RARITY AND EXPONENTIALITY: AN EXTENSION OF KEILSON’S THEOREM, WITH APPLICATIONS

RUDOLF GRÜBEL * ** AND MARCUS REICH,* *** Universität Hannover

Abstract

We generalize a theorem due to Keilson on the approximate exponentiality of waiting times for rare events in regenerative processes. We use the result to investigate the limit distribution for a family of first entrance times in a sequence of Ehrenfest urn models. As a second application, we consider approximate pattern matching, a problem arising in molecular biology and other areas.

Keywords: Exponential distribution; first entrance time; Levenshtein distance; limit distribution; Ornstein–Uhlenbeck process; pattern matching; random string; regenerative process

2000 Mathematics Subject Classification: Primary 60F05
Secondary 60J10

1. Introduction

Suppose that we are interested in the first entrance time
\[ \rho_B = \inf \{ t \in T : X_t \in B \} \]

of a stochastic process \( X = (X_t)_{t \in T}, T \subset \mathbb{R}_+ \), into some subset \( B \) of its state space. If \( X \) can be decomposed into regenerative cycles then \( \rho_B \) is the random sum of the length of the cycles that missed \( B \), plus the part of the last cycle up to the first hit. If the probability of hitting \( B \) in a single cycle is small and if cycle lengths are not too heavy tailed, then we might expect that the distribution of \( \rho_B/E\rho_B \) is close to \( \text{Exp}(1) \), the exponential distribution with mean 1.

We generalize this result to a sequence of stochastic processes. Section 2 gives the basic limit theorem and, in Section 3, we apply this to the Ehrenfest urn model. In Section 4, we consider a question arising in molecular biology and other areas: we obtain asymptotic exponentiality for the time until a given pattern first occurs approximately, with respect to the Levenshtein distance, in a random string. This second application also shows that our generalization of Keilson’s theorem can be applied to more general stopping times than the above entry times, which depend on the value of a single variable of the process.

Received 12 May 2004; revision received 10 October 2004.
* Postal address: Institut für Mathematische Stochastik, Universität Hannover, Postfach 60 09, D-30060 Hannover, Germany.
** Email address: rgrubel@stochastik.uni-hannover.de
*** Email address: reich@stochastik.uni-hannover.de
2. A limit theorem

Our extension of Keilson’s result is somewhat similar to the transition from the central limit theorem for a fixed sequence of random variables to the Lindeberg theorem; see, e.g., Theorem 27.1 and Theorem 27.2 of Billingsley (1986). In its abstract formulation, the result does not even mention entrance times and we consider the following formal framework. For each \( n \in \mathbb{N} \), we have probability measures \( \mu_n, \mu_{1,n}, \) and \( \mu_{2,n}, \) concentrated on the positive half-line, and real numbers \( p_n \) and \( q_n \) with \( 0 < p_n < 1, p_n + q_n = 1 \), such that

\[
\mu_n = p_n \mu_{1,n} + q_n \mu_{2,n} \quad \text{for all } n \in \mathbb{N}, \quad \text{with } \lim_{n \to \infty} q_n = 0. \tag{1}
\]

We further assume that each \( \mu_n \) has finite mean \( m_n \), and that

\[
\lim_{\eta \to \infty} \sup_{n \in \mathbb{N}} \int_{x > \eta m_n} x m_n \mu_n(dx) = 0. \tag{2}
\]

We can now state our first result; its proof combines Keilson’s ideas with techniques from the proof of Lindeberg’s theorem as given in Billingsley (1986).

**Theorem 1.** Suppose that, for each \( n \in \mathbb{N} \), \( Y_{n,k}, Z_{n,k}, \) and \( I_{n,k}, k \in \mathbb{N} \), are independent random variables with \( Y_{n,k} \sim \mu_{1,n}, Z_{n,k} \sim \mu_{2,n} \) and \( P(I_{n,k} = 0) = p_n, P(I_{n,k} = 1) = q_n \) for all \( k \in \mathbb{N} \). Let

\[
S'_n := \sum_{k=1}^{N_n-1} Y_{n,k} \quad \text{and} \quad S''_n := S'_n + Z_{n,N_n}, \quad \text{with } N_n := \inf\{k \in \mathbb{N} : I_{n,k} = 1\}.
\]

Then, for every sequence \( (S_n)_{n \in \mathbb{N}} \) of random variables satisfying

\[
S'_n \leq S_n \leq S''_n \quad \text{for all } n \in \mathbb{N},
\]

we have

\[
\lim_{n \to \infty} \frac{q_n E S'_n}{m_n} = 1 \quad \text{and} \quad \lim_{n \to \infty} P\left( \frac{S_n}{E S_n} \leq x \right) = 1 - e^{-x} \quad \text{for all } x \geq 0.
\]

**Proof.** The representation (1) implies that \( q_n \int \mu_{2,n} \leq \int \mu_{2,n} \) for all nonnegative measurable functions \( f \). Hence, for every \( \eta > 0 \),

\[
\frac{q_n E Z_{n,1}}{m_n} - \frac{q_n}{m_n} \int_{x \leq \eta m_n} x \mu_{2,n}(dx) - \frac{q_n}{m_n} \int_{x > \eta m_n} x \mu_{2,n}(dx) \leq q_n \eta + \int_{x > \eta m_n} \frac{x}{m_n} \mu_n(dx).
\]

By choosing \( \eta \) sufficiently large and using (2), and then choosing \( n \) sufficiently large and using the second part of (1), we obtain

\[
\lim_{n \to \infty} \frac{q_n E Z_{n,1}}{m_n} = 0. \tag{3}
\]

