THE ASYMPTOTICS OF WAITING TIMES BETWEEN STATIONARY PROCESSES

BY
IOANNIS KONTOYIANNIS AND AMIR DEMBO
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 96
JUNE 1997

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION
GRANT NCR-9628193

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
The Asymptotics of Waiting Times Between Stationary Processes

BY

IOANNIS KONTOYIANNIS

AND

AMIR DEMBO

Stanford University

June 19, 1997

Given two independent realizations of the stationary processes $X = \{X_n ; n \geq 1\}$ and $Y = \{Y_n ; n \geq 1\}$, 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 by some distortion criterion.

We study the asymptotics of $W_n(D)$ for large $n$ under various mixing conditions on $X$ and $Y$. 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.

\footnote{AMS 1991 subject classification. Primary—60F10, 60F15; secondary—60F05, 60F17, 94A17.}
\footnote{Key words and phrases. Waiting times, string matching, large deviations, relative entropy, strong approximation, strong, weak convergence, mixing processes, functional limit theorems, almost sure invariance principle.}
\footnote{I.K. was partially supported by grants NSF #NCR-9205663, JSEP #DAAH04-94-G-0058, ARPA #J-FBI-94-218-2.}
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), Marton and Shields (1995), Kontoyiannis (1996) 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 \( X = \{ X_n ; n \geq 1 \} \) and \( Y = \{ Y_n ; n \geq 1 \} \) be two processes taking values in the state-spaces \( A_X \) and \( A_Y \), and distributed according to the measures \( P \) and \( Q \), respectively. By \( x \in A_X^\infty \), \( x = (x_1, x_2, \ldots) \), we denote an infinite realization of \( X \), 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 \( Y \) and \( Q \).

Consider a sequence of non-negative distortion measures \( d_n(\cdot, \cdot) \) on \( A_X^n \times A_Y^n \), \( n \geq 1 \); given a finite string \( x_1^n \in A_X^n \) and a positive constant \( D \), write \( B(x_1^n, D) \) for the ball of radius \( D \) around \( x_1^n \):

\[
B(x_1^n, D) = \{ y_1^n \in A_Y^n : d_n(x_1^n, y_1^n) \leq D \}.
\]

For example, a sequence \( \{d_n\} \) of single-letter distortion measures is defined by fixing a non-negative distortion measure \( d \) on \( A_X \times A_Y \), and letting

\[
d_n(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^n d(x_i(y_i), y_1^n \in A_X^n, y_1^n \in A_Y^n.
\]

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

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

In the case of no distortion and when \( A_X \) and \( A_Y \) are finite sets (so that \( W_n \) 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 \( n \),

\[
\frac{1}{n} \log W_n \rightarrow R(P, Q) \quad \text{P} \times \text{Q} \text{-a.s.,}
\]

when \( X \) is stationary ergodic and \( Y \) satisfies certain mixing conditions (Wyner and Ziv (1989), Shields (1993), Marton and Shields (1995)); here and throughout the paper log denotes the natural logarithm.

The rate in the exponent \( R(P, Q) \) can be expressed in terms of relative entropy; for example, when \( X \) and \( Y \) are both independent and identically distributed (i.i.d.) sequences with marginals \( P_1 \) and \( Q_1 \), respectively, then \( R(P, Q) = H(P_1) + H(P_1|Q_1) \), where \( H(P_1) = E[-\log P(X_1)] \) is the entropy of \( X \) 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 \(X\) and \(Y\), 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 (1996)).

Our purpose in this paper is to extend these asymptotic results to \(W_n(D)\) instead of \(W_n\) (Corollaries 2, 3 and 4 below). Little has been done in this direction: Recently, Yang and Kieffer (preprint) showed that (1) holds for \(W_n(D)\) when \(X\) and \(Y\) have finite state spaces and \(\{d_n\}\) is a sequence of single-letter distortion measures, with a rate function \(R(P, Q, D)\) which is given as the solution to a variational problem in terms of relative entropy (see Theorem 2 below). Similar results were obtained by Luczak 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 processes with general state-spaces. In this paper we address both of these issues. The novelty in our approach is the use of large deviations techniques to relate the \(Q\)-probability of the ball \(B(X^n_1, D)\) around the random center \(X^n_1\), to an associated random walk induced by \(X^n_1\).

