ILLINOIS JOURNAL OF MATHEMATICS
Volume 37, Number 2, Summer 1993

LARGE DEVIATIONS FOR NONSTATIONARY ARRAYS AND SEQUENCES[1]

J. R. BAXTER, N. C. JAIN AND T. O. SEPPÄLÄINEN

1. Introduction

In the present paper we shall prove several results which apply to empirical distributions and empirical processes for nonstationary sequences of random variables. Our first result, Theorem 5.1, which deals with triangular arrays, will be derived from a theorem of Kifer [8], which gives a criterion for the large deviation principle to hold. Kifer's result is stated below in a general form as Theorem 3.5. A geometrical proof of Theorem 3.5 is given in [4]. Theorem 5.1 applies in particular to arrays of independent variables, as is pointed out in Corollary 5.4. Another criterion for the large deviation principle to hold is given in Theorem 6.5, which is a generalization of a result proved in [4]. Applications of Theorem 6.5 are given in Corollaries 7.1 and 8.1. Corollary 8.1 implies a large deviation result in the nonstationary hypermixing case, Theorem 9.13. In Section 10 it is shown that the results of this paper can be applied to the case of an independent sequence whose distributions are quasi-regular, in particular when the distributions are generated by a stationary random process.

A compactification argument will be used in the proofs of Theorems 5.1 and Corollaries 7.1 and 8.1. This step uses some simple compactification results from [4], which are stated in Proposition 4.1 and Proposition 4.9.

2. The LDP

Throughout this paper, for any σ-algebra $\mathscr{F}$, we let $\mathscr{M}(\mathscr{F})$, $\mathscr{M}_+(\mathscr{F})$, and $\mathscr{M}_1(\mathscr{F})$ denote the space of all bounded signed measures on $\mathscr{F}$, the space of all bounded nonnegative measures on $\mathscr{F}$, and the space of all probability measures on $\mathscr{F}$, respectively. By a scaling sequence $(r(n))$ we will mean a sequence of positive integers such that $r(n) \to \infty$ as $n \to \infty$. We will begin by stating the large deviation principle in a suitable form.

Received July 6, 1991
1991 Mathematics Subject Classification. Primary 60F10, 60B12.
[1]This work was supported in part by a grant from the National Science Foundation.

DEFINITION 2.1 (LARGE DEVIATION PRINCIPLE.) Let $(Y, \mathscr{S})$ be any measurable space. Let $\mathscr{T}$ be any Hausdorff topology on Y. Let (μ_n) be a sequence in $\mathscr{M}_1(\mathscr{S})$ and let $(r(n))$ be a scaling sequence. Let $I: Y \to [0, \infty]$ be a lower semicontinuous function. We will say that the large deviation principle holds for the sequences $(\mu_n), (r(n))$, with rate function I, using the topology $\mathscr{T}$, if for any set $A \in \mathscr{S}$ with closure $\bar{A}$ and interior A°,

$$\limsup_{n\to\infty} \frac{1}{r(n)} \log \mu_n(A) \leq -\inf_{y\in\bar{A}} I(y), \tag{2.2}$$

$$\liminf_{n\to\infty} \frac{1}{r(n)} \log \mu_n(A) \geq -\inf_{y\in A^\circ} I(y). \tag{2.3}$$

Condition (2.2) is called the upper bound, and condition (2.3) is called the lower bound. By definition, a rate function in this paper is always lower semicontinuous. If it happens that $\{I \leq c\}$ is a compact set for all $c \in R$, then we will say that I is a good rate function. If the scaling sequence is not mentioned, we take it by default to be the sequence $r(n) = n$.

We will use abstract spaces and topologies in our work, essentially because our method involves a compactification of the original space under consideration. In concrete examples we must then show that our topology agrees with the standard one. For example, if our original space is the space of probability measures on a Polish space, the topology we use on the compactification should induce a topology on the original space that is consistent with the Prohorov metric. The next proposition shows that for applications we need only check that our abstract topology is at least as fine as the standard topology, because a finer topology cannot give a weaker result. Indeed we may obtain a stronger result in this way, but that is not the main motivation of this approach.

