimov involving summable mixing rates (for strong mixing); recall the comments immediately after that theorem.

We will first prove Theorem 4.1(i). The proof given below will be a somewhat modified version of the argument for [8, Theorem 3.1]. The argument will proceed through a series of lemmas. The first lemma is of a standard form. (In closely related contexts, a very similar fact was used in [3, Proof of Lemma 5] and [4, pp. 7–8]; see also [7, Theorem 26.4(I)].)

Lemma 4.1. *In the context of (P1)–(P10) (with all assumptions there satisfied), the sequence $((Z_k, \bar{X}_k),\ k \in \mathbb{Z})$ is, under the probability measure $\mathbb{P}$, strictly stationary.*

Sketch of the proof. Suppose that j is any integer. Define the integer-valued random variable $T := \max\{k \in \mathbb{Z}\colon \zeta_k \le j\}$. The entire array $((Z_k,\ k \ge j+1), (\bar{X}_k,\ k \ge j+1))$ can be represented as $\phi((Z_k,\ k \ge j+1), (X_{T+1}, X_{T+2}, X_{T+3}, \dots))$, where the (measurable) function $\phi\colon \{M, I, D\}^{\mathbb{N}} \times (0, \infty)^{\mathbb{N}} \to \{M, I, D\}^{\mathbb{N}} \times [0, \infty)^{\mathbb{N}}$ does not depend on j. Under $\mathbb{P}$, regardless of j, by the assumptions in (P2)–(P5) and an elementary argument, the sequence $(X_{T+1}, X_{T+2}, X_{T+3}, \dots)$ is independent of $\sigma(Z, T)$ $(= \sigma(Z))$ and has the same distribution as the sequence $(X_1, X_2, X_3, \dots)$. Lemma 4.1 then follows easily.

Lemma 4.2. *Suppose that $L \ge 3$ is an integer. Suppose that, for each $\ell \in \{1, 2, \dots, L\}$, $s_\ell \in \{M, I\}$, N_ℓ is a positive integer, and B_ℓ is a Borel subset of $[0, \infty)$. For each $J \in \{1, 2, \dots, L-1\}$, define the event (see (3.7) and (3.3))*

$$F_J := \{Z_0 \in \{M, I\}\} \cap \left[\bigcap_{\ell=1}^{L} (\{V_{-J+\ell} = (s_\ell, N_\ell)\} \cap \{\bar{Y}_{-J+\ell} \in B_\ell\})\right]. \tag{4.1}$$

Then $\mathbb{P}(F_1) = \mathbb{P}(F_2) = \cdots = \mathbb{P}(F_{L-1})$.

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

Proof. Define the integers $m_0 := 0$ and $m_\ell := N_1 + N_2 + \cdots + N_\ell$ for $\ell \in \{1, 2, \ldots, L\}$. The integers m_ℓ will sometimes be written below as $m(\ell)$. Define the set $S := \{1, 2, \ldots, m_L\} - \{m_1, m_2, \ldots, m_L\}$. Suppose that $J \in \{1, 2, \ldots, L-1\}$.

By (3.3), $\{Z_0 \in \{M, I\}\} = \{\xi_0 = 0\}$. As a consequence, we have the equality of events

$$\begin{aligned}
&\{Z_0 \in \{M, I\}\} \cap \left[\bigcap_{\ell=1}^{L} \{\xi_{-J+\ell} - \xi_{-J+\ell-1} = N_\ell\}\right] \\
&\quad = \bigcap_{\ell=0}^{L} \{\xi_{-J+\ell} = -m_J + m_\ell\} \\
&\quad = \left[\bigcap_{\ell=0}^{L} \{Z_{-m(J)+m(\ell)} \in \{M, I\}\}\right] \cap \left[\bigcap_{u \in S} \{Z_{-m(J)+u} = D\}\right]. \qquad (4.2)
\end{aligned}$$