Our first result generalizes a strong approximation theorem of Kontoyiannis (1996) which was used to derive the asymptotic results for \(W_n\), and which stated that the waiting time \(W_n\) is asymptotically almost surely close to the reciprocal of the probability \(Q(X^n_1)\) that the string \(X^n_1\) appears in \(Y\). For \(W_n(D)\), Theorem 1 relates the asymptotic behavior of \(\log W_n(D)\) to that of \(-\log Q(B(X^n_1, D))\).

**Theorem 1.** Suppose \(Y\) is a stationary process with uniform-mixing coefficients \(\{\phi(k)\}\) that satisfy \(\sum \phi(k) < \infty\), and assume that for all \(x^n_1 \in A^n_X\) the ball \(B(x^n_1, D)\) is \(Q\)-measurable and \(Q(B(X^n_1, D)) > 0\), \(P\)-a.s. If \(\{c(n)\}\) is an arbitrary sequence of non-negative constants such that \(\sum ne^{-c(n)} < \infty\), then

\[
|\log[W_n(D)Q(B(X^n_1, D))]| \leq c(n) \quad \text{eventually a.s.}
\]

Recall that the uniform-mixing coefficients of \(Y\) are defined by \(\phi(k) = \sup\{|Q(B|A) - Q(B)|\}\) where the supremum is taken over all events \(B \in \sigma(Y^\infty_k)\) and all \(A \in \sigma(Y^0_{-\infty})\) with \(Q(A) \neq 0\); see Bradley (1986) for an extensive discussion. In particular, from Theorem 1 we see that

\[
\log W_n(D) - [-\log Q(B(X^n_1, D))] = o(\sqrt{n}) \quad P - \text{a.s.,}
\]

but, unlike in the case of no distortion, the structure of \(-\log Q(B(X^n_1, D))\) itself is no longer that of a random walk. Nevertheless, we can relate \(-\log Q(B(X^n_1, D))\) 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 \(X^n_1\) (Theorems 2 and 3, below). From that we can read off the exact asymptotic behavior of \(-\log Q(B(X^n_1, D))\), and, via (2), the behavior of the waiting times \(W_n(D)\) (Corollaries 2, 3 and 4).
For simplicity, we assume hereafter that $Y$ is i.i.d. and that $\{d_n\}$ is a sequence of single-letter distortion measures generated by a $d : A_X \times A_Y \to \mathbb{R}$, such that the random variables $d(X_1, Y_1)$ have compact support. In particular we assume,

\[ D_{\min} = \text{ess inf } d(X_1, Y_1) = 0 \]
\[ D_{\max} = \text{ess sup } d(X_1, Y_1) < \infty, \]

and let $D_{av} = Ed(X_1, Y_1)$. Of interest is the range of values of $D$, for which $W_n(D)$ exhibits an exponential behavior.

**Theorem 2.** Let $X$ be a stationary ergodic process and $Y$ be i.i.d. Then for $D \in (0, D_{av})$ we have

\[ -\log Q(B(X^n_1, D)) - n R(\hat{P}_n) = o(\sqrt{n}) \quad P - \text{a.s.}, \]

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

\[ R(\hat{P}_n, Q, D) = \inf \int H(\nu(\cdot | x) | Q(\cdot)) d\hat{P}_n(x), \]

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

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

**Corollary 1.** Assume that $X$ is stationary ergodic, $Y$ is i.i.d., and $D \in (0, D_{av})$. Then

\[ R(\hat{P}_n) \to R(P) \quad P - \text{a.s.,} \]

and hence

\[ \frac{1}{n} \log W_n(D) \to R(P, Q, D) \quad P \times Q - \text{a.s.} \]

Next we investigate the behavior of $\sqrt{n}[R(\hat{P}_n) - R(P)]$. For any probability measure $\mu$ on $A_X$ and any $\lambda \in \mathbb{R}$, let

\[ \Lambda_\mu(\lambda) = \int \log \left\{ \int e^{\lambda d(x, y)} dQ(y) \right\} d\mu(x). \]

Write $\Lambda(\cdot) = \Lambda_P(\cdot)$ when $\mu = P$, $\Lambda_{d_x}(\cdot) = \Lambda_x(\cdot)$ for any $x \in A_X$, and $\Lambda_{X_1}(\cdot) = \Lambda_{X_1}(\cdot) - Ep[\Lambda_{X_1}(\cdot)]$. Theorem 3 provides an explicit approximation of $\sqrt{n}[R(\hat{P}_n) - R(P)]$ by a random walk induced by $X^n_1$.