PROPOSITION 2.4. *Let $(Y, \mathscr{S})$ be a measurable space. Let $\mathscr{T}$, $\mathscr{T}_c$ be Hausdorff topologies on Y such that $\mathscr{T}_c \subset \mathscr{T}$ and for every $V \in \mathscr{T}_c$ and every $y \in V$ there exists a set $U \in \mathscr{S} \cap \mathscr{T}_c$, such that $y \in U$ and the closure $\bar{U}$ of U is a subset of V. Let (μ_n) be a sequence in $\mathscr{M}_1(\mathscr{S})$ and let $(r(n))$ be a scaling sequence. Suppose that the large deviation principle holds for the sequences (μ_n), $(r(n))$, using the topology $\mathscr{T}$, with good rate function J. Then J is lower semicontinuous with respect to $\mathscr{T}_c$, and J is the unique rate function such that the large deviation principle holds for the sequences (μ_n), $(r(n))$, using the topology $\mathscr{T}_c$.*

The proposition is a straightforward consequence of the definitions, so the proof is omitted.

3. The Pressure

We will now recall some terminology and basic results from [4]. Our results will always involve a locally convex topological space in some way, so we fix the notations for this space in the next statement, which describes our general setting. This setting is similar to that used in [1]. For other results using topological vector spaces see [2].

SETTING 3.1. *Assume that a real vector space Y_1 and a convex subset of Y_1, called Y, are given. Let a linear space $\mathbf{L}$ of linear functionals on Y_1 be given, which separates the points of Y_1. Let $\mathscr{T}_1$ be the weakest topology with respect to which every functional in $\mathbf{L}$ is continuous. Let $\mathscr{G}_1 = \sigma(\mathbf{L})$. Let $\mathscr{T} = \{A \cap Y: A \in \mathscr{T}_1\}$, $\mathscr{G} = \{A \cap Y: A \in \mathscr{G}_1\}$.*

We note that with the topology $\mathscr{T}_1$, Y_1 becomes a locally convex linear topological space, and its dual is simply $\mathbf{L}$. Also, if it happens that there is a countable collection L_0 of functionals in $\mathbf{L}$ such that any functional in L can be approximated uniformly on Y by members of L_0, then $\mathscr{T}$ is metrizable, and $\mathscr{G} = \sigma(\mathscr{T})$.

When working with the setting of Setting 3.1, we will usually be given a sequence (μ_n) in $\mathscr{M}_1(\mathscr{G})$ and a scaling sequence $(r(n))$. In this case, for any $\varphi \in \mathbf{L}$, we will define

$$\overline{\mathscr{P}}(\varphi) = \limsup_{n\to\infty} \frac{1}{r(n)} \log \int e^{r(n)\varphi}\, d\mu_n, \tag{3.2}$$

$$\underline{\mathscr{P}}(\varphi) = \liminf_{n\to\infty} \frac{1}{r(n)} \log \int e^{r(n)\varphi}\, d\mu_n. \tag{3.3}$$

When $\overline{\mathscr{P}} = \underline{\mathscr{P}}$ on $\mathbf{L}$ we will say (following the terminology of [8]) that the pressure exists, and we will define the pressure $\mathscr{P}$ to be the common value of $\overline{\mathscr{P}}$ and $\underline{\mathscr{P}}$. As in the statement of the large deviation principle, if the scaling sequence $(r(n))$ is not mentioned, we will take it by default to be $r(n) = n$. When $\mathscr{P}$ exists, we will define the Legendre transform J by

$$J(y) = \sup\{\varphi(y) - \mathscr{P}(\varphi) \colon \varphi \in \mathbf{L}\}. \tag{3.4}$$

An illustration of these general definitions is given in Setting 4.7.

An appropriate form of Kifer's theorem for our purposes is the following (cf. [4]).