From (1) it follows that \( m_n = p_n E Y_{n,1} + q_n E Z_{n,1} \), so that (3) implies

\[
\lim_{n \to \infty} \frac{p_n E Y_{n,1}}{m_n} = 1. \tag{4}
\]
Obviously, \( P(N_n = i) = p_n^{i-1}q_n \) for all \( i \in \mathbb{N} \). Also, \( N_n \) and \((Y_{n,k})_{k \in \mathbb{N}}\) are independent, meaning that the characteristic function \( \phi_n(\theta) = \mathbb{E}\exp(i\theta S'_n) \) of the sum \( S'_n \) can be written as

\[
\phi_n(\theta) = \frac{q_n}{1 - p_n \phi_n'(\theta)} \quad \text{with} \quad \phi_n'(\theta) := \mathbb{E}e^{i\theta Y_n}. 
\]

Using the fact that \(|e^{iy} - 1 - iy| \leq \min\{2|y|, y^2\} \) for all \( y \in \mathbb{R} \) (see, e.g. Billingsley (1986, Equation (26.4))), we obtain

\[
\begin{align*}
\left|\frac{p_n}{q_n} \phi_n'(\frac{q_n \theta}{m_n}) - \left(1 + i\theta \frac{q_n \mathbb{E}Y_{n,1}}{m_n}\right)\right| &\leq \frac{p_n}{q_n} \mathbb{E}\min\left\{\frac{2|\theta q_n Y_{n,1}|}{m_n}, \left(\frac{\theta q_n Y_{n,1}}{m_n}\right)^2\right\} \\
&\leq \frac{p_n q_n \theta^2}{m_n} \int_{x \leq \eta m_n} x^2 \mu_{1,n}(dx) + 2|\theta| \int_{x > \eta m_n} x \mu_{1,n}(dx) \\
&\leq \frac{p_n q_n \eta \theta^2}{m_n} + 2|\theta| \int_{x > \eta m_n} x \mu_n(dx).
\end{align*}
\]

For the second term in the upper bound we can use (2), so that, by taking \( \eta \) and then \( n \) sufficiently large, and using the second part of (1) again, we obtain

\[
\begin{align*}
\lim_{n \to \infty} \frac{p_n}{q_n} \phi_n(\frac{q_n \theta}{m_n}) - \left(1 + i\theta \frac{q_n \mathbb{E}Y_{n,1}}{m_n}\right) &= 0.
\end{align*}
\]

Together with (4), this gives

\[
\begin{align*}
\lim_{n \to \infty} \phi_n(\frac{q_n \theta}{m_n})^{-1} &= \lim_{n \to \infty} \left(1 - i\theta \frac{p_n \mathbb{E}Y_{n,1}}{m_n} - \frac{p_n}{q_n} \phi_n'(\frac{q_n \theta}{m_n}) \right) \\
&= 1 - i\theta
\end{align*}
\]

for all \( \theta \in \mathbb{R} \). Together with the continuity theorem for characteristic functions, this implies the convergence in distribution of \( q_n S'_n/m_n \) to \( \text{Exp}(1) \). Since \( N_n \) is independent of \((Z_{n,k})_{k \in \mathbb{N}}\), it follows from (3) that \( \lim_{n \to \infty} \mathbb{E}(q_n Z_{n,N_n}/m_n) = 0 \), which gives the convergence in probability of \( q_n Z_{n,N_n}/m_n \) to 0. Therefore, by Slutsky’s theorem, \( q_n S'_n/m_n \) also has limit distribution \( \text{Exp}(1) \).

Finally, Wald’s equation yields

\[
\mathbb{E}S'_n = \left(\frac{1}{q_n} - 1\right) \mathbb{E}Y_{n,1}, \quad \mathbb{E}S''_n = \left(\frac{1}{q_n} - 1\right) \mathbb{E}Y_{n,1} + \mathbb{E}Z_{n,1}.
\]

Together with (3) and (4), this gives \( \lim_{n \to \infty} q_n \mathbb{E}S'_n/m_n = 1 \), which completes the proof of the theorem.

Condition (2) can be rephrased as the uniform integrability of the standardized variables \( V_n/\mathbb{E}V_n \), where \( V_n := I_{n,1} Y_{n,1} + (1 - I_{n,1}) Z_{n,1} \) has distribution \( \mu_n \). This is implied, for example, by

\[
\sup_{n \in \mathbb{N}} \int_{m_n}^{x}(\frac{x}{m_n})^{1+\delta} \mu_n(dx) < \infty \quad \text{for some} \quad \delta > 0.
\]
For another sufficient condition that works well with our later applications, let $S_a(\mu)$, $a > 0$, be the measure defined by

$$S_a(\mu)([0, x]) = \mu([0, ax]) \quad \text{for all } x \geq 0.$$

Clearly, if $X$ has distribution $\mu$ then $S_a(\mu)$ is the distribution of $X/a$. We write '$w \rightarrow$' for the weak convergence of probability measures.

**Condition 1.** There exist a sequence $(a_n)_{n \in \mathbb{N}}$ of positive real numbers and a probability measure $\mu_\infty$ on $[0, \infty)$, with positive, finite first moment $m_\infty$, such that

$$m_n/a_n \to m_\infty \quad \text{and} \quad S_{a_n}(\mu_n) \overset{w}{\to} \mu_\infty \quad \text{as } n \to \infty.$$

Suppose that $V_n$ has distribution $\mu_n$. If Condition 1 is satisfied then the family $V_n/a_n$ is uniformly integrable, by Theorem 5.4 of Billingsley (1968). Multiplying these by $a_n/m_n$ does not destroy the uniform integrability, since the sequence converges to a finite limit. Hence, Condition 1 implies (2).