**Theorem 3.** Let $X$ be stationary ergodic, $Y$ be i.i.d., and $D \in (0, D_{av})$. Then for some $\lambda = \lambda(D) < 0$ such that $\Lambda'(\lambda) = D$ we have

\[ n[R(\hat{P}_n) - R(P)] + \sum_{i=1}^{n} \Lambda_{X_i}(\lambda) = o(\sqrt{n}) \quad P - \text{a.s.} \]
In particular, combining (2) with Theorems 2 and 3 gives
\[
[\log W_n(D) - nR(P, Q, D)] + \sum_{i=1}^{n} \tilde{\Lambda}_X(\lambda) = o(\sqrt{n}) \quad P - \text{a.s.},
\]
and it is now straightforward to harvest a series of corollaries; the following is an immediate consequence of combining (3) with well-known CLT results (see, for example, Theorem 1.7 in Peligrad (1986)).

**Corollary 2 (CLT).** Let \( X \) be a stationary process with strong-mixing coefficients \( \{\alpha(k)\} \) that satisfy \( \sum \alpha(k) < \infty \), and let \( Y \) be i.i.d. For \( D \in (0, D_{av}) \) and \( \lambda = \lambda(D) \) the following series converges,
\[
\sigma^2 = E_P \{\tilde{\Lambda}_X(\lambda)^2\} + 2 \sum_{k=2}^{\infty} E_P \{\tilde{\Lambda}_X(\lambda)\tilde{\Lambda}_X(\lambda)\},
\]
and when \( \sigma^2 > 0 \),
\[
\frac{\log W_n(D) - nR(P)}{\sqrt{n}} \overset{d}{\to} N(0, \sigma^2).
\]
Moreover, when \( \sigma^2 > 0 \), the sequence of processes,
\[
\left\{ \frac{w(nt; D)}{\sigma \sqrt{n}} ; t \in [0, 1] \right\}, \quad n \geq 1,
\]
converges in distribution to standard Brownian motion, where \( w(t; D) = [\log W_{nt}(D) - [nt]R(P, Q, D)] \) for \( t \geq 1 \), and \( w(t; D) = 0 \) for \( t < 1 \).

The strong-mixing coefficients of \( X \) are defined, as usual, by \( \alpha(k) = \sup [|P(A \cap B) - P(A)P(B)| ; A \in \sigma(X_{-\infty}^0), B \in \sigma(X_{k}^\infty)] \); see Bradley (1986) for details. Corollary 3 is a consequence of (3) combined with the LIL (see Theorem 1 in Oodaira and Yoshihara (1971)).

**Corollary 3 (LIL).** Let \( X \) be a stationary process with strong-mixing coefficients \( \{\alpha(k)\} \) that satisfy \( \sum \alpha(k)^{1-\delta} < \infty \) for some \( \delta > 0 \). Suppose \( Y \) is i.i.d., let \( D \in (0, D_{av}) \) and \( \sigma^2 > 0 \) as in (4). Then, with probability one, the set of limit points of the sequence
\[
\left\{ \frac{\log W_n(D) - nR(P)}{\sqrt{2n \log \log n}} \right\}, \quad n \geq 3
\]
coincides with the interval \( [-\sigma, \sigma] \). Moreover, with 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 (3) and an almost sure invariance principle proved by Philipp and Stout (1975), Theorem 4.1.
Corollary 4 (Almost sure invariance principle). Let \( X \) be a stationary process with uniform-mixing coefficients \( \{\phi(k)\} \) that satisfy \( \sum \phi(k) < \infty \). Let \( Y \) be i.i.d., \( D \in (0, D_{av}) \), and \( \sigma^2 > 0 \) as in (4). Then there exists a Brownian motion \( \{B(t) \mid t \geq 0\} \) such that

\[
w(t; D) - \sigma B(t) = o(\sqrt{t}) \quad a.s.
\]

(5)

As usual, we interpret (5) 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 (5) holds. For some of the numerous corollaries that can be derived from almost sure invariance principles as the one in (5) see Strassen (1964) and Ch. 1 of Philipp and Stout (1975).

