The Asymptotics of Waiting Times Between Stationary Processes, Allowing Distortion

Short title: **Waiting Times Allowing Distortion**

BY AMIR DEMBO[1] AND IOANNIS KONTOYIANNIS[2]

Stanford University

Appeared in *Annals of Applied Probability,* vol. 9, pp. 413– 429, May 1999

Given two independent realizations of the stationary processes $\boldsymbol{X} = \{X_n \ ; \ n \geq 1\}$ and $\boldsymbol{Y} = \{Y_n \ ; \ n \geq 1\}$, our main quantity of interest is the waiting time $W_n(D)$ until a D-close version of the initial string $(X_1, X_2, \ldots, X_n)$ first appears as a contiguous substring in $(Y_1, Y_2, Y_3, \ldots)$, where closeness is measured with respect to some "average distortion" criterion.

We study the asymptotics of $W_n(D)$ for large n under various mixing conditions on $\boldsymbol{X}$ and $\boldsymbol{Y}$. We first prove a strong approximation theorem between $\log W_n(D)$ and the logarithm of the probability of a D-ball around $(X_1, X_2, \ldots, X_n)$. Using large deviations techniques, we show that this probability can, in turn, be strongly approximated by an associated random walk, and we conclude that: (*i*) $n^{-1} \log W_n(D)$ converges almost surely to a constant R determined by an explicit variational problem; (*ii*) $[\log W_n(D) - R]$, properly normalized, satisfies a central limit theorem, a law of the iterated logarithm, and, more generally, an almost sure invariance principle.

[1]partially supported by NSF grant #DMS-9403553.

[2]partially supported by grants NSF #NCR-9628193, JSEP #DAAH04-94-G-0058, ARPA #J-FBI-94-218-2.

AMS 1991 *subject classification.* Primary—60F15; secondary—60F10, 94A17.

Key words and phrases. Waiting times, string matching, large deviations, relative entropy, strong approximation, almost sure invariance principle.

1. Introduction and Main Results. The problem of analyzing the asymptotic behavior of waiting times between stationary processes has received a lot of attention in the literature over the past few years (see Wyner and Ziv (1989), Shields (1993), Szpankowski (1993), Marton and Shields (1995), Kontoyiannis (1998) and the references therein), primarily because of its important applications in several fields, most notably in data compression and the analysis of string matching algorithms in DNA sequence analysis. These applications are outlined in the next section.

Let $\boldsymbol{X} = \{X_n \; ; \; n \geq 1\}$ and $\boldsymbol{Y} = \{Y_n \; ; \; n \geq 1\}$ be two processes taking values in the Polish spaces $(A_X^\infty, \mathcal{F}_X)$ and $(A_Y^\infty, \mathcal{F}_Y)$, and distributed according to the probability measures P and Q, respectively. We will assume throughout the paper that the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. By $x = (x_1, x_2, \ldots) \in A_X^\infty$ we denote an infinite realization of $\boldsymbol{X}$, and for $1 \leq i \leq j \leq \infty$ we write x_i^j for the substring $(x_i, x_{i+1}, \ldots, x_j)$. Similarly we write X_i^j for the vector $(X_i, \ldots, X_j)$, and likewise for $\boldsymbol{Y}$.

Given a measurable function $\rho(\cdot,\cdot) : A_X \times A_Y \to [0, \infty)$, the "distortion" between two finite strings $x_1^n \in A_X^n$ and $y_1^n \in A_Y^n$ is measured by:

$$\rho_n(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^n \rho(x_i, y_i). \tag{1}$$

For $x_1^n \in A_X^n$ and $D \geq 0$ we write $B(x_1^n, D)$ for the ball of radius D around x_1^n:

$$B(x_1^n, D) = \{y_1^n \in A_Y^n \; : \; \rho_n(x_1^n, y_1^n) \leq D\}.$$

Given $D \geq 0$ and two independent infinite realizations x, y from $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, our main quantity of interest is the *waiting time* $W_n(D)$ until a D-close version of x_1^n first appears in y:

$$W_n(D) \;=\; W_n(x_1^n, y, D) \;=\; \inf\{k \geq 1 \; : \; y_k^{k+n-1} \in B(x_1^n, D)\}.$$

In the special case where A_X and A_Y are finite sets and W_n stands for the first time an *exact* copy of the string x_1^n appears in y, it is known that W_n increases exponentially with n,

$$\frac{1}{n} \log W_n \to R \quad P \times Q - \text{a.s.}, \tag{2}$$

when $\boldsymbol{X}$ is stationary ergodic and $\boldsymbol{Y}$ satisfies certain mixing conditions (Wyner and Ziv (1989), Shields (1993), Marton and Shields (1995), Kontoyiannis (1998)); here and throughout the paper log denotes the natural logarithm. The constant R can be expressed in terms of relative entropy; for example, when $\boldsymbol{X}$ is composed of independent and identically distributed random variables (an "i.i.d. process") with marginal distribution P_1, and $\boldsymbol{Y}$ is an i.i.d. process with marginal Q_1, then $R = R(P_1, Q_1) = H(P_1) + H(P_1|Q_1)$, where $H(P_1) = E[-\log P(X_1)]$ is the entropy of $\boldsymbol{X}$ and $H(\cdot|\cdot)$ denotes the relative entropy between two probability measures:

$$H(\mu|\nu) = \begin{cases} \int d\mu \log \frac{d\mu}{d\nu}, & \text{when} \frac{d\mu}{d\nu} \text{ exists} \\ \infty, & \text{otherwise.} \end{cases}$$

Moreover, under more restrictive conditions on the mixing properties of $\boldsymbol{X}$ and $\boldsymbol{Y}$, it is known that $[\log W_n - nR]$ satisfies a central limit theorem (CLT) (Wyner (1993)), and a law of the iterated logarithm (LIL), as well as the functional counterparts of these results (Kontoyiannis (1998)).

Our purpose in this paper is to extend these asymptotic results to $W_n(D)$ (see Corollaries 1 through 4, below). Little has been done in this direction: Recently, Yang and Kieffer (1998) showed that (2) holds for $W_n(D)$ when A_X and A_Y are finite sets, with $R = R(P_1, Q_1, D)$ given as the solution to a variational problem in terms of relative entropy (see Theorem 2 below). Related results were obtained by Łuczak and Szpankowski (1997), but neither of these papers addressed the problem of determining the second-order asymptotic properties of $\log W_n(D)$, and also left open the question of whether analogous results can be established for general spaces A_X and A_Y. In this paper we address both of these issues.

The first step in our analysis (carried out in Theorem 1) is to show that the waiting time $W_n(D)$ until a D-close match for X_1^n occurs in $\boldsymbol{Y}$ is approximately equal to the reciprocal of the probability $Q(B(X_1^n, D))$ that such a match indeed occurs. In the case when no distortion is allowed, $Q(B(X_1^n, D))$ simply reduces to $Q(X_1^n)$, and applying the Shannon-McMillan-Breiman Theorem and its second-order refinements one gets a complete picture of the asymptotic behavior of W_n (cf. Kontoyiannis (1998)). But when distortion is allowed, the asymptotic behavior (particularly the second-order behavior) of the probabilities $Q(B(X_1^n, D))$ is not quite obvious *a priori*. The novelty in the approach we employ here is the use of large deviations techniques to obtain corresponding results for $Q(B(X_1^n, D))$ in place of $Q(X_1^n)$: Theorems 2 and 3 relate $Q(B(X_1^n, D))$ to an associated random walk on $\mathbb{R}$ induced by X_1^n, and they provide natural generalizations of the Shannon-McMillan-Breiman theorem and its subsequent refinements [by Ibragimov (1962) and by Philipp and Stout (1975)] for processes with values in general spaces, and to the case when distortion is allowed.