### 3. An application to the Ehrenfest urn model

We now return to the problem outlined in Section 1, assuming that $T = N_0 := \mathbb{N} \cup \{0\}$ for convenience. For each $n \in \mathbb{N}$, let $(X_{n,m})_{m \in N_0}$ be a stochastic process with an associated sequence $(\tau_{n,k})_{k \in \mathbb{N}_0}$ of regeneration times. By this we mean that, for each $n \in \mathbb{N}$, $(\tau_{n,k})_{k \in \mathbb{N}_0}$ is a sequence of stopping times with respect to the natural filtration of $(X_{n,m})_{m \in \mathbb{N}_0}$, and that the segments $(X_{n,m})_{m=\tau_{n,k},\tau_{n,k}+1,\ldots,\tau_{n,k+1}-1}$, $k \in \mathbb{N}_0$, are independent and identically distributed for each $n \in \mathbb{N}$. In particular, the lengths $(L_{n,k})_{k \in \mathbb{N}}$, $L_{n,k} := \tau_{n,k} - \tau_{n,k-1}$, of the segments are independent and identically distributed. We additionally assume that $\tau_{n,0} = 0$ for all $n \in \mathbb{N}$.

Let $B_n$ be a measurable subset of the state space of the process $(X_{n,m})_{m \in \mathbb{N}_0}$ and let $\rho_n := \inf \{m \in \mathbb{N}_0 : X_{n,m} \in B_n \}$ be the corresponding entrance time. We assume that $B_n$ will be visited in a segment of the process with positive probability since, otherwise, $\rho_n = \infty$ with probability 1. The following dictionary provides the connection to the framework introduced in Section 2:

- $\mu_n \leftrightarrow$ the distribution of $L_{n,1}$,
- $q_n \leftrightarrow$ the probability that $X_{n,m} \in B_n$ for some $m \in [0, \ldots, \tau_{n,1} - 1]$,
- $\mu_{1,n} \leftrightarrow$ the conditional distribution of $L_{n,1}$, given that $X_{n,m} \notin B_n$ for all $m \in [0, \ldots, \tau_{n,1} - 1]$,
- $\mu_{2,n} \leftrightarrow$ the conditional distribution of $L_{n,1}$, given that $X_{n,m} \in B_n$ for some $m \in [0, \ldots, \tau_{n,1} - 1]$.

The process starts with a random number of cycles that avoid $B_n$ and then has a cycle in which $B_n$ is visited; clearly $S_n' \leq \rho_n \leq S_n''$, in the notation of Theorem 1. The theorem therefore shows that entrance times are asymptotically exponential if the following two conditions are met: first, the probability of a hit in a single cycle converges to 0; and second, the cycle lengths for the sequence of processes, normed to have expectation 1, are uniformly integrable.

As mentioned above, for the case in which we have one process only (and a fixed cycle-length distribution) this asymptotic exponentiality has been obtained by Keilson (1966). We now give a first example that shows that our generalization can be useful.
We consider the following variant of the Ehrenfest urn model. There are two urns, each containing \( n \) balls at time \( m = 0 \). At time \( m \in \mathbb{N} \), one of the \( 2n \) balls is selected uniformly and at random, and moved to the other urn. Let \( X_{n,m} \) be the absolute value of the difference, times one-half, in ball-count of the two urns after \( m \) such steps. Clearly, \( (X_{n,m})_{m \in \mathbb{N}_0} \) is a Markov chain with state space \( S_n := \{0, 1, \ldots, n\} \), origin 0 and transition probabilities

\[
p^{(n)}_{ij} = \begin{cases} 
1 & \text{for } i = 0, j = 1, \\
\frac{(n - i)}{(2n)} & \text{for } j = i + 1, i = 1, \ldots, n - 1, \\
\frac{(n + i)}{(2n)} & \text{for } j = i - 1, i = 1, \ldots, n, \\
0 & \text{otherwise}.
\end{cases}
\]

Kac (1959) and Kemperman (1961) are canonical references for such models and their impact on understanding the transition from the time-reversible laws of classical mechanics to the irreversible laws of thermodynamics. In this context, the very short time needed to go from a state with one empty urn to a state in which the two urns contain roughly the same number of balls – short in comparison to the time taken for the opposite transition – plays a fundamental role. Here we are interested in the onset of asymptotic exponentiality for the time needed to go from a balanced state, i.e. \( X_{n,0} = 0 \), to a state in which the difference in contents first reaches a certain value. As a consequence, we discuss the above variant; the symmetry of the original model obviously implies that the modification does not destroy the Markov property.

We consider the first entrance into

\[ B_n := \{b_n, b_n + 1, \ldots, n\}, \]

where the sequence \( (b_n)_{n \in \mathbb{N}} \) increases to infinity. The following result shows that asymptotic exponentiality holds whenever \( b_n \) grows faster than \( n^{1/2} \).

**Theorem 2.** With \( (X_{n,m})_{m \in \mathbb{N}_0}, (b_n)_{n \in \mathbb{N}}, \) and \((\rho_n)_{n \in \mathbb{N}}\) as above, we have

\[
\lim_{n \to \infty} P\left(\frac{\rho_n}{E\rho_n} \leq x \right) = 1 - e^{-x} \quad \text{for all } x \geq 0,
\]

provided that \( \lim_{n \to \infty} n/b_n^2 = 0 \).

For the proof we need some auxiliary results. Let \( Y = (Y_m)_{m \in \mathbb{N}_0} \) be an irreducible and aperiodic Markov chain with finite state space \( E \) and transition probabilities \( (p_{ij})_{i,j \in E} \). For \( i \in E \), let

\[
T_i := \min\{m \in \mathbb{N}_0 : Y_m = i\} \quad \text{and} \quad T_i^+ := \min\{m \in \mathbb{N} : Y_m = i\}
\]

be the first entrance and the first return time to \( i \), respectively; \( \pi = (\pi_i)_{i \in E} \) denotes the stationary distribution of the process. Our first lemma follows from Section I.11, Corollary 1 of Chung (1967); see also Chapter 2, Corollary 8 of Aldous and Fill (2004). As usual, we write \( P_i(\cdot) \) and \( E_i(\cdot) \) for probability and expectation with respect to having started at \( i \in E \).