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

2. Applications.

In this section we discuss two areas of applications in which the question of the asymptotic behavior of \( W_n(D) \) arises naturally.

Data Compression. In the past few years, the analysis of several data compression schemes based on string matching, such as the celebrated Lempel-Ziv algorithm, has been reduced to studying the following idealized scenario (see Wyner and Ziv (1994), Steinberg and Gutman (1993), Yang and Kieffer, Luczak and Sspankowski (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 \( Y \sim Q \), and the encoder’s task is to communicate the “message” \( x^n = (x_1, x_2, \ldots, x_n) \) to the decoder, within some prescribed accuracy \( D \) with respect to some distortion measure \( d_n \). This is done as follows: The encoder scans the database until a \( D \)-close version of \( x^n \) 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 compression ratio of the code in nats-per-symb (by Corollary 1) is given by

\[
\frac{\log W_n(D) + O(\log \log W_n(D))}{n} \rightarrow R(P, Q, D) \quad a.s.
\]

For example, in the case of lossless coding \( (D = 0) \) of an i.i.d. process \( X \), \( R(P, Q, 0) \) reduces to \( H(P) + H(P | Q_1) \), which is interpreted as the optimal limiting compression ratio \( H(P) \), plus the additional “penalty” term \( H(P | Q_1) \) 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, choosing the measure \( Q \) so as to minimize the above rate we obtain the rate-distortion function \( r(D) = \inf_Q R(P, Q, D) \) of the process \( X \) with respect to the distortion measures \( \{d_n\} \) (see Berger (1971) for a general discussion).
Moreover, once the limiting compression ratio is identified, from Corollaries 2, 3 and 4 we get further information about the rate at which it is achieved ("redundancy"), 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 $x_1^n$ and a long but finite "database" sequence $y_1^m$, find the longest contiguous substring in the database that matches an initial portion $x_1^j$ of the template, within distortion $D$ (with respect to some sequence $\{d_n\}$ of distortion measures). The length $L_m(D)$ of the longest such match is of interest here:

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

As was already observed by Wyner and Ziv (1989) for the case $D = 0$, there is a duality relationship between $L_m(D)$ and $W_n(D)$: $W_n(D) \leq m$ if and only if $L_m(D) \geq n$. From this it is easy to see that the asymptotics of $L_m(D)$ can be read off from the corresponding results for $W_n(D)$: The almost sure convergence of $(1/n) \log W_n(D)$ to $R(P)$ is equivalent to

$$\frac{L_m(D)}{\log m} \to \frac{1}{R(P)} \quad \text{a.s.,}$$

and the (one-dimensional) CLT and LIL for $\log W_n(D)$ (Corollaries 2 and 3) translate to:

**Corollary 5.** (i) Under the assumptions of Corollary 2, with $\sigma^2$ as in (4) and $R = R(P, Q, D)$,

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

(ii) Under the assumptions of Corollary 3, with $\sigma^2$ as in (4) and $R = R(P, Q, D)$,

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

3. Strong Approximation.

**Proof of Theorem 1.** Consider the joint process $(X, Y)$ distributed according to the product measure $P = P \times Q$. For an arbitrary constant $K > 1$ and any opening string $x_1^n$ s.t. $Q(B(x_1^n, D)) > 0$,

$$P(W_n(D) < K | X_1^n = x_1^n) \leq \sum_{j=1}^{[K-1]} Q(Y_j^{j+n-1} \in B(x_1^n, D)) \leq K Q(B(x_1^n, D)).$$

In particular, setting $K = e^{-c(n)/Q(B(x_1^n, D))}$ gives

$$P(\log[W_n(D)Q(B(X_1^n, D))] < -c(n) | X_1^n = x_1^n) \leq e^{-c(n)}. \quad (6)$$
(Since $W_n(D) \geq 1$ there is no need to consider $K = e^{-c(n)}/Q(B(x^n_1, D)) \leq 1$.) Summing (6) over $n$, by the Borel-Cantelli Lemma we get
\[
\log[W_n(D)Q(B(X^n_1, D))] \geq -c(n) \quad \text{eventually a.s.} \tag{7}
\]