Referring to (4.1) and applying both equalities in (4.2) carefully, we obtain

$$\begin{aligned}
F_J &= \{Z_0 \in \{M, I\}\} \\
&\quad \cap \left[\bigcap_{\ell=1}^{L} (\{Z_{\xi(-J+\ell)} = s_\ell\} \cap \{\xi_{-J+\ell} - \xi_{-J+\ell-1} = N_\ell\} \cap \{\bar{X}_{\xi(-J+\ell)} \in B_\ell\})\right] \\
&= \left[\bigcap_{\ell=0}^{L} \{\xi_{-J+\ell} = -m_J + m_\ell\}\right] \cap \left[\bigcap_{\ell=1}^{L} (\{Z_{-m(J)+m(\ell)} = s_\ell\} \cap \{\bar{X}_{-m(J)+m(\ell)} \in B_\ell\})\right] \\
&= \{Z_{-m(J)} \in \{M, I\}\} \cap \left[\bigcap_{\ell=1}^{L} (\{Z_{-m(J)+m(\ell)} = s_\ell\} \cap \{\bar{X}_{-m(J)+m(\ell)} \in B_\ell\})\right] \\
&\quad \cap \left[\bigcap_{u \in S} \{Z_{-m(J)+u} = D\}\right]. \qquad (4.3)
\end{aligned}$$

By Lemma 4.1, the probability (under $\mathbb{P}$) of the last expression in (4.3) does not depend on J ($\in \{1, 2, \ldots, L-1\}$). Thus, Lemma 4.2 holds.

Lemma 4.3. *The sequence* $((V_\ell, \bar{Y}_\ell),\ \ell \in \mathbb{Z})$ *of* $((\{M, I\} \times \mathbb{N}) \times [0, \infty))$*-valued random variables is strictly stationary under* $\mathbb{P}_0$.

Proof. Suppose that $j \in \mathbb{Z}$ and $n \in \mathbb{N}$. It suffices to prove that, under $\mathbb{P}_0$, the 'random vectors' $((V_\ell, \bar{Y}_\ell),\ \ell \in \{j+1, j+2, \ldots, j+n\})$ and $((V_\ell, \bar{Y}_\ell),\ \ell \in \{j+2, j+3, \ldots, j+n+1\})$ have the same distribution (on $(\{M, I\} \times \mathbb{N} \times [0, \infty))^n$).

Let J and L be positive integers such that $\{J-1, J\} \subset \{1, 2, \ldots, L-1\}$ and $\{j+1, j+2, \ldots, j+n\} \subset \{-J+1, -J+2, \ldots, -J+L\}$. It suffices to prove that, under $\mathbb{P}_0$, the random vectors $((V_\ell, \bar{Y}_\ell),\ \ell \in \{-J+1, -J+2, \ldots, -J+L\})$ and $((V_\ell, \bar{Y}_\ell),\ \ell \in \{-J+2, -J+3, \ldots, -J+L+1\})$ have the same distribution. However, this holds by (3.6), Lemma 4.2, and a trivial calculation. Thus, Lemma 4.3 holds.

4.1. Proof of Theorem 4.1(i)

By (P8) (see Remark 3.1(b)), the sequence U is, under $\mathbb{P}_0$, independent of the sequence $((V_\ell, \bar{Y}_\ell),\ \ell \in \mathbb{Z})$. It follows from (P8) and Lemma 4.3 (again see Remark 3.1(b)) that, under $\mathbb{P}_0$, the sequence $((V_\ell, \bar{Y}_\ell, U_\ell),\ \ell \in \mathbb{Z})$ is strictly stationary. Now Theorem 4.1(i) holds by (3.5) and (3.8).

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

The proof of Theorem 4.1(ii) will be based on the following lemma. In what follows, $\mathbb{E}_0(\cdot)$ denotes the expected value with respect to the probability measure $\mathbb{P}_0$. The indicator function of a given event A will be denoted by $\mathbf{1}(A)$.

Lemma 4.4. *In the context of (P1)–(P10) (with all the assumptions there satisfied), suppose also that* $\sum_{n=1}^{\infty} \beta(X, n; \mathbb{P}) < \infty$. *Define the sequence* $(H(n),\ n \in \{0\} \cap \mathbb{N})$ *of nonnegative numbers as follows. For each* $n \geq 0$, $H(n) := \beta(X, n+1; \mathbb{P})$.