**Lemma 1.** For all \( i, j \in E \) with \( i \neq j \),

\[
E_i T_j + E_j T_i = \frac{1}{\pi_i p_i(T_j < T_i^+)}. 
\]
We now further specify $Y$ to have state space $E = \{0, 1, \ldots, n\}$ and transition probabilities $p_{ij} = 0$ for $|i - j| \neq 1$ and $p_{ij} > 0$ for $|i - j| = 1$; note that the urn model described above satisfies these conditions. Such a process is an irreducible birth–death process and can be regarded as a random walk on a weighted graph with vertices $E$ and edge set $\{(i, i + 1) : i = 0, 1, \ldots, n - 1\}$. The following lemma is Chapter 5, Proposition 3(a) of Aldous and Fill (2004).

Lemma 2. For all $i$ and $j$ with $0 < i < j \leq n$,

$$P_i(T_j < T_0) = \frac{\sum_{l=0}^{j-1} (\pi_l p_{l,l+1})^{-1} - 1}{\sum_{l=0}^{j-1} (\pi_l p_{l,l+1})^{-1}}.$$

For the last of our auxiliary results, we recall that $U = (U_t)_{t \geq 0}$ is an Ornstein–Uhlenbeck process (with standard parameters) if $U$ is a (strong) Markov process with state space $\mathbb{R}$, continuous paths, and infinitesimal generator

$$Af(x) = \frac{1}{2} f''(x) - x f'(x)$$

for functions $f : \mathbb{R} \to \mathbb{R}$ that are twice continuously differentiable. As in the Markov chain case, we write $P_x(\cdot)$ and $E_x(\cdot)$ for probability and expectation with respect to having started at $x \in \mathbb{R}$.

Lemma 3. Let $\tau_c := \inf\{t > 0 : U_t = c\}$ be the first hitting time of $c \in \mathbb{R}$, and let

$$\psi(x) := \sqrt{\pi} \int_0^x e^{-y^2} \, dy + 2 \int_0^x e^{-y^2} \int_0^y e^{-z^2} \, dz \, dy.$$

Then, for all $a, b \in \mathbb{R}$ with $a < b$,

$$E_a \tau_b = \psi(b) - \psi(a). \quad (5)$$

This result is taken from Nobile et al. (1985), who do not give a proof but refer to the physics literature. As there are different conventions with regard to standard parameters and also with respect to the definition of the error function, we sketch an argument that leads to (5). Suppose that $f$ is twice continuously differentiable and satisfies $Af \equiv 1$. Itô’s formula then implies that $(Y_t)_{t \geq 0}$, with $Y_t := f(U_t) - t$, is a local martingale with respect to the natural filtration associated with $(U_t)_{t \geq 0}$. The differential equation

$$\frac{1}{2} h''(x) - x h'(x) = 1,$$

for all $x \in \mathbb{R}$, has the general solution

$$h(x) = \eta e^{x^2} + 2 e^{x^2} \int_0^x e^{-y^2} \, dy, \quad \eta \in \mathbb{R}.$$ 

With $\eta = \pi^{1/2}$, we can find a suitable upper bound in $(-\infty, b]$ for $|f|$, with $f(x) := \int_0^x h(y) \, dy$, $x \in \mathbb{R}$, so that $Y$, stopped at $\tau_b$, is a martingale and the optional stopping theorem applies. This yields

$$f(a) = E_a Y_0 = E_a Y_{\tau_b} = f(b) - E_a \tau_b,$$

and (5) follows on noting that $f = \psi$. 

Downloaded from https://www.cambridge.org/core. IP address: 54.191.40.80, on 20 Aug 2017 at 02:23:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S002190020000425
Proof of Theorem 2. We want to use Theorem 1 with Condition 1 as a sufficient condition for (2). As in our other application (to be given in the next section) the construction of a suitable family of regeneration times is a crucial step. Here, we let $L_{n,1}$ be the time of the first visit of the process $(X_{n,m})_{m\in\mathbb{N}_0}$ to 0 after its first visit to $a_n := \lceil n^{1/2} \rceil$, where $\lceil \cdot \rceil$ denotes the smallest integer greater than or equal to its argument. We have to show that the probability of an interesting event, i.e. a visit to $B_n$, occurring within a regenerative cycle tends to 0, and that the distribution of the suitably rescaled length of the regenerative cycles converges in mean and in distribution. In order to be able to use Lemmas 1 and 2 for this purpose, we first investigate the asymptotics of the stationary probabilities and some associated quantities.

The stationary distribution associated with the $n$th model $(X_{n,m})_{m\in\mathbb{N}_0}$ is given by $\pi^{(n)} = (\pi_0^{(n)}, \ldots, \pi_n^{(n)})$, where

$$
\pi_0^{(n)} = \frac{1}{2^{2n}} \left( \frac{2n}{n} \right), \quad \pi_i^{(n)} = \frac{1}{2^{2n-1}} \left( \frac{2n}{n + i} \right) \quad \text{for } i = 1, \ldots, n;
$$

see, e.g. Brémaud (1999, pp. 76f.) (the minor modification needed for our variant should be obvious). Hence, for fixed $C > 0$ and with $c_n := \lceil C n^{1/2} \rceil$,

$$
\pi_0^{(n)} P_{0,1}^{(n)} = \left( \frac{2n}{n} \right) \frac{1}{2^{2n}} \sim \frac{1}{\sqrt{n}},
$$

$$
\pi_l^{(n)} P_{l,l+1}^{(n)} = \frac{1}{2^{2n}} \left( \frac{2n}{n + l} \right) \left( 1 - \frac{l}{n} \right) \sim \frac{1}{2^{2n}} \left( \frac{2n}{n + l} \right),
$$

the latter uniformly in $l = 1, \ldots, c_n$. The local form of the normal approximation for the binomial distribution leads to

$$
\lim_{n \to \infty} \sup_{1 \leq l \leq c_n} \left| \frac{2n}{n + l} - \left( 1 - \frac{l}{n} \right) \right| = 1 = 0.
$$