THEOREM 3.5. *Let the assumptions of Setting 3.1 hold, and let Y be compact. Let a sequence (μ_n) in $\mathscr{M}_1(\mathscr{G})$ and a scaling sequence $(r(n))$ be given. Suppose that $\overline{\mathscr{P}} = \underline{\mathscr{P}} = \mathscr{P}$. Let J be defined by (3.4). Then $J \equiv \infty$ on*

$Y_1 - Y$, *and for each* $\varphi \in \mathbf{L}$,

$$\mathscr{P}(\varphi) = \sup\{\varphi(y) - J(y): y \in Y\}. \tag{3.6}$$

Suppose that for each $\varphi \in \mathbf{L}$, there is a unique $x \in Y$ such that

$$\varphi(x) - J(x) = \sup\{\varphi(y) - J(y): y \in Y\}, \tag{3.7}$$

or, equivalently, suppose that for each $\varphi \in \mathbf{L}$ there is a unique $x \in Y$ such that x is tangent to $\mathscr{P}$ at φ. Then the large deviation principle holds for the sequences (μ_n), $(r(n))$, with rate function J, using the topology $\mathscr{T}$.

For references to earlier results involving the pressure functional, see [8].

4. Compactification

Theorem 3.5 is stated for a compact space Y. This suggests that we should examine the extent to which a noncompact space can be compactified by adding ideal elements, so that a large deviation result on the compactification can be inherited by the original space. The next result is straightforward. It is proved in [4] as part of Theorem 1.21 of that paper.

PROPOSITION 4.1. *Let* $(Y, \mathscr{S})$ *be any measurable space. Let* $\mathscr{T}$ *be any Hausdorff topology on* Y, *such that for every point* $y \in Y$ *and every* $V \in \mathscr{T}$ *with* $y \in V$, *there exists a set* $U \in \mathscr{T} \cap \mathscr{S}$ *with* $y \in U$ *and* $U \subset V$. *Let* Y_0 *be a subset of* Y, *let* $\mathscr{S}_0 = \{A \cap Y_0 \colon A \in \mathscr{S}\}$, *and let* $\mathscr{T}_0 = \{A \cap Y_0 \colon A \in \mathscr{T}\}$. *Let a sequence* (ν_n) *in* $\mathscr{M}_1(\mathscr{S}_0)$ *and a scaling sequence* $(r(n))$ *be given. Let* μ_n *be the measure on* $\mathscr{S}$ *obtained from* ν_n *by setting*

$$\mu_n(A) = \nu_n(A \cap Y_0) \tag{4.2}$$

for every $A \in \mathscr{S}$. *Suppose that the large deviation principle holds for the sequences* (μ_n), $(r(n))$, *using the topology* $\mathscr{T}$, *with good rate function* I. *If*

$$I \equiv \infty \text{ on } Y - Y_0, \tag{4.3}$$

then the large deviation principle holds for the sequences (ν_n), $(r(n))$, *using the topology* $\mathscr{T}_0$, *with good rate function* I_0, *where* I_0 *is the restriction of* I *to* Y_0.

In applying Proposition 4.1 it is natural to consider when Equation (4.3) will hold. First we consider the notion of exponential tightness.

DEFINITION 4.4. Let $(Y, \mathscr{S})$ be any measurable space. Let $\mathscr{T}$ be any Hausdorff topology on Y. Let (μ_n) be a sequence in $\mathscr{M}_1(\mathscr{S})$ and let $(r(n))$ be a scaling sequence. We will say that exponential tightness holds for the sequences (μ_n), $(r(n))$, using the topology $\mathscr{T}$, if for any real number c, there exists a compact set K such that

$$\limsup_{n\to\infty} \frac{1}{r(n)} \log \mu_n(K^c) \leq -c. \tag{4.5}$$

The next proposition is immediate from the definitions.

PROPOSITION 4.6. *Under the assumptions of Proposition* 4.1, *if exponential tightness holds for the sequences* (ν_n), $(r(n))$, *using the topology* $\mathscr{T}_0$ *on* Y_0, *then condition* (4.3) *holds, and the large deviation principle holds for the sequence* (ν_n), $(r(n))$, *using the topology* $\mathscr{T}_0$, *with good rate function* I_0.

A slightly less trivial condition can also be given. First we shall formally describe our basic setting for results concerning occupation measures.