For the corresponding upper bound we use a standard second-moment blocking argument, similar to the one in Yang and Kieffer (preprint). Fix $n$ large enough so that $e^{c(n)} \geq n + 1$, and $x^n_1 \in A^n_X$ with $Q(B(x^n_1, D)) > 0$. Then for any $K \geq n + 1$ we have
\[
P(W_n(D) > K | X^n_1 = x^n_1) = Q(S_n = 0) \leq \frac{\text{Var}_Q(S_n)}{E_Q S_n^2} \tag{8}
\]
where $S_n = \sum_{j=0}^{V(K,n)} I_n(j)$, $I_n(j)$ is the indicator function of the event \{\(Y_{j+1}^{(j+1)n} \in B(x^n_1, D)\)}, and $V(K,n) = [(K - 1)/n]$. By stationarity,
\[
E_Q S_n = [V(K,n) + 1]Q(B(x^n_1, D)) \tag{9}
\]
and $E_Q(I_n(0)I_n(j)) \leq Q(B(x^n_1, D))\phi((j - 1)n + 1) + Q(B(x^n_1, D))$, so that
\[
\text{Var}_Q(S_n) = \sum_{j,k=0}^{V(K,n)} \text{Cov}_Q(I_n(j)I_n(k)) 
\leq [V(K,n) + 1]Q(B(x^n_1, D)) \left[1 + 2 \sum_{j=1}^{V(K,n)} \phi((j - 1)n + 1)\right]. \tag{10}
\]
Writing $\Phi = 1 + 2 \sum \phi(k)$, and substituting (9) and (10) in (8) we get
\[
P(W_n(D) > K | X^n_1 = x^n_1) \leq \frac{\Phi}{[V(K,n) + 1]Q(B(x^n_1, D))}. \tag{11}
\]
Choosing $K = e^{c(n)}/Q(B(x^n_1, D)) < \infty$ we have $[V(K,n) + 1]Q(B(x^n_1, D)) > e^{c(n)}/2n$, and (11) yields, for $n$ large enough,
\[
P(\log[W_n(D)Q(B(X^n_1, D))] > c(n) | X^n_1) \leq 2\Phi ne^{-c(n)} \quad P - \text{a.s.}
\]
Since the above bound is uniform over $X^n_1$ and summable, by the Borel-Cantelli lemma we obtain the desired upper bound, which, combined with (7) completes the proof. $\Box$

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 Théorem 3 are given.

**Lemma 1.** Let $\mu$ be an arbitrary distribution $A_X$, $\lambda \in \mathbb{R}$, and define $D_{\min}^\mu$, $D_{av}^\mu$ and $D_{\max}^\mu$ like $D_{\min}$, $D_{av}$ and $D_{\max}$, respectively, with $X_1 \sim \mu$. 

8
(i) \(|A_\mu(\lambda)| \leq |\lambda|D^\mu_{\text{max}}.\)

(ii) The Fenchel-Legendre transform of \(A_\mu\),

\[
A^*_\mu(x) = \sup_{\lambda \in \mathbb{R}} [\lambda x - A_\mu(\lambda)]
\]

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

(iii) \(A_\mu \in C^\infty, A'_\mu(0) = D^\mu_{\text{av}}, A''_\mu(\lambda) \geq 0\) for all \(\lambda \leq 0\), and \(A'_\mu(\lambda) \downarrow D^\mu_{\text{min}}\) as \(\lambda \to -\infty.\)

(iv) For each \(D \in (D^\mu_{\text{min}}, D^\mu_{\text{av}})\), there exists a unique \(\lambda < 0\) such that \(A'_\mu(\lambda) = D\), and \(A^*_\mu(D) = \lambda D - A_\mu(\lambda).\)

(v) For (almost) any \(x \in A_X, A_x \in C^\infty\), and its derivatives are uniformly bounded over (almost) all \(x \in A_X\) and all \(\lambda\) in a compact subset of \(\mathbb{R}\).