Together with the elementary fact that

$$
\lim_{n \to \infty} \sup_{1 \leq l \leq c_n} \left| \frac{a_n}{b_n} - 1 \right| = 0 \Rightarrow \lim_{n \to \infty} \sup_{1 \leq l \leq c_n} \left| \frac{b_n}{a_n} - 1 \right| = 0
$$

for arbitrary $a_n, b_n \in \mathbb{R}$, this implies that

$$
\lim_{n \to \infty} \sup_{1 \leq l \leq c_n} \left| \frac{\pi_l^{(n)} P_{l,l+1}^{(n)}}{\sqrt{\pi} n \exp(l^2/n)} - 1 \right| = 0.
$$

Hence, we have shown that

$$
\lim_{n \to \infty} \frac{1}{n} \sum_{l=0}^{c_n} \left( \pi_l^{(n)} P_{l,l+1}^{(n)} \right)^{-1} = \lim_{n \to \infty} \sum_{l=0}^{c_n} \frac{1}{\sqrt{n}} \exp \left( \left( \frac{l}{\sqrt{n}} \right)^2 \right)
$$

$$
= \sqrt{\pi} \int_0^C e^{x^2} \, dx \quad \text{for all } C > 0.
$$

(6)

Now let $q_n$ be the probability that a visit to $b_n$ occurs between two regenerative times. Given that we start at 0, this necessarily happens after the first visit to $a_n$. Hence, we have

$$
q_n = P_{a_n} (T_{b_n}^{(n)} < T_0^{(n)}),
$$
provided that \( n \) is large enough, and the assumption on the rate of growth of \( (b_n)_{n \in \mathbb{N}} \), (6), and Lemma 2 together imply that

\[
\limsup_{n \to \infty} q_n \leq \frac{\int_0^1 e^{x^2} \, dx}{\int_C^1 e^{x^2} \, dx} \quad \text{for all } C > 1.
\]

By letting \( C \to \infty \), we see that (1) is satisfied.

For the second part of the argument, i.e. the verification of Condition 1 for the distribution of the length of the regenerative cycles, we first note that the behaviour of the first moment can also be obtained from the above calculations: we have

\[
E_0 L_{n,1} = E_0 T_{a_n}^{(n)} + E_{a_n} T_0^{(n)}, \quad P_0 (T_{a_n}^{(n)} = T_0^{(n)}),
\]

so that

\[
E_0 L_{n,1} = \sum_{i=0}^{a_n-1} \left( \mu_i^{(n)} \right) p_{i,i+1}^{(n)} - n \sqrt{\pi} \int_0^1 e^{x^2} \, dx \quad (7)
\]

by Lemma 1, Lemma 2, and (6).

For the convergence in distribution of the rescaled regeneration times, we invoke a functional limit theorem. Our process can be regarded as a simple function of another chain that converges to an Ornstein–Uhlenbeck process. To make this precise, let \( (\tilde{X}_{n,m})_{m \in \mathbb{N}_0} \) be a Markov chain with state space \( \{-n, -n+1, \ldots, n-1, n\} \) and transition probabilities

\[
p_{i,i+1} = \frac{n-i}{2n} \quad \text{for } i < n, \quad p_{i,i-1} = \frac{n+i}{2n} \quad \text{for } i > -n.
\]

Then, \( (X_{n,m})_{m \in \mathbb{N}_0} \) is equal in distribution to \( (\tilde{X}_{n,m})_{m \in \mathbb{N}_0} \). In addition, let \( (N_t)_{t \geq 0} \) be a Poisson process with unit intensity, independent of \( (\tilde{X}_{n,m})_{m \in \mathbb{N}_0} \), and let \( (\tilde{Z}_{n,t})_{t \geq 0} \) be defined by

\[
\tilde{Z}_{n,t} = \tilde{X}_{n,N_t}, \quad t \geq 0.
\]

This results in a continuous-time Markov chain to which the results of Rosenkrantz and Dorea (1980) can be applied: the process \( (Z_{n,t})_{t \geq 0} \) with

\[
Z_{n,t} := \frac{1}{\sqrt{n}} \tilde{Z}_{n,t}, \quad t \geq 0,
\]

converges in distribution, as \( n \to \infty \), in the space of càdlàg functions on \([0, \infty)\) to a standard-parameter Ornstein–Uhlenbeck process \( U = (U_t)_{t \geq 0} \) with origin at 0.

Let \( \tau \) and \( \tau_n, n \in \mathbb{N} \), be the times of the first visit to 0 after the first exit from \((-1, 1)\) of \( (U_t)_{t \geq 0} \) and \( (Z_{n,t})_{t \geq 0} \), respectively. The process convergence implies the convergence in distribution of \( \tau_n \) to \( \tau \) as \( n \to \infty \). From the construction, it follows that \( L_{n,1} = N_{\tau_n}, \) so that

\[
\frac{1}{n} L_{n,1} = \tau_n \frac{N_{\tau_n}}{n \tau_n}.
\]

Furthermore, \( N_i / t \to 1 \) almost surely as \( t \to \infty \) and, hence, \( L_{n,1} / n \) converges in distribution to \( \tau \). In order to be able to use Condition 1, and in view of (7), it therefore remains to show that

\[
E_0 \tau = \pi^{1/2} \int_0^1 e^{x^2} \, dx.
\]
We regard the regeneration sequence as fixed and assume that (2) is satisfied. If $R$ and exponentiality