Our first result is a strong approximation theorem stating that the waiting time $W_n(D)$ is asymptotically almost surely close to the reciprocal of the probability $Q(B(X_1^n, D))$:

Theorem 1. *Suppose $\boldsymbol{Y}$ is a stationary process with ϕ-mixing coefficients that satisfy $\sum \phi(k) < \infty$, and assume that $Q(B(X_1^n, D)) > 0$ eventually $P-$a.s. If $\{c(n)\}$ is an arbitrary sequence of non-negative constants such that $\sum n e^{-c(n)} < \infty$, then*

$$|\log[W_n(D)Q(B(X_1^n, D))]| \;\leq\; c(n) \quad \textit{eventually } P{\times}Q{-}a.s.$$

It will be evident from the proof of Theorem 1 that the result remains valid for general sequences of distortion measures $\{\rho_n\}$, not necessarily of the form of (1), under mild regularity conditions.

Recall that the ϕ-mixing coefficients of $\boldsymbol{Y}$ are defined by $\phi(k) = \sup\{|Q(B|A) - Q(B)|\}$ where the supremum is taken over all integers $r \geq 1$ and all pairs of events A and B such that $B \in \sigma(Y_{r+k}^{\infty})$, $A \in \sigma(Y_1^r)$, and $Q(A) \neq 0$; see Bradley (1986) for an extensive discussion.

From Theorem 1 we get that

$$\log W_n(D) - [-\log Q(B(X_1^n, D))] \;=\; o(\sqrt{n}) \quad P{\times}Q - \text{a.s.} \tag{3}$$

In contrast with the case of exact matching (i.e., when no distortion is allowed), here, $-\log Q(B(X_1^n, D))$ cannot be readily expanded as the partial sum of the logarithms of conditional probabilities. Nevertheless, we can relate $-\log Q(B(X_1^n, D))$ to a different random walk, which arises as a functional of the empirical measure $\hat{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ induced on A_X by X_1^n (Theorems 2 and 3). From that, we can read off the exact asymptotic behavior of $-\log Q(B(X_1^n, D))$, and, via (3), the behavior of the waiting times $W_n(D)$ (Corollaries 1 through 4).

Let

$$D_{\min} = \; E_P[\operatorname*{ess\,inf}_{Y_1} \; \rho(X_1, Y_1)],$$

and for simplicity, assume hereafter that $\boldsymbol{Y}$ is an i.i.d. process, and that

$$D_{\max} = \operatorname*{ess\,sup}_{(X_1,Y_1)} \rho(X_1,Y_1) \;\in\; (D_{\min},\infty).$$

For $\boldsymbol{X}$ stationary and ergodic, by the ergodic theorem, $W_n(D) = 1$ eventually $P\times Q-$almost surely, for any $D > D_{\rm av} = E\rho(X_1,Y_1)$, whereas $W_n(D) = \infty$ eventually $P\times Q-$almost surely, for any $D < D_{\min}$. Of interest is the range $D \in (D_{\min}, D_{\rm av})$ where $W_n(D)$ exhibits exponential behavior.

Theorem 2. *Let $\boldsymbol{X}$ be a stationary ergodic process and $\boldsymbol{Y}$ be an i.i.d. process. Then for $D \in (D_{\min}, D_{\rm av})$ we have,*

$$-\log Q(B(X_1^n,D)) - nR(\hat{P}_n) \;=\; o(\sqrt{n}) \quad P-a.s.,$$

where $R(\hat{P}_n) = R(\hat{P}_n, Q_1, D)$ is defined by the following variational problem:

$$R(\hat{P}_n, Q_1, D) \;=\; \inf \int H(\nu(\cdot|x)|Q_1(\cdot))d\hat{P}_n(x),$$

and the infimum is taken over all probability measures ν on $A_X\times A_Y$ such that the A_X-marginal of ν is $\hat{P}_n$ and $\int \rho(x,y)\,d\nu(x,y) \le D$.

See Proposition 1 in Section 3 for an alternative characterization of $R(\hat{P}_n, Q_1, D)$. An easy consequence of Theorem 2 is the following generalization of (2).

Corollary 1. *Assume that $\boldsymbol{X}$ is stationary ergodic, $\boldsymbol{Y}$ is an i.i.d. process, and $D \in (D_{\min}, D_{\rm av})$. Then $R(\hat{P}_n) \to R(P_1)$ $P-$almost surely, and hence*

$$\frac{1}{n}\log W_n(D) \to R(P_1, Q_1, D) \quad P\times Q - a.s.$$

Next we investigate the behavior of $\sqrt{n}[R(\hat{P}_n) - R(P_1)]$. As it turns out (see Proposition 1 in Section 4), the function $R(P_1) = R(P_1, Q_1, D)$ is the convex dual of the log-moment generating function $\Lambda_{P_1}(\lambda)$, where, for any probability measure μ on A_X and any $\lambda \in \mathbb{R}$, $\Lambda_\mu(\lambda)$ is defined as

$$\Lambda_\mu(\lambda) \;=\; \int \log\left\{\int e^{\lambda\rho(x,y)}dQ_1(y)\right\} d\mu(x).$$

Write $\Lambda(\cdot) = \Lambda_{P_1}(\cdot)$ when $\mu = P_1$, $\Lambda_x(\cdot) = \Lambda_{\delta_x}(\cdot)$ for any $x \in A_X$, and $\bar{\Lambda}_{X_i}(\cdot) = \Lambda_{X_i}(\cdot) - \int \Lambda_x(\cdot) dP_1(x)$. Theorem 3 provides an explicit approximation of $\sqrt{n}[R(\hat{P}_n) - R(P_1)]$ by a random walk induced by X_1^n. [Recall that the α-mixing coefficients of $\boldsymbol{X}$ are defined by $\alpha(k) = \sup\{|P(A \cap B) - P(A)P(B)| \ ; \ A \in \sigma(X_1^r),\ B \in \sigma(X_{r+k}^\infty),\ r \geq 1\}$; see Bradley (1986) for details.]

Theorem 3. *Let $\boldsymbol{X}$ be a stationary process with α-mixing coefficients that satisfy $\sum \alpha(k) < \infty$, let $\boldsymbol{Y}$ be an i.i.d. process, and $D \in (D_{\min}, D_{\text{av}})$. Then for $\lambda = \lambda(D) < 0$ such that $\Lambda'(\lambda) = D$ we have*

$$n[R(\hat{P}_n) - R(P_1)] + \sum_{i=1}^n \bar{\Lambda}_{X_i}(\lambda) \ = \ o(\sqrt{n}) \quad P - a.s.$$

In particular, combining (3) with Theorems 2 and 3 gives

$$[\log W_n(D) - nR(P_1, Q_1, D)] + \sum_{i=1}^n \bar{\Lambda}_{X_i}(\lambda) \ = \ o(\sqrt{n}) \quad P{\times}Q - \text{a.s.}, \tag{4}$$

and it is now straightforward to harvest a series of corollaries. The following is an immediate consequence of combining (4) with well-known CLT results (see, for example, Theorem 1.7 in Peligrad (1986)).