SETTING 4.7. *Let* $(\Psi, \mathscr{D})$ *be a measurable space, and* W *be a vector space of bounded* $\mathscr{D}$*-measurable functions on* Ψ, *containing the constant functions, such that* $\sigma(W) = \mathscr{D}$. *We consider the uniform closure* V *of* W *as a Banach space with the supremum norm, and let* Y_1 *denote the dual space. For each* $f \in W$, *let* φ_f *be the functional on* Y_1 *defined by* $\varphi_f(y) = y(f)$, *and let* $\mathbf{L}$ *be the space of such functionals. Let* $\mathscr{T}_1$ *be the topology on* Y_1 *induced by* $\mathbf{L}$. *Then* $\mathbf{L}$ *is the dual of the linear topological space* $(Y_1, \mathscr{T}_1)$. *Let* Y *be the set of all members of* Y_1 *which are positive and have norm* 1. *Then the assumptions of Setting* 3.1 *hold, if we define* $\mathscr{S}_1$, $\mathscr{T}$, *and* $\mathscr{S}$ *as in Setting* 3.1. *The topology* $\mathscr{T}$ *coincides with the weak*-topology on* Y, *so the space* $(Y, \mathscr{T})$ *is compact by Alaoglu's Theorem. We identify a signed measure in* $\mathscr{M}(\mathscr{D})$ *with the corresponding functional in* Y_1. *In this sense we regard* $\mathscr{M}_1(\mathscr{D})$ *as a subset* Y_0 *of* Y. *Let*

$$\mathscr{T}_0 = \{A \cap Y_0 : A \in \mathscr{T}_1\}, \quad \mathscr{S}_0 = \{A \cap Y_0 : A \in \mathscr{S}_1\}.$$

For brevity, we will often identify a function $f \in W$ *with the functional* $\varphi_f \in \mathbf{L}$ *which it induces. For example, we will often write* $\overline{\mathscr{P}}(\varphi_f)$ *as* $\overline{\mathscr{P}}(f)$, *and so on. If* γ *is a measure in* $\mathscr{M}_1(\mathscr{D})$, *the notations* $\varphi_f(\gamma)$, $\gamma(f)$, *and* $\int f\,d\gamma$ *will all have the same meaning.*

This setting is the natural specialization of Setting 3.1 to the case of occupation measures. A similar setting is considered in [1].

The space Y_0 of Setting 4.7 is the space of interest in most of the results of this paper. We introduce the more abstract space Y as a convenient compactification of Y_0.

Remark 4.8. When Ψ is a Polish space, it is easy to see that we may choose W to be the span of a countable set of bounded continuous functions, such that $\mathscr{T}_0$ is the topology induced by the Prohorov metric.

The next result gives a sufficient condition for Equation (4.3) to hold. It has a straightforward proof, given in [4].

PROPOSITION 4.9. *In Setting* 4.7, *assume that the uniform closure V of W is a lattice, that is, closed under finite sup and inf. Let (ν_n) be a sequence of measures in $\mathscr{M}_1(\mathscr{S}_0)$ and let $(r(n))$ be a scaling sequence. Assume that $\overline{\mathscr{P}} = \underline{\mathscr{P}} = \mathscr{P}$ holds in Setting* 4.7, *and that in addition*

$$\mathscr{P}(f_n) \searrow 0 \text{ for all } f_n \in W \text{ such that } f_n \searrow 0 \text{ pointwise}. \tag{4.10}$$

In this case we have

$$J \equiv \infty \text{ on } Y - Y_0,$$

where we define J as usual by (3.4).

Condition 4.10 will hold, for example, if there is a probability measure ρ, a continuous function G_1 on R, and a continuous function G_2 defined on an interval containing the range of G_1, such that

$$\mathscr{P}(f) \leq G_2\left(\int G_1(f)\, d\rho\right) \tag{4.11}$$

for all $f \in W$. In applications we will consider cases where

$$\mathscr{P}(f) \leq c_1 \log \int e^{c_2|f|}\, d\rho + c_3. \tag{4.12}$$

5. A large deviation theorem for arrays

We now can state our first large deviation result.