Suppose that N *is an integer such that* $N \geq 2$. *Then*

$$\beta(\Upsilon, N; \mathbb{P}_0) \leq \beta(V, N; \mathbb{P}_0) + 2\mathbb{E}_0 H\left(\sum_{i=1}^{N-1} \mathbf{1}(Z_{\xi(i)} = M)\right).$$

Proof. The proof of this lemma will proceed through a series of steps.

Step 1. Refer to the integer $N \geq 2$ in the hypothesis (of Lemma 4.4). Define the nonnegative integer-valued random variable T by

$$T := \text{card}\{k \in \mathbb{N}\colon 1 \leq k \leq \xi_N - 1 \text{ and } Z_k \in \{M, D\}\} = \max\{j \in \{0\} \cup \mathbb{N}\colon \zeta_j < \xi_N\}. \quad (4.4)$$

(The second equality in (4.4) holds by (3.1).) Define the (one-sided) sequence $X^* := (X_1^*, X_2^*, X_3^*, \dots)$ of random variables as

$$X_k^* := X_{k+T} \quad \text{for all } k \geq 1. \quad (4.5)$$

Step 2. Let us first look at the random variable $\bar{Y}_N$. Suppose that $\omega \in \Omega$. If $Z_{\xi(N)(\omega)}(\omega) = I$ then $\xi_N(\omega) \notin \{\zeta_k(\omega)\colon k \in \mathbb{Z}\}$ by (3.1), and $\bar{Y}_N(\omega) = 0$ by (3.4) and (3.2). If instead $Z_{\xi(N)(\omega)}(\omega) = M$ (the only other possibility, by (3.3)) then, for some $q \geq 1$, $\xi_N(\omega) = \zeta_q(\omega)$; hence, $q = T(\omega) + 1$ by (4.4), and, hence, $\bar{Y}_N(\omega) = \bar{X}_{\zeta(q)(\omega)}(\omega) = X_q(\omega) = X_{T(\omega)+1}(\omega) = X_1^*(\omega)$ by (3.4), (3.2), and (4.5). Thus, $\bar{Y}_N = 0\,\mathbf{1}(Z_{\xi(N)} = I) + X_1^*\,\mathbf{1}(Z_{\xi(N)} = M)$. Hence, by (3.7),

$$\sigma(\bar{Y}_N) \subset \sigma(V_N, X_1^*). \quad (4.6)$$

Step 3. Suppose that ℓ is any integer such that $\ell > N$. Our task in step 3 is to obtain some sort of analog of (4.6) for $\bar{Y}_\ell$.

First define the random variable

$$\tau := \text{card}\{k \in \mathbb{N}\colon \xi_N \leq k \leq \xi_\ell \text{ and } Z_k \in \{M, D\}\}. \quad (4.7)$$

Then $\tau = [\sum_{i=N}^{\ell} \mathbf{1}(Z_{\xi(i)} = M)] + [\sum_{i=N+1}^{\ell} (\xi_i - \xi_{i-1} - 1)]$. For a given $\omega \in \Omega$, by (3.3), the first sum on the right-hand side is simply the number of indices k in the set in (4.7) such that $Z_k(\omega) = M$, and the second sum is simply the number of indices k in that set such that $Z_k(\omega) = D$. (Either sum can be 0.) From this expression for τ, we have, by (3.7),

$$\sigma(\tau) \subset \sigma(V_N, V_{N+1}, \dots, V_\ell). \quad (4.8)$$

Now suppose that $\omega \in \Omega$. Consider first the case where $Z_{\xi(\ell)(\omega)}(\omega) = M$. Then, for some $q \geq 1$, $\xi_\ell(\omega) = \zeta_q(\omega)$; and by (4.4), (4.7), and (3.1), $q = T(\omega) + \tau(\omega)$. Hence, by (3.4), (3.2), and (4.5), $\bar{Y}_\ell(\omega) = \bar{X}_{\zeta(q)(\omega)}(\omega) = X_q(\omega) = X_{\tau(\omega)}^*(\omega)$. Also, in the case where $Z_{\xi(\ell)(\omega)}(\omega) = M$, we have $\tau(\omega) \geq 1$ by (4.7). If instead $Z_{\xi(\ell)(\omega)}(\omega) = I$ (the only other possibility) then $\bar{Y}_\ell(\omega) = 0$ by (3.4) and (3.2).