Corollary 2 (CLT). *Let $\boldsymbol{X}$ be a stationary process with α-mixing coefficients such that $\sum \alpha(k) < \infty$, let $\boldsymbol{Y}$ be an i.i.d. process, and $D \in (D_{\min}, D_{\text{av}})$. Then, for $\lambda = \lambda(D)$, the following series converges,*

$$\sigma^2 \ = \ E_P\left\{\bar{\Lambda}_{X_1}(\lambda)^2\right\} + 2\sum_{k=2}^\infty E_P\left\{\bar{\Lambda}_{X_1}(\lambda)\bar{\Lambda}_{X_k}(\lambda)\right\}, \tag{5}$$

and

$$\frac{\log W_n(D) - nR(P_1)}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, \sigma^2).$$

Moreover, when $\sigma^2 > 0$, the sequence of processes,

$$\left\{ \frac{w(nt; D)}{\sigma\sqrt{n}} \ ; \ t \in [0,1] \right\}, \qquad n \geq 1,$$

converges in distribution to standard Brownian motion, where $w(t; D) = [\log W_{\lfloor t \rfloor}(D) - \lfloor t \rfloor R(P_1, Q_1, D)]$ for $t \geq 1$, and $w(t; D) = 0$ for $t < 1$.

Similarly, Corollary 3 is a consequence of (4) combined with the LIL (Rio (1995)).

Corollary 3 (LIL). *Let $\boldsymbol{X}$ be a stationary process with α-mixing coefficients such that $\sum \alpha(k) < \infty$, $\boldsymbol{Y}$ be an i.i.d. process, and $D \in (D_{\min}, D_{\mathrm{av}})$. Then, for σ^2 as in (5), with $P{\times}Q-$probability one the set of limit points of the sequence*

$$\left\{ \frac{\log W_n(D) - nR(P_1)}{\sqrt{2n \log\log n}} \right\}, \quad n \geq 3$$

coincides with the interval $[-\sigma, \sigma]$. Moreover, when $\sigma^2 > 0$, with $P{\times}Q-$probability one, the sequence of sample paths

$$\left\{ \frac{w(nt; D)}{\sqrt{2n\ \log\log n}}\ ;\ t \in [0,1] \right\}, \quad n \geq 3,$$

is relatively compact in the topology of uniform convergence on $D[0,1]$, and the set of its limit points is the collection of all absolutely continuous functions $r : [0,1] \to \mathbb{R}$, such that $r(0) = 0$ and $\int_0^1 (dr/dt)^2 dt \leq \sigma^2$.

Finally, Corollary 4 follows from (4) and an almost sure invariance principle proved by Philipp and Stout (1975), Theorem 4.1.

Corollary 4 (Almost sure invariance principle). *Let $\boldsymbol{X}$ be a stationary process with ϕ-mixing coefficients that satisfy $\sum \sqrt{\phi(k)} < \infty$, $\boldsymbol{Y}$ be an i.i.d process, and $D \in (D_{\min}, D_{\mathrm{av}})$. Then, with $\sigma^2 > 0$ as in (5), there exists a Brownian motion $\{B(t)\ ;\ t \geq 0\}$ such that*

$$w(t; D) - \sigma B(t) \;=\; o(\sqrt{t}) \quad P{\times}Q - a.s., \tag{6}$$

As usual we interpret (6) as saying that, without changing its distribution, $w(t; D)$ can be redefined on a richer probability space that contains a Brownian motion such that (6) holds. For some of the numerous corollaries that can be derived from almost sure invariance principles as the one in (6) see Strassen (1964) and Ch. 1 of Philipp and Stout (1975).

Remark 1. In Corollary 1, $W_n(D)$ can be replaced by $1/Q(B(X_1^n, D))$ to give a natural generalization of the Shannon-McMillan-Breiman theorem (analogous to the one obtained by Yang and Kieffer for finite sets A_X, A_Y) for the case when distortion is allowed, and for processes with values in general spaces. In a similar fashion, from Corollaries 2 and 4 we can obtain corresponding generalizations of Ibragimov's (1962) CLT-refinement, and Philipp and Stout's (1975) almost sure invariance principle, respectively.

Remark 2. Similar results as those obtained for the waiting times $W_n(D)$ can also be obtained for the sequence of *recurrence times* $R_n(D)$: Given $D \geq 0$ and a realization x from a doubly-infinite process $\boldsymbol{X} = \{X_n \; ; \; n \in \mathbb{Z}\}$, $R_n(D)$ is defined as the first time a D-close version of x_{-n}^{-1} appears in x_0^{∞}:

$$R_n(D) \;=\; R_n(x, D) \quad = \quad \inf\{k \geq 0 \,:\, x_k^{k+n-1} \in B(x_{-n}^{-1}, D)\}.$$

Theorems 2 and 3 remain valid in this case with X_1^n replaced by X_{-n}^{-1} and $Q = P$, which forces us to assume that $\boldsymbol{X}$ is an i.i.d. process. Under this assumption it is easy to see that Theorem 1 also remains essentially unchanged, so that, combining Theorems 1, 2 and 3 as before, we recover the exact same asymptotic behavior for $R_n(D)$ as that for $W_n(D)$ (Corollaries 1 through 4).

In the next section we outline two areas of applications of our results, in Section 3 we prove Theorem 1, in Section 4 we prove our main results, Theorems 2 and 3, and in Section 5 we prove Theorem 4.

2. Applications.

In this section we outline two potential applications of our results about the asymptotic behavior of $W_n(D)$.

Data Compression. The analysis of several data compression schemes based on string matching, such as the celebrated Lempel-Ziv algorithm, is typically reduced to studying the following idealized scenario (see Wyner and Ziv (1989, 1991), Steinberg and Gutman (1993), the discussion in Yang and Kieffer (1998), Łuczak and Szpankowski (1997) and the references therein): An encoder and a decoder have available to them a common infinite "database" $y = (y_1, y_2, \ldots)$ generated by an i.i.d. process $\boldsymbol{Y} \sim Q$, and the encoder's task is to communicate the "message" $x_1^n = (x_1, x_2, \ldots, x_n)$ to the decoder, within some prescribed accuracy D with respect to a sequence $\{\rho_n\}$ of distortion measures of the form of (1). This is done as follows; the encoder scans the database until a D-close version of x_1^n is found in y, and then "tells" the decoder the position $W_n(D)$ where this match occurs. To describe $W_n(D)$ it takes $\log W_n(D) + O(\log\log W_n(D))$ nats (or bits, if the logarithms are taken to be base-2), and therefore the limiting compression ratio of the code in nats-per-symbol (by Corollary 1) is given by

$$\frac{\log W_n(D) + O(\log\log W_n(D))}{n} \;\to\; R(P_1, Q_1, D) \quad \text{a.s.}$$

For example, in the case of lossless coding of an i.i.d. "message source" $\boldsymbol{X}$, $R(P_1, Q_1, 0)$ reduces to $H(P_1) + H(P_1|Q_1)$, which is interpreted as the optimal limiting compression ratio $H(P_1)$, plus the additional "penalty" term $H(P_1|Q_1)$ induced by the fact that the database was generated by the sub-optimal distribution Q instead of P. Similarly, in the case of lossy coding we may choose to generate the database y according to the product measure Q for which $R(P_1, Q_1, D)$ is minimal; for an i.i.d. process $\boldsymbol{X}$ the limiting compression ratio of this code, $r(D) = \inf_{Q_1} R(P_1, Q_1, D)$, equals the optimal compression ratio, namely the rate-distortion function of $\boldsymbol{X}$ with respect to $\{\rho_n\}$ (see Berger (1971) for details).