**Proposition 1.** In the notation of Lemma 1, let \(\mu\) be an arbitrary probability measure on \(A_X\), \(\lambda\) be a negative constant, and \(D \in (D^\mu_{\text{min}}, D^\mu_{\text{av}}).\) Then, \(R(\mu, Q, D) = A^*_\mu(D),\) i.e.

\[
\inf_{\nu \in \mathcal{M}(\mu, D)} \int H(\nu(\cdot|x), Q(\cdot))d\mu(x) = \sup_{\lambda \in \mathbb{R}} \left[ \lambda D - \int \log \left\{ \int e^{\lambda d(x,y)}dQ(y) \right\} d\mu(x) \right], \tag{12}
\]

where the infimum is taken over all probability measures on \(A_X \times A_Y\) such that the \(A_X\)-marginal of \(\nu\) is \(\mu\), and \(\int d(x,y)d\nu(x,y) \leq D.\)

**Proof of Proposition 1.** We know by Lemma 1 that the supremum on the right-hand-side of (12) is achieved by a unique \(\lambda < 0.\) To show that the left-hand-side is no greater than \(A^*_\mu(D)\) we just take \(\nu\) to be the composition of \(\mu\) with the kernel \(\nu(\cdot|x) = Z(\lambda, x)^{-1}Q(\cdot)e^{\lambda d(x,\cdot)}\) where \(Z(\lambda, x) = \int e^{\lambda d(x,y)}dQ(y).\) It is easy to check that \(\int d(x,y)d\nu(x,y) = A'_\mu(\lambda) = D\) and

\[
\int H(\nu(\cdot|x), Q(\cdot))d\mu(x) = \lambda D - \int \log Z(\lambda, x) d\mu(x) = A^*_\mu(D).
\]

To prove the reverse inequality in (12) we recall that for any \(\nu(\cdot|x)\) and any bounded measurable function \(\phi : A_Y \to \mathbb{R},\)

\[
H(\nu(\cdot|x), Q(\cdot)) \geq \int \phi(y)d\nu(y|x) - \log \left\{ \int e^{\phi(y)}dQ(y) \right\}
\]

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

**Proof of Theorem 2.** Let \(D^\nu_{\text{av}}(n) = \int d(x,y)\hat{d}P_n(x)dQ(y),\) so that, by the ergodic theorem,

\[
D^\nu_{\text{av}}(n) \to D_{\text{av}}\quad P - \text{a.s.} \tag{13}
\]
Similarly let \( D_{\min}^{(n)} = \text{ess sup}_{(X_1,Y_1) \sim \hat{P}_n \times Q} d(X_1,Y_1) \), so that
\[
D_{\min}^{(n)} \to D_{\min} \quad P - \text{a.s.}
\] (14)

Given a realization of the \( X \) process such that (13) and (14) hold, for \( n \) large enough the given \( D \) will be strictly between \( D_{\min}^{(n)} \) and \( D_{\min}^{(n)} \), so by Lemma 1 we can choose, for each \( n \), a negative \( \lambda_n \) such that
\[
\Lambda'_{\hat{P}_n}(\lambda_n) = D, \quad \Lambda'_{\hat{P}_n}(D) = \lambda_n D - \Lambda_{\hat{P}_n}(\lambda_n), \quad \text{and} \quad \Lambda''_{\hat{P}_n}(\lambda_n) > 0. \]
We similarly choose \( \lambda < 0 \) such that \( \Lambda'(\lambda) = D \), and claim that
\[
\lambda_n \to \lambda \quad P - \text{a.s.}
\] (15)

To see this suppose, for example, that \( \limsup \lambda_n \leq \lambda - \epsilon \), for some \( \epsilon > 0 \), so that, eventually, \( \lambda_n \leq \lambda - \epsilon/2 \).
Then by the ergodic theorem and the monotonicity of \( \Lambda' \), we get a contradiction: \( D = \Lambda'_{\hat{P}_n}(\lambda_n) = \limsup \Lambda'_{\hat{P}_n}(\lambda_n) \leq \limsup \Lambda'_{\hat{P}_n}(\lambda - \epsilon/2) = \lim n^{-1} \sum_{i=1}^{n} \Lambda'_{z_i}(\lambda - \epsilon/2) = \Lambda'(\lambda - \epsilon/2) < \Lambda'(\lambda) = D \).
The case \( \liminf \lambda_n > \lambda \) is ruled out similarly.