Then (for our given $\ell > N$) putting all these pieces together,

$$\bar{Y}_\ell = 0\,\mathbf{1}(Z_{\xi(\ell)} = I) + X_\tau^*\,\mathbf{1}(Z_{\xi(\ell)} = M) = 0 + \sum_{t=1}^{\infty} [X_t^*\,\mathbf{1}(\tau = t)\,\mathbf{1}(Z_{\xi(\ell)} = M)],$$

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

and, hence, by (4.8) and (3.7),

$$\sigma(\bar{Y}_\ell) \subset \sigma(V_N, V_{N+1}, \dots, V_\ell) \vee \sigma(X_1^*, X_2^*, X_3^*, \dots). \tag{4.9}$$

Step 4. Combining (4.6) and (4.9), we now have

$$\sigma(\bar{Y}_N, \bar{Y}_{N+1}, \bar{Y}_{N+2}, \dots) \subset \sigma(V_N, V_{N+1}, V_{N+2}, \dots) \vee \sigma(X_1^*, X_2^*, X_3^*, \dots). \tag{4.10}$$

Define the two random arrays $\mathbb{A}$ and $\mathbb{B}$ as follows:

$$\mathbb{A} := ((X_k,\ k \le 0); (V_k,\ k \le 0); (U_k,\ k \le 0)), \tag{4.11}$$

$$\mathbb{B} := (X^*; (V_k,\ k \ge N); (U_k,\ k \ge N)). \tag{4.12}$$

By (3.5), $\sigma(Y_\ell) \subset \sigma(\bar{Y}_\ell, U_\ell)$ for each integer ℓ. We now have, by (3.8) and (4.10),

$$\sigma(\Upsilon_\ell, \ell \ge N) \subset \sigma(\mathbb{B}). \tag{4.13}$$

We need some sort of analog of (4.13) for $\sigma(\Upsilon_\ell,\ \ell \le 0)$ and $\sigma(\mathbb{A})$. To this end, we will need to work with the probability measure $\mathbb{P}_0$ (in (3.6)). Some more notation will be needed. For the events A and B, the notation $A \doteq B$ will mean that $\mathbb{P}_0(A \triangle B) = 0$, where '$\triangle$' denotes the symmetric difference. For an event A and a σ-field $\mathcal{B}$, the notation $A \dot{\in} \mathcal{B}$ will mean that $A \doteq B$ for some $B \in \mathcal{B}$. For σ-fields $\mathcal{A}$ and $\mathcal{B}$, the notation $\mathcal{A} \dot{\subset} \mathcal{B}$ will mean that $A \dot{\in} \mathcal{B}$ for every $A \in \mathcal{A}$, and the notation $\mathcal{A} \doteq \mathcal{B}$ will mean that $\mathcal{A} \dot{\subset} \mathcal{B}$ and $\mathcal{B} \dot{\subset} \mathcal{A}$.

Refer to (3.7) and both equalities in (3.6). For $\omega \in \{\xi_0 = 0\}$, the 'ordered pairs' $(V_k(\omega),\ k \le 0)$ determine ('measurably') the set of integer $\{\xi_k(\omega),\ k \le 0\}$ as well as $Z_j(\omega)$ (M or I) for j in that set, and, hence, determine $Z_j(\omega)$ for all $j \le 0$ (since $Z_j(\omega) = D$ for integers $j \le 0$ that are not in that set). Combining this with (3.3), we obtain $\sigma(V_k,\ k \le 0) \doteq \sigma(Z_k,\ k \le 0)$. Now $\sigma(\bar{X}_k,\ k \le 0) \subset \sigma(X_k, Z_k,\ k \le 0)$ by (3.1) and (3.2); hence, $\sigma(\bar{Y}_k,\ k \le 0) \subset \sigma(X_k, Z_k,\ k \le 0)$ by (3.3) and (3.4), and, hence, also $\sigma(Y_k,\ k \le 0) \subset \sigma(X_k, Z_k, U_k,\ k \le 0)$ by (3.5). Thus, by (4.11) and (3.8),