Once the compression ratio is identified, from Corollaries 2, 3 and 4 we get further information about the rate at which it is achieved (the "redundancy" of the code), about the limiting distribution of the size of the encoded data, and so on.

DNA Sequence Analysis. In the analysis of DNA or protein sequences the following problem is of interest (see Karlin and Ost (1988), Pevzner, Borodovsky and Mironov (1991), Arratia and Waterman (1994) and the references therein): Given a template $x_1, x_2, \ldots$ and a long but finite "database" sequence y_1^m, find the longest contiguous substring in the database that matches an initial portion x_1^ℓ of the template within accuracy D, with respect to the average of some score function $\rho(\cdot,\cdot)$. The length $L_m(D)$ of the longest such match is of interest here:

$$L_m(D) = L_m(x, y, D) = \sup\{n \geq 1 \ : \ y_j^{j+n-1} \in B(x_1^n, D), \ \text{ for some } \ j = 1, 2, \ldots, m\}.$$

Clearly, there is a duality relationship between $L_m(D)$ and $W_n(D)$: $L_m(D) \geq n$ if and only if $W_k(D) \leq m$ for some $k \geq n$. This relationship is exploited in the last section, where we read off the asymptotics of $L_m(D)$ from the corresponding results for $W_n(D)$, explicitly identifying the asymptotic mean, variance and distribution of $L_m(D)$:

Theorem 4. *(i) Under the assumptions of Corollary 1,*

$$\frac{L_m(D)}{\log m} \to \frac{1}{R(P_1, Q_1, D)} \qquad P{\times}Q - a.s.$$

(ii) Under the assumptions of Corollary 2, with $\sigma^2 > 0$ as in (5) and $R = R(P_1, Q_1, D)$,

$$\frac{L_m(D) - \frac{\log m}{R}}{\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0, \sigma^2 R^{-3})$$

(iii) Under the assumptions of Corollary 3, with $\sigma^2 > 0$ as in (5) and $R = R(P_1, Q_1, D)$,

$$\limsup_{m\to\infty} \frac{L_m(D) - \frac{\log m}{R}}{\sqrt{2\log m\,\log\log\log m}} = \sigma R^{-3/2} \quad P{\times}Q - a.s.$$

3. Strong Approximation.

Proof of Theorem 1. Write $\mathbf{P}$ for the product measure $P{\times}Q$, and for each integer $m \geq 1$ let $G_m = \{x : \; Q(B(x_1^n, D)) > 0 \text{ for all } n \geq m\}$.

For the upper bound we use a standard second-moment blocking argument (similar to the one by Yang and Kieffer). Choose and fix any integer $m \geq 1$, pick an arbitrary $x \in G_m$, and let $n \geq m$ be large enough so that $e^{c(n)} \geq n+1$. Let $K \geq n+1$ and write $S_n = \sum_{j=0}^{V(K,n)} I_n(j)$, where $I_n(j)$ is the indicator function of the event $\{Y_{jn+1}^{(j+1)n} \in B(x_1^n, D)\}$, and $V(K,n) = \lfloor (K-1)/n \rfloor$. Then

$$\mathbf{P}(W_n(D) > K \,|\, X_1^n = x_1^n) \;\leq\; Q(S_n = 0) \;\leq\; \frac{\mathrm{Var}_Q(S_n)}{(E_Q S_n)^2}. \tag{7}$$

By stationarity,

$$E_Q S_n = [V(K,n)+1] Q(B(x_1^n, D)) \tag{8}$$

and $E_Q(I_n(0) I_n(j)) \leq Q(B(x_1^n, D))[\phi((j-1)n+1) + Q(B(x_1^n, D))]$, so that

$$\begin{aligned} \mathrm{Var}_Q(S_n) &= \sum_{j,k=0}^{V(K,n)} \mathrm{Cov}_Q(I_n(j), I_n(k)) \\ &\leq [V(K,n)+1] Q(B(x_1^n, D)) \left[1 + 2 \sum_{j=1}^{V(K,n)} \phi((j-1)n+1) \right]. \end{aligned} \tag{9}$$

Writing $\Phi = 1 + 2\sum \phi(k)$, and substituting (8) and (9) in (7) we get

$$\mathbf{P}(W_n(D) > K \,|\, X_1^n = x_1^n) \;\leq\; \frac{\Phi}{[V(K,n)+1]\, Q(B(x_1^n, D))}. \tag{10}$$

Choosing $K = e^{c(n)}/Q(B(x_1^n, D))$ we have $[V(K,n)+1]Q(B(x_1^n, D)) > e^{c(n)}/2n$, and (10) yields

$$\mathbf{P}(\log[W_n(D) Q(B(X_1^n, D))] > c(n) \,|\, X_1^n = x_1^n) \;\leq\; 2\Phi n e^{-c(n)}.$$

Since the above bound is uniform over $x \in G_m$ and summable, by the Borel-Cantelli lemma we obtain that

$$\log[W_n(D)Q(B(x_1^n, D))] \;\leq\; c(n) \quad \text{eventually for } P{\times}Q\text{−almost all } (x,y) \in G_m \times A_Y^\infty. \tag{11}$$

For the lower bound, we observe that for an arbitrary constant $K > 1$ and any $x \in G_m$,

$$\mathbf{P}(W_n(D) < K \,|\, X_1^n = x_1^n) \;\leq\; \sum_{j=1}^{\lfloor K \rfloor} Q(Y_j^{j+n-1} \in B(x_1^n, D)) \;\leq\; K\, Q(B(x_1^n, D)). \tag{12}$$

Since $W_n(D) \geq 1$, this inequality holds also for $K \in [0,1]$. In particular, setting $K = e^{-c(n)}/Q(B(x_1^n, D))$ gives

$$\mathbf{P}(\log[W_n(D)Q(B(X_1^n, D))] < -c(n) \,|\, X_1^n = x_1^n) \leq e^{-c(n)},$$

and summing this over n, by the Borel-Cantelli lemma we get

$$\log[W_n(D)Q(B(X_1^n, D))] \;\geq\; -c(n) \quad \text{eventually for } P{\times}Q\text{−almost all } (x,y) \in G_m \times A_Y^\infty. \tag{13}$$

Finally, combining (11) and (13) with the assumption that $P\{\cup_m G_m\} = 1$ completes the proof. □

4. Large Deviations. Lemma 1 below provides some easily checked facts needed in the proofs of Theorems 2 and 3. The variational characterization of the rate function R in terms of relative entropy is established next in Proposition 1, and the proofs of Theorem 2, Corollary 1 and Theorem 3 are given.

Lemma 1. *Let μ be an arbitrary probability measure on A_X, $\lambda \in \mathbb{R}$, and define $0 \leq D^\mu_{\min} < D^\mu_{\text{av}} < D^\mu_{\max} < \infty$ like $D_{\min}$, D_{av} and $D_{\max}$, respectively, with $X_1 \sim \mu$.*

(i) $|\Lambda_\mu(\lambda)| \leq |\lambda| D^\mu_{\max}$.

(ii) The Fenchel-Legendre transform of Λ_μ,

$$\Lambda^*_\mu(x) = \sup_{\lambda \in \mathbb{R}} [\lambda x - \Lambda_\mu(\lambda)]$$

exists and is finite for all $x \in (D^\mu_{\min}, D^\mu_{\text{av}})$.