401

Ornstein–Uhlenbeck process in the case that $c (A_n)n$

simple symmetric random walk. Hence, if we take the lengths

light on condition (2): the 'raw', i.e. unrescaled, Ehrenfest models converge in distribution to a

infinite first moment). With respect to these excursions, the sets

takes a simple symmetric random walk to return to 0 (which is finite with probability 1, but has

0 as our basic regeneration intervals, we find that

for Markov chains implies that $E_0 \min \{\tau_0, \, \rho\} = \rho$, so that $E_0 = E_0 + \frac{1}{2} (E_1 \tau_1 + E_{-1} \tau_1)$,

so that $E_0 \rho = \frac{1}{2} (E_0 \tau_1 - E_{-1} \tau_0)$ and, therefore,

Using the symmetry properties of $\psi$, we see that Lemma 3 provides the required formula

for $E_0 \tau$. This completes the proof of Theorem 2.

Nobile et al. (1985) obtained asymptotic exponentiality for the hitting times of $c$ of an

Ornstein–Uhlenbeck process in the case that $c \to \infty$. Note that this result, together with

the weak convergence of the Ehrenfest models to an Ornstein–Uhlenbeck process, which we

also used in the proof, does not imply the statement in Theorem 2 unless we are prepared to

interchange limits uncritically. Urn models of the above type also appear as models of biological

populations. Our result may be considered as dealing with these in a somewhat more direct

manner, i.e. not via limiting properties of the limit process.

Certain aspects of asymptotic exponentiality in a sequence of Markov chains become trans-

parent in our approach. For example, let $q_n(B_n)$ be the probability that $B_n$ is visited during

an $L_n$-segment, where we use the general framework summarized in the above dictionary. We regard the regeneration sequence as fixed and assume that (2) is satisfied. If $(B_n)n\in\mathbb{N}$ is

rare (with respect to $(L_n)n\in\mathbb{N}$) in the sense that $q_n(B_n) > 0$ with $\lim_{n\to\infty} q_n(B_n) = 0$, then the

entrance times into $B_n$ are asymptotically exponential. Now, if there are two such sequences

$(A_n)n\in\mathbb{N}$ and $(B_n)n\in\mathbb{N}$ then, since

\[ 0 < q_n(A_n) \land q_n(B_n) \leq q_n(A_n \cup B_n) \leq q_n(A_n) + q_n(B_n) \to 0 \quad \text{as } n \to \infty, \]

we also have asymptotic exponentiality for the entrance times associated with the unions

$(A_n \cup B_n)n\in\mathbb{N}$. Similarly, we may consider subsets, potentially changing the rate $m_n$

in the process.

Together with the formula for the Laplace transform of the first exit time given in Borodin and

Salminen (1996), the weak convergence to an Ornstein–Uhlenbeck process shows that we do not have asymptotic exponentiality if $b_n \sim C n^{1/2}$ with $0 < C < \infty$ fixed. The condition on the rate of growth of $(b_n)n\in\mathbb{N}$ in Theorem 2 is therefore sharp. This example also throws some

light on condition (2): the 'raw', i.e. unrescaled, Ehrenfest models converge in distribution to a

simple symmetric random walk. Hence, if we take the lengths $(L_{n,k})k\in\mathbb{N}$ of the excursions from 0

as our basic regeneration intervals, we find that $L_{n,1}$ converges in distribution to the time it

takes a simple symmetric random walk to return to 0 (which is finite with probability 1, but has

infinite first moment). With respect to these excursions, the sets $B_n := \{ j \in \mathbb{N} : j \geq n^{1/2} \}$ are

asymptotically rare, in the sense explained in the previous paragraph. The main limit theorem

for Markov chains implies that $E_0 L_{n,1} \sim (\pi n)^{1/2}$ and it is easy to see that $L_{n,1} / n^{1/2}$ is not

uniformly integrable: as $L_{n,1}$ itself converges in distribution, $L_{n,1} / n^{1/2}$ would converge to 0 in

probability, and uniform integrability would give $E_0 L_{n,1} \to 0$, in contradiction to the above.
4. An application to approximate pattern matching

Suppose that we have a finite alphabet $\Sigma_1 = \{\sigma_1, \ldots, \sigma_d\}$ and a sequence $X = (X_i)_{i \in \mathbb{N}}$ of independent and identically distributed, $\Sigma_1$-valued random variables with $P(X_1 = \sigma_i) > 0$ for $i = 1, \ldots, d$. Suppose further that we are given a pattern $s = (s_1, \ldots, s_n) \in \Sigma_1^n$ of length $|s| = n$ from this alphabet; let

$$\tau(s) := \inf\{m \geq n : (X_{m-n+1}, \ldots, X_m) = s\}$$

be the time of the first occurrence of the pattern in the random string $X$. The analysis of this situation is one of the classical topics of applied probability (under the heading ‘monkey typing Shakespeare’ it even appears in some undergraduate probability courses). The expectation or, more generally, the distribution of $\tau(s)$ is treated in Feller (1968, Chapter XIII.7), Li (1980), Gerber and Li (1981), Guibas and Odlyzko (1981), and elsewhere. These results, especially the expressions obtained for the generating functions associated with pattern waiting times, can be used to show that, for any sequence $(s(n))_{n \in \mathbb{N}}$ of patterns with length $|s(n)|$ growing to infinity, $\tau(s(n))$ is asymptotically exponential in the sense that the distribution of $\tau(s(n))/E\tau(s(n))$ converges weakly to $\text{Exp}(1)$; see Rudander (1996).

Interest in this problem and its ramifications has increased over recent years due to its relevance to molecular biology, where $X$ might be the model for some genome sequence and $s$ corresponds to a particular gene; see, e.g. Waterman (1995). However, the results concerning exact occurrence seem to be of limited relevance in this area; instead, interest is in the statistical significance of observations of approximate occurrences of the given pattern. In this subsection, we show that our approach can be used, under certain conditions, to ‘bootstrap’ the result on asymptotic exponentiality of exact occurrence to asymptotic exponentiality of approximate occurrence.