$$\sigma(\Upsilon_k, k \le 0) \dot{\subset} \sigma(\mathbb{A}). \tag{4.14}$$

Step 5. On the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let Γ be a random variable which (under $\mathbb{P}$) is uniformly distributed on the interval $[0, 1]$ and is independent of $\sigma(X, Z, U)$ (recall (P2), (P3), and (P8)). Now, of course, by the hypothesis (of Lemma 4.4), the sequence $\beta(X, n; \mathbb{P}) \to 0$ as $n \to \infty$ (absolute regularity). At this point, we will apply the 'coupling' result of Berbee [1, Theorem 4.4.7, p. 104, the Note, p. 106], which is closely related to the 'maximal coupling' result of Goldstein [11]. As a convenient reference for Berbee's result, we cite [6, Theorem 20.7, p. 277 and Lemma A1651, pp. 477–478]. The latter lemma in that reference simply involves the use of the random variable Γ above as a randomizer, and it is simply (a special case of) the version in Dudley and Philipp [10, Lemma 2.11] of a theorem of Skorohod [17]. Thereby there exists a sequence $X' := (X'_k,\ k \in \mathbb{Z})$ of random variables with the following properties (under $\mathbb{P}$):

$$\text{the distributions of the sequences } X \text{ and } X' \text{ are identical (under } \mathbb{P}), \tag{4.15}$$

$$\text{the } \sigma\text{-fields } \sigma(X') \text{ and } \sigma(X_k,\ k \le 0) \text{ are independent (under } \mathbb{P}), \tag{4.16}$$

$$\mathbb{P}\,(\text{there exist } k \ge n \text{ such that } X'_k \ne X_k) = \beta(X, n; \mathbb{P}) \text{ for all } n \in \mathbb{N}, \tag{4.17}$$

$$\sigma(X') \subset \sigma(X, \Gamma). \tag{4.18}$$

By (P1)–(P8) and the properties of the random variable Γ, under $\mathbb{P}$, the σ-fields $\sigma(\Gamma)$, $\sigma(X)$, $\sigma(Z)$, and $\sigma(U)$ are independent. By (3.6) and a trivial argument, this condition holds under

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

$\mathbb{P}_0$ as well. Furthermore, by (3.6) and a trivial argument, the distribution of the random array (Γ, X, X', U) (on $[0,1] \times \mathbb{R}^{\mathbb{Z}} \times \mathbb{R}^{\mathbb{Z}} \times [0,1]^{\mathbb{Z}}$) is the same under $\mathbb{P}_0$ as it is under $\mathbb{P}$. In particular, (4.15)–(4.17) all hold with $\mathbb{P}$ replaced with $\mathbb{P}_0$ (see also (3.9)). Thus, under $\mathbb{P}_0$, the following statements hold. By (4.18), the σ-fields $\sigma(X, X')$, $\sigma(Z)$, and $\sigma(U)$ are independent; hence, by (4.16), the σ-fields $\sigma(X')$, $\sigma(X_k,\ k \le 0)$, $\sigma(Z)$, and $\sigma(U)$ are independent, and, hence, the σ-field $\sigma(X')$ is independent of the σ-field $\sigma(U, Z) \vee \sigma(X_k,\ k \le 0)$.

Step 6. Now referring to (4.4), define analogously to (4.5) the (one-sided) sequence $X'^* := (X_1'^*, X_2'^*, X_3'^*, \dots)$ of random variables as $X_k'^* := X'_{k+T}$ for all $k \ge 1$.