(iii) $\Lambda_\mu \in C^\infty$, $\Lambda'_\mu(0) = D^\mu_{\text{av}}$, $\Lambda''_\mu(\lambda) > 0$ *for all* $\lambda \in \mathbb{R}$, *and* $\Lambda'_\mu(\lambda) \downarrow D^\mu_{\min}$ *as* $\lambda \to -\infty$.

(*iv*) *For each* $D \in (D^{\mu}_{\min}, D^{\mu}_{\rm av})$, *there exists a unique* $\lambda < 0$ *such that* $\Lambda'_{\mu}(\lambda) = D$, *and* $\Lambda^*_{\mu}(D) = \lambda D - \Lambda_{\mu}(\lambda)$.

(*v*) *For* μ*-almost any* $x \in A_X$, $\Lambda_x \in C^{\infty}$, *and its derivatives are uniformly bounded over* μ*-almost all* $x \in A_X$ *and all* λ *in a compact subset of* $\mathbb{R}$.

Proposition 1. *In the notation of Lemma 1, let* μ *be an arbitrary probability measure on* A_X *and* $D \in (D^{\mu}_{\min}, D^{\mu}_{\rm av})$. *Then,* $R(\mu, Q_1, D) = \Lambda^*_{\mu}(D)$, *i.e.,*

$$\inf \int H(\nu(\cdot|x)|Q_1(\cdot))d\mu(x) \;=\; \sup_{\lambda\in\mathbb{R}} \left[\lambda D - \int \log\left\{\int e^{\lambda\rho(x,y)}dQ_1(y)\right\} d\mu(x)\right], \tag{14}$$

where the infimum is taken over all probability measures ν *on* $A_X{\times}A_Y$ *such that the* A_X*-marginal of* ν *is* μ *and* $\int \rho(x,y)\, d\nu(x,y) \le D$.

Proof of Proposition 1. By Lemma 1 we may fix $\lambda < 0$ for which the supremum on the right side of (14) is achieved. Consider the probability measure ν defined by

$$\frac{d\nu(x,y)}{d\mu{\times}Q_1} \;=\; \frac{d\nu(y|x)}{dQ_1} \;=\; \frac{e^{\lambda\rho(x,y)}}{\int e^{\lambda\rho(x,z)}dQ_1(z)}$$

in the left side of (14). The A_X-marginal of ν is μ, $\int \rho(x,y)d\nu(x,y) = \Lambda'_{\mu}(\lambda) = D$, and

$$\int H(\nu(\cdot|x)|Q_1(\cdot))d\mu(x) \;=\; \lambda D - \int \log\left[\int e^{\lambda\rho(x,y)}dQ_1(y)\right] d\mu(x) \;=\; \Lambda^*_{\mu}(D),$$

and hence the left side of (14) is no greater than $\Lambda^*_{\mu}(D)$. To prove the reverse inequality we recall that for any probability measure ν and any bounded measurable function $\phi : A_Y \to \mathbb{R}$,

$$H(\nu(\cdot|x)|Q_1(\cdot)) \ge \int \phi(y)d\nu(y|x) - \log\left\{\int e^{\phi(y)}dQ_1(y)\right\}$$

(c.f. Lemma 3.2.13 in Deuschel and Stroock (1989)). In particular, choosing $\phi(\cdot) = \lambda\rho(x,\cdot)$ and then integrating both sides with respect to μ yields the required inequality and completes the proof. □

Proof of Theorem 2. Let $D^{(n)}_{\rm av} = \int \rho(x,y)\, d\hat{P}_n(x)dQ_1(y)$, so that, by the ergodic theorem,

$$D^{(n)}_{\rm av} \to D_{\rm av} \quad P-\text{a.s.} \tag{15}$$

Similarly let $D_{\min}^{(n)} = E_{\hat{P}_n}[\operatorname{ess\,inf}_{Y_1}\ \rho(X_1, Y_1)]$, so that

$$D_{\min}^{(n)} \to D_{\min} \quad P-\text{a.s.} \tag{16}$$

Given a realization of the $\boldsymbol{X}$ process such that both (15) and (16) hold, for n large enough the given D will be strictly between $D_{\min}^{(n)}$ and $D_{\text{av}}^{(n)}$, so by Lemma 1 we can choose, for each n, a negative λ_n such that $\Lambda'_{\hat{P}_n}(\lambda_n) = D$, $\Lambda^*_{\hat{P}_n}(D) = \lambda_n D - \Lambda_{\hat{P}_n}(\lambda_n)$, and $\Lambda''_{\hat{P}_n}(\lambda_n) > 0$. We similarly choose $\lambda < 0$ such that $\Lambda'(\lambda) = D$, and claim that

$$\lambda_n \to \lambda \quad P-\text{a.s.} \tag{17}$$

To see this suppose, for example, that $\liminf_{n\to\infty} \lambda_n \le \lambda - \epsilon$, for some $\epsilon > 0$, so that $\lambda_{n_k} \le \lambda - \epsilon/2$ for some $n_k \to \infty$. Then, by the ergodic theorem and the strict monotonicity of Λ' we get a contradiction:

$$D = \liminf_{n\to\infty} \Lambda'_{\hat{P}_n}(\lambda_n) \le \limsup_{n\to\infty} \Lambda'_{\hat{P}_n}(\lambda - \epsilon/2) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \Lambda'_{X_i}(\lambda - \epsilon/2) = \Lambda'(\lambda - \epsilon/2) < \Lambda'(\lambda) = D.$$

The case $\limsup_{n\to\infty} \lambda_n > \lambda$ is ruled out similarly.

Before we move to the main part of the proof, we need to show that

$$\Lambda''_{\hat{P}_n}(\lambda_n) \to \Lambda''(\lambda) > 0 \quad P-\text{a.s.} \tag{18}$$

Writing

$$|\Lambda''_{\hat{P}_n}(\lambda_n) - \Lambda''(\lambda)| \le \frac{1}{n}\sum_{i=1}^{n} |\Lambda''_{X_i}(\lambda_n) - \Lambda''_{X_i}(\lambda)| + \left| \frac{1}{n}\sum_{i=1}^{n} \Lambda''_{X_i}(\lambda) - \Lambda''(\lambda) \right|, \tag{19}$$

we can bound the first term above, for any $\epsilon > 0$ and n large enough, by

$$\operatorname*{ess\,sup}_{X_1} |\Lambda''_{X_1}(\lambda_n) - \Lambda''_{X_1}(\lambda)| \le |\lambda_n - \lambda| \operatorname*{ess\,sup}_{X_1} \sup_{\lambda-\epsilon\le\xi\le\lambda+\epsilon} |\Lambda'''_{X_1}(\xi)|$$

and this converges to zero, by (17) and part (v) of Lemma 1. As for the second term of (19), by the ergodic theorem it converges to zero, $P-$almost surely.