THEOREM 5.1. *Let $(\Psi, \mathscr{D})$ be a measurable space, and let $(r(n))$ be a sequence of positive integers with $r(n) \to \infty$ as $n \to \infty$. For each n, assume that ξ_j^n, $j = 1, \ldots, r(n)$, Ψ-valued random variables, are defined on the same probability space $(\Omega_n, \mathscr{F}_n, P_n)$, and let*

$$S_n = \delta_{\xi_1^n} + \cdots + \delta_{\xi_{r(n)}^n}.$$

Let W be an algebra of bounded measurable functions on Ψ which contains the *constant functions, such that* $\sigma(W) = \mathscr{D}$. *Let the notations of Setting* 4.7 *hold, in particular the definition of the topology* $\mathscr{T}_0$. *Then* $S_n/r(n)$ *is a random variable taking values in the measurable space* $Y_0 = \mathscr{M}_1(\mathscr{D})$, *with* σ*-algebra* $\mathscr{G}_0$. *Let* ν_n *denote the distribution of* $S_n/r(n)$.

Assume that the pressure $\mathscr{P}$ *exists on* W. *Let* V *denote the uniform closure of* W. *For every* $f \in V$, *and every positive integer* n, *let*

$$f[n] = f \circ \xi_1^n + \cdots + f \circ \xi_{r(n)}^n.$$

Then

$$\mathscr{P}(\varphi_f) = \mathscr{P}(f) = \lim_{n \to \infty} \frac{1}{r(n)} \log E_n e^{f[n]}$$

for each $f \in V$, *where* E_n *denotes expectation with respect to the probability measure* P_n. *Define* J *by* (3.4). *Let* J_0 *be the restriction of* J *to* Y_0.

Assume either that exponential tightness holds for the sequences (ν_n), $(r(n))$, *using the topology* $\mathscr{T}_0$, *or else that* (4.10) *holds. Then* J_0 *is a good rate function.*

For every $f \in V$, *and every positive integer* n, *let*

$$P_n^f = \frac{e^{f[n]} P_n}{E_n e^{f[n]}},$$

and let E_n^f *denote the corresponding expectation. For any square-integrable random variable* η *on* Ω_n, *let* $\mathrm{Var}_n^f(\eta)$ *denote the variance of* η *using the probability measure* P_n^f. *For* $f, g \in W$, *and real* t, *let*

$$b(t) = \sup_n \frac{1}{r(n)} \mathrm{Var}_n^{(f+tg)}(g[n]). \tag{5.2}$$

For all $f, g \in W$, *assume that* b *is a locally bounded function of* t. *Then the sequences* (ν_n), $(r(n))$ *satisfy the large deviation principle with rate function* J_0, *using the topology* $\mathscr{T}_0$.

We will prove Theorem 5.1 shortly. First we note:

Remark 5.3. Since by Proposition 4.1 a finer topology gives a stronger result, the strongest conclusion of the theorem holds when W is as large as possible, namely when W is the space of all bounded measurable functions on Ψ. However, the hypotheses may be easier to check when W is smaller, for example when W is the span of a countable set of functions.

Proof of Theorem 5.1. Fix $f, g \in W$. Let $u(t) = \mathscr{P}(f + tg)$ for any real t. By Hölder's Inequality, u is convex. For every real t, let

$$u_n(t) = \frac{1}{r(n)} \log E_n e^{(f+tg)[n]}.$$

Condition (5.2) implies that there is a constant c_4 such that

$$u_n''(t) \leq c_4$$

for all n and all $t \in (-1, 1)$. Consider any real numbers t_1, t_2, t_3, t_4 with

$$-1 < t_1 < t_2 < 0 < t_3 < t_4 < 1,$$

we see easily that

$$\frac{u_n(t_4) - u_n(t_3)}{t_4 - t_3} - \frac{u_n(t_2) - u_n(t_1)}{t_2 - t_1} \leq c_4(t_4 - t_1).$$

Letting $n \to \infty$ and allowing t_4 and t_1 to approach the origin shows that the function u is differentiable at 0. Of course the same proof shows that u is differentiable everywhere.