Consider an arbitrary event $A \subset \sigma(U, Z) \vee \sigma(X_k,\ k \le 0)$ such that $\mathbb{P}_0(A) > 0$. For each integer $t \ge 0$ such that $\mathbb{P}_0(A \cap \{T = t\}) > 0$, we now have (note that $\sigma(T) \subset \sigma(Z)$)

$$\begin{aligned}\mathcal{L}_0(X'^* \mid A \cap \{T = t\}) &= \mathcal{L}_0((X'_{t+1}, X'_{t+2}, X'_{t+3}, \dots) \mid A \cap \{T = t\}) \\ &= \mathcal{L}_0(X'_{t+1}, X'_{t+2}, X'_{t+3}, \dots) \\ &= \mathcal{L}_0(X'_1, X'_2, X'_3, \dots), \end{aligned} \tag{4.19}$$

where $\mathcal{L}_0(\cdot)$ and $\mathcal{L}_0(\cdot \mid \cdot)$ respectively denote the distribution and conditional distribution under $\mathbb{P}_0$. Since the last term in (4.19) is 'constant' (not depending on A or t), it follows by a simple standard calculation that, for each such event A, $\mathcal{L}_0(X'^* \mid A) = \mathcal{L}_0(X'_1, X'_2, X'_3, \dots)$, and also (consider the case $A = \Omega$) $\mathcal{L}_0(X'^*) = \mathcal{L}_0(X'_1, X'_2, X'_3, \dots)$. Consequently, the sequence X'^* is (under $\mathbb{P}_0$) independent of the σ-field $\sigma(U, Z) \vee \sigma(X_k,\ k \le 0)$.

Analogously to (4.12), define the random array $\mathbb{B}'$ as follows:

$$\mathbb{B}' := (X'^*; (V_k,\ k \ge N); (U_k,\ k \ge N)). \tag{4.20}$$

Referring to the last sentence of the preceding paragraph and the third sentence after (4.18) (which together with (P8)(i) yields $\beta(U, N; \mathbb{P}_0) = 0$), we have, by (4.11), (4.20), Remark 3.1(b), and Lemma 2.1,

$$\beta(\sigma(\mathbb{A}), \sigma(\mathbb{B}'); \mathbb{P}_0) = \beta(V, N; \mathbb{P}_0). \tag{4.21}$$

Also, by (4.4), (4.12), (4.20), and (4.17) (and the fact that $\sigma(T) \subset \sigma(Z)$), with the sums below taken over all nonnegative integers t such that $\mathbb{P}_0(T = t) > 0$, we have (recall the sequence $H(\cdot)$ in the statement of Lemma 4.4)

$$\begin{aligned}&\mathbb{P}_0(\mathbb{B}' \ne \mathbb{B}) \\ &\quad = \mathbb{P}_0(X'^* \ne X^*) \\ &\quad = \sum \mathbb{P}_0(X'^* \ne X^* \mid T = t)\mathbb{P}_0(T = t) \\ &\quad = \sum \mathbb{P}_0((X'_{t+1}, X'_{t+2}, X'_{t+3}, \dots) \ne (X_{t+1}, X_{t+2}, X_{t+3}, \dots) \mid T = t)\mathbb{P}_0(T = t) \\ &\quad = \sum \mathbb{P}_0((X'_{t+1}, X'_{t+2}, X'_{t+3}, \dots) \ne (X_{t+1}, X_{t+2}, X_{t+3}, \dots))\mathbb{P}_0(T = t) \\ &\quad = \sum \beta(X, t+1; \mathbb{P}_0)\mathbb{P}_0(T = t) \\ &\quad = \mathbb{E}_0 H(T). \end{aligned}$$

Hence, by (4.11)–(4.14), (4.21), and Lemma 2.2 (and Theorem 4.1(i), proved above)

$$\begin{aligned}\beta(\Upsilon, N; \mathbb{P}_0) &\le \beta(\sigma(\mathbb{A}), \sigma(\mathbb{B}); \mathbb{P}_0) \\ &\le \beta(\sigma(\mathbb{A}), \sigma(\mathbb{B}'); \mathbb{P}_0) + 2\mathbb{E}_0 H(T) \\ &= \beta(V, N; \mathbb{P}_0) + 2\mathbb{E}_0 H(T). \end{aligned} \tag{4.22}$$

https://doi.org/10.1239/jap/1402578640 Published online by Cambridge University Press