Now choose and fix a realization $\{x_i\}$ of $\boldsymbol{X}$ such that the statements (15) (16) (17) and (18) all hold. Define $\zeta_i = \rho(x_i, Y_i)$, $T_n = \sum_{i=1}^{n} \zeta_i$, and $\hat{T}_n = T_n/n$, with μ_n denoting the law of ζ_1^n. With a slight abuse of notation we write $\hat{P}_n$ for the (non-random, since x_1^∞ is fixed) empirical measure induced by x_1^n on A_X. In this notation, $Q(B(x_1^n, D)) = \Pr(\hat{T}_n \leq D)$, and, if we define

$$J_n \;=\; e^{n\Lambda^*_{\hat{P}_n}(D)} \Pr(\hat{T}_n \leq D),$$

then in view of Proposition 1 the statement of the theorem can be rephrased as

$$\log J_n = o(\sqrt{n}) \quad P - \text{a.s.} \tag{20}$$

The upper-bound part of (20) follows from

$$\begin{aligned} J_n &= e^{n\Lambda^*_{\hat{P}_n}(D)} E\left\{1_{\{\hat{T}_n \leq D\}}\right\} \;\leq\; e^{n\Lambda^*_{\hat{P}_n}(D)} E\left\{e^{n\lambda_n(\hat{T}_n - D)}\right\} \\ &= e^{n[\Lambda^*_{\hat{P}_n}(D) - \lambda_n D]} E\left\{e^{\lambda_n T_n}\right\} \;=\; 1 \end{aligned}$$

(by the choice of λ_n and the definition of $\Lambda_{\hat{P}_n}$).

Turning to the proof of the lower bound, suppose n is large enough so that λ_n exists, and define a new probability measure ν_n by

$$\frac{d\nu_n}{d\mu_n}(z_1^n) = \exp\left\{\lambda_n \sum_{i=1}^n z_i - n\Lambda_{\hat{P}_n}(\lambda_n)\right\}.$$

Let

$$G_n = -\frac{\sum_{i=1}^n [\zeta_i - E_{\nu_n}\zeta_i]}{\sqrt{n\Lambda''_{\hat{P}_n}(\lambda_n)}}, \quad \text{when } \zeta_1^n \sim \nu_n.$$

It is easy to see that G_n is a partial sum process of zero mean random variables, normalized so that $\mathrm{Var}(G_n) = 1$. Observe that when ζ_1^n is distributed according to ν_n,

$$\hat{T}_n \stackrel{\mathcal{D}}{=} D - \sqrt{\frac{\Lambda_{\hat{P}_n}{}''(\lambda_n)}{n}}\, G_n,$$

so that we can expand

$$\begin{aligned} J_n &= e^{n\Lambda^*_{\hat{P}_n}(D)} E_{\nu_n}\left\{1_{\{\hat{T}_n \leq D\}} e^{-n\lambda_n \hat{T}_n + n\Lambda_{\hat{P}_n}(\lambda_n)}\right\} \\ &= E_{\nu_n}\left\{1_{\{G_n \geq 0\}} e^{\lambda_n \sqrt{n\Lambda''_{\hat{P}_n}(\lambda_n)}\, G_n}\right\} \\ &\geq E_{\nu_n}\left\{1_{\{0 < G_n < \delta\}} e^{-\beta_n \sqrt{n}\, G_n}\right\} \\ &\geq e^{-\beta_n \sqrt{n}\,\delta} \Pr_{\nu_n}(0 < G_n < \delta), \end{aligned} \tag{21}$$

for any $\delta > 0$, and where $\beta_n = -\lambda_n \sqrt{\Lambda''_{\hat{P}_n}(\lambda_n)} > 0$ and $\beta_n = O(1)$, by (17) and (18).

Since the random variables ζ_i are uniformly bounded, and also $\Lambda''_{\hat{P}_n}(\lambda_n)$ is bounded away from zero by (18), it is easy to check that the Lindeberg condition for the CLT is satisfied by G_n, from which it follows that the probability $\Pr_{\nu_n}(0 < G_n < \delta) \to \rho > 0$ as $n \to \infty$. Now choose $M > 0$ large enough so that $M - \beta_n$ is bounded away from zero, and get from (21) that

$$\liminf_{n\to\infty} \log\left[e^{M\sqrt{n}\delta} J_n\right] \geq \log \rho > -\infty,$$

i.e.,

$$\liminf_{n\to\infty} \sqrt{n}\left[M\delta + \frac{1}{\sqrt{n}} \log J_n\right] > -\infty,$$

from which we conclude that

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}} \log J_n \geq -M\delta.$$

Since $\delta > 0$ was arbitrary and $M > 0$ was chosen independent of δ, letting $\delta \downarrow 0$ completes the proof. □

Proof of Corollary 1. Since $D > D_{\min}$, by (16) also $D > D_{\min}^{(n)}$ eventually $P-$almost surely. Consequently, $Q(B(X_1^n, D)) > 0$ eventually $P-$almost surely. Thus, Corollary 1 follows by combining Theorem 2 with (3), provided we show that $R(\hat{P}_n) \to R(P_1)$ almost surely, or, equivalently (by Proposition 1), that $\Lambda^*_{\hat{P}_n}(D) \to \Lambda^*(D)$ almost surely. Recall that for all n large enough $\Lambda^*_{\hat{P}_n}(D) = \lambda_n D - \Lambda_{\hat{P}_n}(\lambda_n)$ and $\Lambda^*(D) = \lambda D - \Lambda(\lambda)$, as in the proof of Theorem 2, where $\lambda_n \to \lambda$ almost surely by (17). So we only have to show that $\Lambda_{\hat{P}_n}(\lambda_n) \to \Lambda(\lambda)$ which comes from an obvious adaptation of the derivation of (18). □

Proof of Theorem 3. Let λ and $\{\lambda_n\}$ be chosen as in the beginning of the proof of Theorem 2, so that, in particular, $\Lambda^*(\lambda) = \lambda D - \Lambda(\lambda)$ and $\Lambda''(\lambda) > 0$. By the continuity of Λ'' we can choose constants $\delta, \eta > 0$ such that $\Lambda''(\lambda + \theta) > \eta$ whenever $|\theta| < \delta$. Also, from (17), we can pick $N = N(X_1^\infty) < \infty$ $P-$almost surely, such that $|\lambda_n - \lambda| < \delta$ for all $n \geq N$.

In view of Proposition 1 it suffices to show that

$$\sqrt{n}\left\{[\Lambda^*_{\hat{P}_n}(D) - \Lambda^*(D)] - [\Lambda(\lambda) - \Lambda_{\hat{P}_n}(\lambda)]\right\} \to 0 \quad . \tag{22}$$

From the definition of $\Lambda^*_{\hat{P}_n}$ and our choice of N, $\Lambda^*_{\hat{P}_n}(D)$ is given by the supremum of $[\theta D - \Lambda_{\hat{P}_n}(\theta)]$ over all $\theta \in (\lambda - \delta, \lambda + \delta)$, so (22) is the same as

$$\sqrt{n} \sup_{|\theta|<\delta} \left[\theta D - \Lambda_{\hat{P}_n}(\theta + \lambda) + \Lambda_{\hat{P}_n}(\lambda)\right] \to 0 \tag{23}$$

Since this supremum is always non-negative (take $\theta = 0$), (23) is equivalent to

$$\liminf_{n\to\infty} \sqrt{n} \inf_{|\theta|<\delta} \frac{1}{n} \sum_{i=1}^{n} [f(\theta, X_i) - f(0, X_i)] \geq 0 \ , \tag{24}$$

where $f(\theta, x) = \Lambda_x(\lambda + \theta) - (\lambda + \theta)D$. By Taylor's theorem we can expand $g(\theta) = n^{-1}\sum_{i=1}^n f(\theta, X_i)$ around $\theta = 0$, to obtain

$$\frac{1}{n}\sum_{i=1}^{n}[f(\theta, X_i) - f(0, X_i)] = \theta A_n + \frac{\theta^2}{2} B_n(\theta) \ , \tag{25}$$

where $A_n = n^{-1}\sum_{i=1}^n f'(0, X_i)$ and $B_n(\theta) = \frac{1}{n}\sum_{i=1}^n f''(\xi_n, X_i)$ for some $\xi_n(\theta)$ such that $|\xi_n| < \delta$.