Let (μ_n) be defined by (4.2). We will show that the sequences (μ_n), $(r(n))$ satisfy the large deviation principle with rate J, using the topology $\mathscr{T}$ on Y. Indeed, let $x \in Y$, such that (3.7) holds, with $\varphi = \varphi_f$. For any real number t, let $h = f + tg$. By the definition of J we have

$$x(h) - J(x) \leq \mathscr{P}(h).$$

By (3.7) we have

$$x(f) - J(x) = \mathscr{P}(f).$$

Thus for all real t,

$$\mathscr{P}(f + tg) - \mathscr{P}(f) \geq tx(g).$$

It follows immediately that $x(g)$ is equal to the derivative of u at 0, and so x is unique. We can then use Theorem 3.5 to conclude that the large deviation principle holds for (μ_n). If exponential tightness holds, then (4.3) holds, by Proposition 4.6. Otherwise, since (4.10) holds, it follows from Proposition 4.9 that (4.3) holds. Thus Theorem 4.1 implies that the large deviation principle also holds for (ν_n), so the theorem is proved.

A useful consequence of Theorem 5.1 is the following.

COROLLARY 5.4. *Suppose that the random variables $\xi_1^n, \dots, \xi_{r(n)}^n$ form an independent sequence for each n. Let ρ_j^n be the distribution of the random variable ξ_j^n. Suppose the pressure*

$$\mathscr{P}(f) = \lim_{n\to\infty} \frac{1}{r(n)} \sum_{j=1}^{r(n)} \log \int e^f \, d\rho_j^n$$

exists for each $f \in W$.

Suppose that there is a probability measure ρ on $(\Psi, \mathscr{D})$, and a real number c_3 such that for each $f \in W$,

$$\limsup_{n\to\infty} \frac{1}{r(n)} \sum_{j=1}^{r(n)} \int f \, d\rho_j^n \le c_3 \int |f| \, d\rho. \tag{5.5}$$

Let ν_n, J_0, $\mathscr{T}_0$ be defined as in Theorem 5.1. Then the large deviation principle holds for the sequences (ν_n), $(r(n))$, with rate function J_0, using the topology $\mathscr{T}_0$.

Proof. In this case we see using the concavity of the log function that (4.10) holds. Condition (5.2) is automatically satisfied, since the functions $g \in W$ are assumed to be bounded.

Remark 5.6. Let us take Ψ to be a Polish space, and let W denote the space of bounded continuous functions on Ψ. If we consider Corollary 5.4 in the special case that $\rho_j^n = \rho^n$ for all $j = 1, \dots, r(n)$, then the existence of the pressure is equivalent to the convergence of $\int f \, d\rho^n$ for all $f \in W$. In this case ρ^n converges weakly as $n \to \infty$, to a limit ρ. The large deviation principle in this situation was proved by different methods as Theorem 3 in [3].

6. A convexity property

Let $(Y, \mathscr{S})$ be any measurable space. Let $\mathscr{T}$ be any Hausdorff topology on Y, such that for every point $y \in Y$ and every $V \in \mathscr{T}$ with $y \in V$, there exists a set $U \in \mathscr{T} \cap \mathscr{S}$ with $y \in U$ and $U \subset V$. Let (μ_n) be a sequence in $\mathscr{M}_1(\mathscr{S})$ and let $(r(n))$ be a scaling sequence. For any $U \in \mathscr{T} \cap \mathscr{S}$, define

$$\overline{K}(U) = \limsup_{n\to\infty} \frac{1}{r(n)} \log \mu_n(U), \tag{6.1}$$

$$\underline{K}(U) = \liminf_{n\to\infty} \frac{1}{r(n)} \log \mu_n(U). \tag{6.2}$$

For any point $y \in Y$, let

$$\bar{\kappa}(y) = \inf\{\bar{K}(U) : U \in \mathscr{T} \cap \mathscr{G}, y \in U\}, \tag{6.3}$$

$$\underline{\kappa}(y) = \inf\{\underline{K}(U) : U \in \mathscr{T} \cap \mathscr{G}, y \in U\}. \tag{6.4}$$

We will call $\bar{\kappa}$ and $\underline{\kappa}$ the upper and lower size functions for the sequences (μ_n), $(r(n))$.

The next theorem is a generalization of a result in [4].