Before we can move into the main part of the proof, we will also need to show that
\[
\Lambda''_{\hat{P}_n}(\lambda_n) \to \Lambda''(\lambda) > 0 \quad P - \text{a.s.}
\] (16)

Writing
\[
|\Lambda''_{\hat{P}_n}(\lambda_n) - \Lambda''(\lambda)| \leq \frac{1}{n} \sum_{i=1}^{n} |\Lambda''_{z_i}(\lambda_n) - \Lambda''(\lambda)| + \left| \frac{1}{n} \sum_{i=1}^{n} \Lambda''_{z_i}(\lambda) - \Lambda''(\lambda) \right|, \] (17)
we can bound the first term above, for \( n \) large enough, by
\[
\text{ess sup}_{X_1} |\Lambda''_{X_1}(\lambda_n) - \Lambda''_{X_1}(\lambda)| \leq \text{ess sup}_{X_1} \sup_{\lambda - \epsilon \leq \xi \leq \lambda + \epsilon} |\Lambda''_{X_1}(\xi)| |\lambda_n - \lambda|
\]
and this converges to zero, by (15) and part (v) of Lemma 1. As for the second term of (17), by the ergodic theorem it converges to zero almost surely.

Now chose and fix a realization \( \{x_i\} \) of \( X \) such that the statements (13) (14) (15) and (16) all hold.
Define \( \zeta_i = d(x_i,Y_i) \), \( T_n = \sum_{i=1}^{n} \zeta_i \), and \( \hat{T}_n = (T_n)/n \), and let \( \mu_n \) denote the law of \( \zeta^n \). In this notation, \( Q(B(x^n_1,D)) = \Pr(\hat{T}_n \leq D) \), 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
\[
\log J_n = o(\sqrt{n}) \quad P - \text{a.s.}
\] (18)
The upper-bound part of (18) follows from

$$J_n = e^{n\Lambda_{\tilde{p}_n}(D)}E\left\{1_{\{T_n \leq D\}}\right\} \leq e^{n\Lambda_{\tilde{p}_n}(D)}E\left\{e^{n\lambda_n(T_n-D)}\right\}$$

$$= e^{n[\Lambda_{\tilde{p}_n}(D)-\lambda_n D]}E\left\{e^{\lambda_n T_n}\right\} = 1$$

(by the choice of $\lambda_n$ and the definition of $\Lambda_{\tilde{p}_n}$).

Turning to the proof of the lower bound, suppose $n$ is large enough so $\lambda_n$ exists, and define a new probability measure $\nu_n$ by

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

Let

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

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

$$\hat{T}_n \overset{D}{=} D - \sqrt{n\Lambda''_{\tilde{p}_n}(\lambda_n)} G_n,$$

so that we can expand

$$J_n = e^{n\Lambda_{\tilde{p}_n}(D)}E_{\nu_n}\left\{1_{\{\hat{T}_n \leq D\}}e^{-n\lambda_n \hat{T}_n + n\Lambda_{\tilde{p}_n}(\lambda_n)}\right\}$$

$$= E_{\nu_n}\left\{1_{\{G_n \geq 0\}}e^{\lambda_n \sqrt{n\Lambda''_{\tilde{p}_n}(\lambda_n)} G_n}\right\}$$

$$\geq E_{\nu_n}\left\{1_{\{0 < G_n < \delta\}}e^{-\alpha_n \sqrt{n} G_n}\right\}$$

$$\geq e^{-\alpha_n \sqrt{n} \delta} \text{Pr}_{\nu_n}(0 < G_n < \delta),$$

for any $\delta > 0$, and where $\alpha_n = -\lambda_n \sqrt{\Lambda''_{\tilde{p}_n}(\lambda_n)} > 0$ and $\alpha_n = O(1)$, by (15) and (16).

Since the random variables $\zeta_i$ are uniformly bounded, and also $\Lambda''_{\tilde{p}_n}(\lambda_n) = O(1)$ by (16), it is easy to check that the Lindeberg condition for the CLT is satisfied by $G_n$, from which it follows that the probability $\text{Pr}_{\nu_n}(0 < G_n < \delta) \to \rho > 0$ as $n \to \infty$. Now choose $M > 0$ large enough so that $M - \alpha_n$ is bounded away from zero, and get from (19) that

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

i.e.,

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

from which we conclude that

$$\liminf_n \frac{1}{\sqrt{n}} \log J_n \geq -M \delta.$$
Since $\delta > 0$ was arbitrary and $M > 0$ was chosen independent of $\delta$, letting $\delta \downarrow 0$ completes the proof. □