The family of functions $\{f''(\xi, \cdot) \ ; \ \xi \in (-\delta, \delta)\}$ is uniformly bounded and equicontinuous (by Lemma 1), so by the uniform ergodic theorem (Rao (1962), Section 6),

$$\sup_{|\xi|<\delta} \left| \frac{1}{n} \sum_{i=1}^{n} f''(\xi, X_i) - E_P f''(\xi, X_1) \right| \to 0 \quad P-\text{a.s.}$$

Therefore, $P-$almost surely, by the choice of δ,

$$\begin{aligned} \liminf_{n\to\infty} \inf_{|\theta|<\delta} B_n(\theta) \ &\geq \ \liminf_{n\to\infty} \left\{ \inf_{|\xi|<\delta} E_P f''(\xi, X_1) - \sup_{|\xi|<\delta} \left| \frac{1}{n}\sum_{i=1}^{n} f''(\xi, X_i) - E_P f''(\xi, X_1) \right| \right\} \\ &\geq \ \inf_{|\xi|<\delta} E_P f''(\xi, X_1) = \inf_{|\xi|<\delta} \Lambda''(\lambda + \xi) \geq \eta > 0 \ . \end{aligned} \tag{26}$$

By our choice of λ, we have $E_P f'(0, X_1) = \Lambda'(\lambda) - D = 0$, so A_n is the partial sum corresponding to the zero-mean stationary process $\{f'(0, X_n) \ ; \ n \geq 1\}$. Since $\sum \alpha(k) < \infty$ and the random variables $f'(0, X_i)$ are bounded, the LIL (Rio (1995)) implies that $\sqrt{n} A_n^2 \to 0$ $P-$almost surely. Since the infimum over $|\theta| < \delta$ of the right side of (25) is bounded below by $-A_n^2 / \inf_{|\theta|<\delta} B_n(\theta)$, combining this with (26) gives (24) and completes the proof. □

5. Duality: Match Lengths. Let R denote $R(P_1, Q_1, D)$. Define $T_n(D) = \inf_{k \geq n} W_k(D)$ and $\tilde{T}_n(D) = \min_{n \leq k \leq 2n} W_k(D)$. As mentioned in Section 2, there is a duality relationship between $T_n(D)$ and $L_m(D)$:

$$L_m(D) \geq n \quad \Longleftrightarrow \quad T_n(D) \leq m\ . \tag{27}$$

When combined with Lemma 2 below, (27) allows us to deduce (i), (ii) and (iii) in Theorem 4 from corresponding results for $\tilde{T}_n(D)$, namely, in the notation, and under the corresponding assumptions, of Theorem 4:

$$
\begin{array}{ll}
(i') & \dfrac{\log \tilde{T}_n(D)}{n} \to R \quad P{\times}Q - \text{a.s.} \\
(ii') & \dfrac{\log \tilde{T}_n(D) - nR}{\sqrt{n}} \stackrel{\mathcal{D}}{\longrightarrow} N(0, \sigma^2) \\
(iii') & \liminf\limits_{n\to\infty} \dfrac{\log \tilde{T}_n(D) - nR}{\sqrt{2n\, \log\log n}} \;=\; -\sigma \quad P{\times}Q - \text{a.s.}
\end{array}
$$

Lemma 2. *Assume that* $\boldsymbol{X}$ *is stationary ergodic,* $\boldsymbol{Y}$ *is an i.i.d. process, and* $D \in (D_{\min}, D_{\text{av}})$. *Then,* $T_n(D) = \tilde{T}_n(D)$ *eventually* $P{\times}Q$*-almost surely.*

Proof of Lemma 2. Note that $T_n(D) \leq \tilde{T}_n(D) \leq W_n(D)$ and that $T_n(D) = \tilde{T}_n(D)$ whenever $T_{2n}(D) > W_n(D)$. Therefore, if

$$\liminf_{n\to\infty} n^{-1} \log T_{2n}(D) \geq \frac{4R}{3} \quad P{\times}Q - \text{a.s.} \tag{28}$$

then, by Corollary 1, $T_n(D) = \tilde{T}_n(D)$ eventually $P{\times}Q$-almost surely. For any $x_1^\infty \in A_X^\infty$, for any positive integer m and any n large enough, by the union bound and (12),

$$\mathbf{P}(T_{2n}(D) \leq m \,|\, X_1^\infty = x_1^\infty) \leq \sum_{k\geq 2n} \mathbf{P}(W_k(D) \leq m \,|\, X_1^k = x_1^k) \leq m \sum_{k \geq 2n} Q(B(x_1^k, D)). \tag{29}$$

It follows from Theorem 2 and Corollary 1 that, with P-probability one, $\lim_{k\to\infty} k^{-1} \log Q(B(X_1^k, D)) = -R$. In particular, eventually P-almost surely, $\sup_{k\geq n} k^{-1} \log Q(B(X_1^k, D)) \leq -3R/4$. Substituting this in (29) with $m = \exp(4Rn/3)$ gives,

$$\mathbf{P}(T_{2n}(D) \leq \exp(4Rn/3) \,|\, X_1^\infty = x_1^\infty) \;\leq\; Ce^{-nR/6} \quad \text{eventually } P\text{-a.s.},$$

for some fixed $C < \infty$. Hence, by the Borel-Cantelli lemma, $T_{2n}(D) > \exp(4Rn/3)$ eventually $P{\times}Q$-almost surely, implying (28) and the conclusion of the lemma. $\square$

Proof of Theorem 4. As already stated, it suffices to prove (i') – (iii'). To this end, first observe that combining Theorem 1 and Theorem 2,

$$\liminf_{n\to\infty} \frac{1}{n} \min_{n\le k\le 2n} \left[\log W_k(D) - kR(\hat{P}_k)\right] \ge 0 \quad P{\times}Q - \text{a.s.}, \tag{30}$$

and from Corollary 1 it follows that

$$\frac{1}{n} \min_{n\le k\le 2n} kR(\hat{P}_k) \to R \quad P{\times}Q - \text{a.s.} \tag{31}$$

By (30) and (31) we have

$$\frac{1}{n}\log \tilde{T}_n(D) \ge \frac{1}{n} \min_{n\le k\le 2n} \left[\log W_k(D) - kR(\hat{P}_k)\right] + \frac{1}{n} \min_{n\le k\le 2n} kR(\hat{P}_k) \to R, \quad P{\times}Q - \text{a.s.}$$

Since $\tilde{T}_n(D) \le W_n(D)$, the corresponding upper bound also holds by Corollary 1, proving (i').