Now, by (4.4), $T \geq \sum_{i=1}^{N-1} \mathbf{1}(Z_{\xi(i)} = M)$. Also, the sequence $H(n)$ (in the statement of Lemma 4.4) is nonincreasing as n increases. It follows that $\mathbb{E}_0 H(T) \leq \mathbb{E}_0 H(\sum_{i=1}^{N-1} \mathbf{1}(Z_{\xi(i)} = M))$. Combining this with (4.22) we obtain Lemma 4.4.

4.2. Proof of Theorem 4.1(ii)

Define the sequence $\Theta := (\mathbf{1}(Z_{\xi(i)} = M),\ i \in \mathbb{Z})$ of random indicator functions. Under $\mathbb{P}_0$, this sequence is strictly stationary by (3.7) and Theorem 4.1(i) (proved above). By (3.6), (P3), and a trivial argument, the sequence Θ is also nondegenerate under $\mathbb{P}_0$. Also, by (3.7) and the hypothesis (of Theorem 4.1(ii)), we have $\sum_{n=1}^{\infty} \beta(\Theta, n; \mathbb{P}_0) < \infty$. Also, by the hypothesis (of Theorem 4.1(ii)), the (nonincreasing) sequence $(H(n),\ n \in \{0\} \cap \mathbb{N})$ of nonnegative numbers in Lemma 4.4 is summable. Hence, for the sequence $H(\cdot)$, by Lemma 2.3, $\sum_{n=2}^{\infty} \mathbb{E}_0 H(\sum_{i=1}^{n-1} \mathbf{1}(Z_{\xi(i)} = M)) < \infty$. Hence, by (3.8), Lemma 4.4, and the hypothesis (of Theorem 4.1(ii)), the conclusion of Theorem 4.1(ii) holds. This completes the proof.

Acknowledgement

The author thanks the anonymous referee for a correction and for helpful comments that improved the exposition.

References

[1] Berbee, H. C. P. (1979). *Random Walks with Stationary Increments and Renewal Theory*. Mathematical Centre, Amsterdam.
[2] Billingsley, P. (1995). *Probability and Measure*, 3rd edn. John Wiley, New York.
[3] Bradley, R. C. (1980). A remark on the central limit question for dependent random variables. *J. Appl. Prob.* **17,** 94–101.
[4] Bradley, R. C. (1989). A stationary, pairwise independent, absolutely regular sequence for which the central limit theorem fails. *Prob. Theory Relat. Fields* **81,** 1–10.
[5] Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*, Vol. 1. Kendrick Press, Heber City, UT.
[6] Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*, Vol. 2. Kendrick Press, Heber City, UT.
[7] Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*, Vol. 3. Kendrick Press, Heber City, UT.
[8] Chaudhuri, P. and Dasgupta, A. (2006). Stationarity and mixing properties of replicating character strings. *Statistica Sinica* **16,** 29–43.
[9] Davydov, Y. A. (1973). Mixing conditions for Markov chains. *Theory Prob. Appl.* **18,** 312–328.
[10] Dudley, R. M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrscheinlichkeitsth* **62,** 509–552.
[11] Goldstein, S. (1979). Maximal coupling. *Z. Wahrscheinlichkeitsth* **46,** 193–204.
[12] Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.
[13] Merlevède, F. and Peligrad, M. (2000). The functional central limit theorem under the strong mixing condition. *Ann. Prob.* **28,** 1336–1352.
[14] Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco.
[15] Rio, E. (2000). *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants* (Math. Appl. **31**). Springer, Berlin.
[16] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42,** 43–47.
[17] Skorohod, A. V. (1976). On a representation of random variables. *Theory Prob. Appl.* **21,** 628–632.
[18] Volkonskii, V. A. and Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theory Prob. Appl.* **4,** 178–197.
[19] Waterman, M. S. (1995). *Introduction to Computational Biology*. Chapman and Hall, New York.