**Proof of Corollary 1.** Corollary 1 follows from combining Theorem 2 with (2), provided we show that $R(\hat{P}_n) \to R(P)$ almost surely, or, equivalently (by Proposition 1), that $\Lambda_{\hat{P}_n}^*(D) \to \Lambda^*(D)$ almost surely. But recall that $\Lambda_{\hat{P}_n}^*(D) = \lambda_n D - \Lambda_{\hat{P}_n}(\lambda_n)$ and $\Lambda^*(D) = \lambda D - \Lambda(\lambda)$, as in the proof of Theorem 2, where $\lambda_n \to \lambda$ almost surely by (15). So we only have to show that $\Lambda_{\hat{P}_n}(\lambda_n) \to \Lambda(\lambda)$ which comes from an obvious modification of the derivation of (16) in the previous proof. □

**Proof of Theorem 3.** Let $\lambda$ 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 $\Lambda''$ we can choose constants $\delta, \eta > 0$ such that $\Lambda''(\lambda + \theta) > \eta$ whenever $|\theta| < \delta$. Also, from (15), we can pick $N = N(x_1^\infty)$ such that $|\lambda_n - \lambda| < \delta$ for all $n \geq N$.

In view of Proposition 1 of the previous section, what we want to prove can be rewritten as

$$\sqrt{n}\left\{[\Lambda_{\hat{P}_n}^*(D) - \Lambda^*(D)] - [\Lambda(\lambda) - \Lambda_{\hat{P}_n}(\lambda)]\right\} \to 0 \text{ a.s.} \quad (20)$$

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 (20) is the same as

$$\sqrt{n} \sup_{|\theta| < \delta} \left[\theta D - \Lambda_{\hat{P}_n}(\theta + \lambda) + \Lambda_{\hat{P}_n}(\lambda)\right] \to 0 \text{ a.s.} \quad (21)$$

Since this supremum is always non-negative (take $\theta = 0$), (21) is equivalent to

$$\liminf_n \sqrt{n} \inf_{|\theta| < \delta} \frac{1}{n} \sum_{i=1}^{n} [f(\theta, x_i) - f(0, x_i)] \geq 0 \text{ a.s.}, \quad (22)$$

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

$$\frac{1}{n} \sum_{i=1}^{n} [f(\theta, x_i) - f(0, x_i)] = \theta \frac{1}{n} \sum_{i=1}^{n} f'(0, x_i) + \frac{\theta^2}{2} \frac{1}{n} \sum_{i=1}^{n} f''(\xi_n, x_i), \quad (23)$$

for some $|\xi_n| < \delta$. It is now obvious that the infimum of the quadratic in the above right-hand-side over $|\theta| < \delta$ is bounded below by $-A_n^2/B_n$, where

$$A_n = \frac{1}{n} \sum_{i=1}^{n} f'(0, x_i) \to E Pf'(0, X_1) = 0, \quad (24)$$

since $E Pf'(0, X_1) = \Lambda'(\lambda) - D = 0$ by our choice of $\lambda$, and

$$B_n = \frac{1}{n} \sum_{i=1}^{n} f''(\xi_n, x_i).$$
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),

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

so that

\[
\liminf_{n} B_n \geq \liminf_{n} \left\{ - \sup_{|\xi| < \delta} \left| \frac{1}{n} \sum_{i=1}^{n} f''(\xi, x_i) - E_P f''(\xi, X_1) \right| + E_P f''(\xi, X_1) \right\} \\
\geq \inf_{|\xi| < \delta} E_P f''(\xi, X_1) \quad \text{a.s.} \\
= \inf_{|\xi| < \delta} \Lambda''(\lambda + \xi) \geq \eta > 0,
\]

(25)

by our initial choice of \( \delta \). Combining (23) with (24) and (25) gives (22), as required.

\[ \square \]

Acknowledgment. We would like to thank W. Szpankowski for sending us an early preprint of Luczak and Szpankowski (1997).

References


DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CA 94305-4065
amir@stat.stanford.edu

INFORMATION SYSTEMS LABORATORY
STANFORD UNIVERSITY
STANFORD CA 94305-4055
yiannis@isli.stanford.edu