Next let $\epsilon > 0$ arbitrary, so that in the notation of Corollary 2,

$$\mathbf{P}\left\{\frac{\log \tilde{T}_n(D)}{\sqrt{n}} - \frac{\log W_n(D)}{\sqrt{n}} < -\epsilon\right\} = \mathbf{P}\left\{\inf_{1\le t\le 2}\left[\frac{w(nt;D)}{\sigma\sqrt{n}} - \frac{w(n;D)}{\sigma\sqrt{n}} + \left(\frac{\lfloor nt\rfloor - n}{\sigma\sqrt{n}}\right)R\right] \le -\frac{\epsilon}{\sigma}\right\}.$$

For any $\delta > 0$ and n large enough this is bounded above by

$$\mathbf{P}\left\{\inf_{1\le t\le 1+\delta}\left[\frac{w(nt;D)}{\sigma\sqrt{n}} - \frac{w(n;D)}{\sigma\sqrt{n}}\right] \le -\frac{\epsilon}{\sigma}\right\} + \mathbf{P}\left\{\inf_{1+\delta\le t\le 2}\left[\frac{w(nt;D)}{\sigma\sqrt{n}} - \frac{w(n;D)}{\sigma\sqrt{n}}\right] \le -\frac{\epsilon}{\sigma} - K\sqrt{n}\right\}, \tag{32}$$

where $K = \delta R/(2\sigma)$. By the functional CLT of Corollary 2 (extended in the obvious way to $t \in [0,2]$), the first term of (32) converges, as $n \to \infty$, to $\Pr\{\inf_{0\le t\le\delta} B_t \le -\epsilon/\sigma\}$, where $\{B_t\}$ is standard Brownian motion, and this can be made arbitrarily small by taking δ small enough. Similarly for any $C > 0$ the second term in (32) is asymptotically bounded above by $\Pr\{\inf_{0\le t\le 1} B_t \le -C\}$ which can also be made arbitrarily small by taking C large enough. Combining these with the fact that $\tilde{T}_n(D) \le W_n(D)$ implies that $[\log \tilde{T}_n(D) - \log W_n(D)] = o(\sqrt{n})$ in probability, which, together with Corollary 2, gives (ii').

We similarly obtain (iii') by applying the functional LIL instead of the functional CLT: Set $s_n = \sigma\sqrt{2n\log\log n}$ noting that,

$$\frac{\log \tilde{T}_n(D)}{s_n} - \frac{\log W_n(D)}{s_n} = \inf_{1\le t\le 2}\left[\frac{w(nt;D)}{s_n} - \frac{w(n;D)}{s_n} + \left(\frac{\lfloor nt\rfloor - n}{s_n}\right)R\right].$$

For any $\delta > 0$ and n large enough this is bounded below by

$$\min\left\{\inf_{1\le t\le 1+\delta}\left[\frac{w(nt;D)}{s_n}-\frac{w(n;D)}{s_n}\right],\ \inf_{1+\delta\le t\le 2}\left[\frac{w(nt;D)}{s_n}-\frac{w(n;D)}{s_n}\right]+K\sqrt{\frac{n}{\log\log n}}\right\}. \tag{33}$$

By the functional LIL of Corollary 3 (extended in the obvious way to $t \in [0, 2]$), the first term in (33) is asymptotically $P\times Q$-almost surely bounded below by

$$\inf_r \inf_{1\le t\le 1+\delta}[r(t)-r(1)] \ge -\sqrt{\delta},$$

where the outermost infimum is taken over all absolutely continuous functions r with $\int_0^2 (dr/dt)^2 dt \le 1$ and $r(0) = 0$. Similarly,

$$\liminf_{n\to\infty}\inf_{1+\delta\le t\le 2}\left[\frac{w(nt;D)}{s_n}-\frac{w(n;D)}{s_n}\right] \ge \inf_r \inf_{1+\delta\le t\le 2}[r(t)-r(1)] \ge -\sqrt{1-\delta} \quad P\times Q - \text{a.s.}\ ,$$

so that the second term in (33) converges to $+\infty$ with probability one, and hence

$$\liminf_{n\to\infty}\frac{\log \tilde{T}_n(D)}{\sigma\sqrt{2n\log\log n}}-\frac{\log W_n(D)}{\sigma\sqrt{2n\log\log n}} \ge -\sqrt{\delta} \quad P\times Q - \text{a.s.}$$

Letting $\delta \downarrow 0$, recalling that $\tilde{T}_n(D) \le W_n(D)$ and applying Corollary 3 gives (iii') and completes the proof. □

Acknowledgment. The authors thank W. Szpankowski for providing a preprint of Łuczak and Szpankowski (1997) and for his comments on an earlier version of the manuscript, and Dirk Tasche for pointing out the relevance of Rio (1995) in the context of Theorem 3.

References

Arratia, R. and Waterman, M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.* **4** 200–225.

Berger, T. (1971). *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice Hall.

Bradley, B. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics* (E. Eberlein and M. S. Taqqu, eds) 165–192.

Deuschel, J. D. and Stroock, D. W. (1989). *Large Deviations.* Academic Press, Boston.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory Prob. Appl.* **7** 349–382.

Karlin, S. and Ost, F. (1988). Maximal length of common words among random letter sequences. *Ann. Probab.* **16** 535–563.

Kontoyiannis, I. (1998). Asymptotic recurrence and waiting times for stationary processes. *J. Theor. Probab.* **11** 795–811.

Łuczak, T. and Szpankowski, W. (1997). A suboptimal lossy data compression based on approximate pattern matching. *IEEE Trans. Inform. Theory* **43** 1439–1451.

Marton, K. and Shields, P. C. (1995). Almost-sure waiting time results for weak and very weak Bernoulli processes. *Ergod. Th. & Dynam. Sys.* **15** 951–960.

Peligrad, M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey). In *Dependence in Probability and Statistics* (E. Eberlein and M. S. Taqqu, eds) 193–223.

Pevzner, P., Borodovsky, M. and Mironov, A. (1991). Linguistic of nucleotide sequences: the significance of deviations from mean statistical characteristics and prediction of the frequency of occurrence of words. *J. Biomol. Struct. Dynam.* **6** 1013–1026.

Philipp, W. and Stout, W. (1975). *Almost sure invariance principles for partial sums of weakly dependent random variables.* Memoirs of the AMS, vol. 2, issue 2, no. 161.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Stat.* **33** 659–680.

Rio, E. (1995). The functional law of the iterated logarithm for stationary strongly mixing sequences. *Ann. Probab.* **23** 1188–1203.

Shields, P. C. (1993). Waiting times: positive and negative results on the Wyner-Ziv problem. *J. Theor. Probab.* **6** 499–519.

Steinberg, Y. and Gutman, M. (1993). An algorithm for source coding subject to a fidelity criterion based on string matching. *IEEE Trans. Inform. Theory* **39** 877–886.

Strassen, V. (1964). An almost sure invariance principle for the law of the iterated logarithm. *Z. Wahrsch. Verw. Gabiete* **3** 23–32.

Szpankowski, W. (1993). Asymptotic properties of data compression and suffix trees. *IEEE Trans. Inform. Theory* **39** 1647–1659.

Wyner, A. J. (1993). *String matching theorems and applications to data compression and statistics.* Ph.D. Thesis, Dept. of Statistics, Stanford University.

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. Inform. Theory* **35** 1250–1258.

Wyner, A. D. and Ziv, J. (1991). Fixed data base version of the Lempel-Ziv data compression algorithm. *IEEE Trans. Inform. Theory* **37** 878–880.

Yang, E.-H. and Kieffer, J. C. (1998). On the performances of data compression algorithms based upon string matching. *IEEE Trans. Inform. Theory* **44** 47–65.

Department of Statistics
Stanford University
Stanford, CA 94305-4065
amir@stat.stanford.edu

Department of Statistics
Purdue University
W. Lafayette, IN 47907
yiannis@stat.purdue.edu