The approach is quite flexible with respect to the distance concept involved. For definiteness, we consider the minimum suffix edit distance

$$d_S(s, (X_1, \ldots, X_m)) := \min_{1 \leq k \leq m} d_L(s, (X_k, \ldots, X_m)),$$

where $d_L(s, x)$ is the string edit or Levenshtein distance between $s$ and $x$, defined to be the minimum number of insert, delete, or replace operations needed to transform $s$ into $x$; see, e.g. Gusfield (1997). Note that while $d_L$ is a metric, $d_S$ is not, as we may have $d_S(a, b) = 0$ for $a \neq b$; also, $d_S$ is not symmetric. The $d_S$ values can be computed recursively. Starting with $d_S(\emptyset, a) = 0$ and $d_S(a, \emptyset) = |a|$, we use

$$d_S((a_1, \ldots, a_k), (x_1, \ldots, x_n)) = \min\{m_1 + 1, m_2 + 1, m_3 + \delta\},$$

where

$$m_1 = d_S((a_1, \ldots, a_{k-1}), (x_1, \ldots, x_n)),$$

$$m_2 = d_S((a_1, \ldots, a_k), (x_1, \ldots, x_{n-1})),$$

$$m_3 = d_S((a_1, \ldots, a_{k-1}), (x_1, \ldots, x_{n-1})),$$

$$\delta = \begin{cases} 0 & \text{if } a_k = x_n, \\ 1 & \text{if } a_k \neq x_n. \end{cases}$$

See Table 1 for an example with $x = (c, c, g, g, a, t, a, t, a, t, g, g, g, a, c, c, g, \ldots)$ and $s = (t, a, c, c, g)$. Inspection of Table 1 shows that the first match up to distance 1
between an $x$ suffix and $s$ occurs at $\rho = 17$ (the $\tau$ variables will be explained below). It is interesting to note that the successive columns below the central horizontal line can be regarded as a Markov chain, as is obvious from the recursion given above and the assumptions on the random string $X$. In this chain, $\rho$ would be the entry time into the set of all states that have component value 1 at the bottom of the column. For the numerical treatment of such problems, this is indeed a convenient construction; see Reich (2004). However, we will work with the original sequence $(X_i)_{i \in \mathbb{N}}$ and we will choose the regeneration times appropriately.

We write $a \subset b$ for strings $a$ and $b$ if $a$ appears as a contiguous substring of $b$; also, $a \circ b$ denotes the concatenation $(a_1, \ldots, a_m, b_1, \ldots, b_n)$ of $a = (a_1, \ldots, a_m) \in \Sigma^m$ and $b = (b_1, \ldots, b_n) \in \Sigma^n$. For later use, we note the following properties of the Levenshtein distance:

$$d_L(a, b) \geq |a| - |b|,$$

(8)

$$a = a' \circ a'' \Rightarrow \text{for all } b, \text{there exist } b' \text{ and } b'' \text{ such that } b = b' \circ b'' \text{ and } d_L(a, b) = d_L(a', b') + d_L(a'', b'').$$

(9)

Obviously, (9) can be extended by induction to concatenations of more than two strings.

Now let $s = (s_1, s_2, \ldots) \in \Sigma^\infty$ be an infinitely long pattern and let $s(n) = (s_1, \ldots, s_n)$ be its prefix of order $n$. In addition, let $(k(n))_{n \in \mathbb{N}}$ be a nondecreasing sequence of nonnegative integers and let

$$\rho_n := \inf \{ m \in \mathbb{N} : d_S(s(n), (X_1, \ldots, X_m)) \leq k(n) \}$$

be the first time that $s(n)$ appears in $X$ up to an approximation distance $k(n)$. Our condition for asymptotic exponentiality involves the notion of a splitting pattern for $(s(n), k(n))$, by which we mean a string $y$ with the property that

$$y \subset x \Rightarrow d_L(s(n), x) > k(n),$$

i.e. the edit distance between $s(n)$ and every finite sequence $x$ with contiguous substring $y$ is greater than $k(n)$. For example, using (9) we see that $y = (g, g, a)$ splits $s = (t, a, c, c, g)$ at distance level 1. The nonoverlapping occurrences of $y$ partition $X$ into regenerative cycles; in Table 1, $\tau_1$ and $\tau_2$ indicate the first two completed occurrences of $(g, g, a)$. The following result roughly says that, if the size of the regenerative cycles becomes negligible with respect to the length $n$ of the pattern, taking into account the desired distance $k(n)$ and the length of the splitting pattern, then asymptotic exponentiality holds.

### Table 1.

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>∅</td>
<td>c</td>
<td>c</td>
<td>g</td>
<td>g</td>
<td>a</td>
<td>t</td>
<td>a</td>
<td>t</td>
<td>a</td>
<td>t</td>
<td>g</td>
<td>g</td>
<td>g</td>
<td>a</td>
<td>c</td>
<td>c</td>
<td>g</td>
</tr>
<tr>
<td>$\tau_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\rho$</td>
<td>∅</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0021900200000425
Theorem 3. Suppose that there exists a sequence \((y(n))_{n \in \mathbb{N}}\) of splitting patterns \(y(n)\) for \((s(n), k(n))\) with nondecreasing length \(|y(n)|\) and the property that

\[
\lim_{n \to \infty} P(\tau(y(n)) \geq n - k(n) - |y(n)|) = 0.
\]

Then,

\[
\lim_{n \to \infty} P\left( \frac{\rho_n}{E\rho_n} \leq x \right) = 1 - e^{-x} \quad \text{for all } x \geq 0.
\]

Proof. Let \(\tau_{n,l}\) be the time of the \(l\)th completed nonoverlapping occurrence of \(y(n)\) in \(X\), with \(\tau_{n,0} \equiv 0\). With \(E_{n,l} := (X_{\tau_{n,l}} - 1 + 1, \ldots, X_{\tau_{n,l}})\), we obtain a decomposition of \(X\) into independent and identically distributed regenerative cycles \(E_{n,1}, E_{n,2}, \ldots\), for any fixed \(n \in \mathbb{N}\). In addition, let \(\Sigma^* := \bigcup_{j=0}^{\infty} \Sigma^j\) be the set of all finite strings with letters from \(\Sigma\) and let

\[B_n := \{ x \in \Sigma^* : d_S(s(n), y(n) \circ (x_1, \ldots, x_i)) \leq k(n) \text{ for some } i \in [1, \ldots, |x|]\}.
\]

With \(\mathcal{L}(X)\) and \(\mathcal{L}(X | A)\) denoting the distribution of \(X\) and the conditional distribution of \(X\) given \(A\), respectively, our dictionary for the present situation is given by

\[
q_n = P(E_{n,1} \in B_n), \quad \mu_n = \mathcal{L}(|E_{n,1}|),
\]

\[
\mu_{n,1} = \mathcal{L}(|E_{n,1}| | E_{n,1} \notin B_n), \quad \mu_{n,2} = \mathcal{L}(|E_{n,1}| | E_{n,1} \in B_n).
\]

In order to have \(E_{n,1} \in B_n\), it is necessary that

\[
d_S(s(n), y(n) \circ (X_1, \ldots, X_i)) \leq k(n) \quad \text{for some } i \leq \tau_{n,1}.
\]

The length of a suffix of the second string is bounded from above by \(|y(n)| + \tau_{n,1}\). By (8), the inequality therefore implies that

\[|y(n)| + \tau_{n,1} \geq n - k(n),\]

so that \(\lim_{n \to \infty} q_n = 0\) follows from \(\tau_{n,1} = \tau(y(n))\), the condition in the theorem, and (8).

For the uniform integrability condition, we distinguish between the two cases \(\sup_{n \in \mathbb{N}} |y(n)| < \infty\) and \(|y(n)| \uparrow \infty\). In the first case, we only have a finite number of possibilities for \(y(n)\); since \(E \tau(y) < \infty\) for every finite string \(y\), uniform integrability follows. For an unbounded sequence of patterns, we have asymptotic exponentiality by Rudander’s (1996) result mentioned above, and this implies uniform integrability as explained at the end of Section 2.

Theorem 1 now implies that

\[
\tilde{\rho}_n = \inf\{m \in \mathbb{N} : d_S(s(n), y(n) \circ (X_1, \ldots, X_m)) \leq k(n)\}
\]

is asymptotically exponential. It is here that we use the splitting condition: the first approximate occurrence of \(s(n)\) with suffix edit distance not exceeding \(k(n)\) takes place entirely within \(y(n) \circ E_{n,l}\), for some \(l \in \mathbb{N}\).

The variable \(\tilde{\rho}_n\) differs from \(\rho_n\) because of the concatenation from the left with \(y(n)\) used for the first regenerative cycle. (Indeed, we chose our example above in such a manner that we
would have \( \tilde{\rho} = 3 \) with \( y = (g, g, a) \); with \( y = (g, g, g) \), which also satisfies the splitting condition, we have \( \tilde{\rho} = \rho = 17 \). However, we have
\[
P(\tilde{\rho}_n \neq \rho_n) = P(E_{n, 1} \in B_n) = q_n \to 0 \quad \text{as } n \to \infty.
\]
Furthermore,
\[
E \tilde{\rho}_n \leq E \rho_n \leq E \tau_{n, 1} + E \tilde{\rho}_n,
\]
where the second inequality can be obtained by considering the post-\( \tau_{n, 1} \) process \( (X_{i, 1})_{i \in \mathbb{N}} \), which has the same distribution as the original \((X_i)_{i \in \mathbb{N}} \). Theorem 1 also gives
\[
q_n \frac{E \tilde{\rho}_n}{E \tau_{n, 1}} \to 1 \quad \text{as } n \to \infty,
\]
so that, since \( q_n = o(1) \), we also have \( E \tau_{n, 1} = o(E \tilde{\rho}_n) \). Finally,
\[
\left| \frac{\rho_n}{E \rho_n} - \frac{\tilde{\rho}_n}{E \tilde{\rho}_n} \right| \leq 1_{(\tilde{\rho}_n \neq \rho_n)} \left| \frac{\rho_n - \tilde{\rho}_n}{E \rho_n} \right|,
\]
where \( 1_{(\cdot)} \) is an indicator function. The right-hand side converges to 0 in probability, because
\[
\lim_{n \to \infty} P(\tilde{\rho}_n = \rho_n) = 1.
\]
Using the familiar properties of convergence in distribution, together with
\[
\lim_{n \to \infty} E \tilde{\rho}_n / E \rho_n = 1,
\]
we now see that the asymptotic exponentiality of \( \tilde{\rho}_n \) implies the asymptotic exponentiality of \( \rho_n \).

If \( \Sigma \) contains a letter \( \sigma_0 \) that does not appear in \( s \), then we can take \( y(n) \) to be a run of \( \sigma_0 \) of length \( k(n) + 1 \); the results on runs and exact pattern matching mentioned at the beginning of this section imply that the condition in the theorem will be satisfied if \( k(n) = o(\log n) \). Runs of specific letters can also be used as splitting patterns if the number of occurrences of particular letters in substrings of \( s \) of length \( l \) grows sufficiently slowly as \( l \to \infty \).

Acknowledgement

We wish to thank the referee for constructive suggestions, which led to improvements in the contents and the readability of the paper.

References