THEOREM 6.5. *Under the assumptions of Setting* 3.1, *let Y be compact. Let $\mathscr{I}$ be a general index set. For each $i \in \mathscr{I}$, suppose that a sequence (μ_n^i) in $\mathscr{M}_1(\mathscr{G})$ and a scaling sequence $(r^i(n))$ are given. Suppose that all the sequences (μ_n^i), $(r^i(n))$ induce the same pressure $\mathscr{P}$. Let J be defined by* (3.4). *Then $J \equiv \infty$ on $Y_1 - Y$, and* (3.6) *holds.*

For each $i \in \mathscr{I}$, let $\bar{\kappa}^i$, $\underline{\kappa}^i$ be the upper and lower size functions for the sequences (μ_n^i), $(r^i(n))$. Let

$$\kappa_0 = \inf_{i \in \mathscr{I}} \underline{\kappa}^i.$$

Suppose that $-\kappa_0$ is convex. Then for each $i \in \mathscr{I}$, the large deviation principle holds for the sequence (μ_n^i), $(r^i(n))$, with rate function J.

Proof. The proof is almost the same as that given in [4] for the special case in which $\mathscr{I}$ consists of a single point. First we note that $J = \infty$ on $Y_1 - Y$ and (3.6) hold by the usual formula for the inverse Legendre transform, or by a standard direct argument (cf. [4], proof of Theorem 10.1 of that paper). Also, as a simple consequence of the definitions of the upper and lower size functions we have

$$J \leq -\bar{\kappa}^i \leq -\underline{\kappa}^i \leq -\kappa_0$$

elsewhere on Y, for each $i \in \mathscr{I}$ (cf. [4], Lemma 2.14).

Let $i \in \mathscr{I}$ be fixed. We consider a typical affine function which is below J. That is, let $\varphi \in \mathbf{L}$, and let c be a real number, such that $\varphi + c \leq J$ everywhere on Y. Let $g = \varphi + c$. Let $A = \{g = J\}$. Suppose that A is nonempty (notice that, by compactness, if A is empty we can increase c until A is nonempty). Then A is convex since $J - g$ is convex, and A is compact since $J - g$ is lower semicontinuous and $A = \{J - g \leq 0\}$. Let x be an extreme point of A. By Theorem 4.8 and Lemma 4.9 of [4], we know that $J(x) = -\underline{\kappa}^i(x)$. Since i was arbitrary, we have shown that $J(x) = -\kappa_0(x)$. It follows that

$$\varphi \in \mathbf{L}, c \in R, \varphi + c \leq J \Leftrightarrow \varphi \in \mathbf{L}, c \in R, \varphi + c \leq -\kappa_0.$$

Since both functions J and $-\kappa_0$ are convex and lower semicontinuous, they are equal. Thus we have shown that for each $i \in \mathscr{I}$,

$$J = -\bar{\kappa}^i = -\underline{\kappa}^i.$$

Since Y is compact, it is then a standard argument to verify that the large deviation theorem holds for each i, so the theorem is proved.

7. Empirical distributions

COROLLARY 7.1. *Let $(\Psi, \mathscr{D})$ be a measurable space, and let ξ_n, $n = 1, 2, \ldots$, be a sequence of Ψ-valued random variables defined on a probability space $(\Omega, \mathscr{F}, P)$. For nonnegative integer i and every positive integer n, let*

$$S_n(i) = \delta_{\xi_{in+1}} + \cdots + \delta_{\xi_{(i+1)n}}.$$

Let W be an algebra of bounded measurable functions on Ψ which contains the constant functions, such that $\sigma(W) = \mathscr{D}$. Let all the notations of Setting 4.7 *hold. Then $S_n(i)/n$ is a random variable taking values in the measurable space $Y_0 = \mathscr{M}_1(\mathscr{D})$, with σ-algebra $\mathscr{G}_0$. Let ν_n^i denote the distribution of $S_n(i)/n$.*

Assume that for each sequence (ν_n^i), the pressure $\mathscr{P}$ exists and does not depend on i. Define J by (3.4). *Let J_0 be the restriction of J to Y_0.*

Assume either that exponential tightness holds for the sequence (ν_n^0), or that (4.10) *holds. Then J_0 is a good rate function on Y_0.*

Assume that for every nonnegative integer i, and every $x, y \in Y$,

$$\underline{\kappa}^i(\bar{x}) \geq \frac{\underline{\kappa}^{2i}(x)}{2} + \frac{\underline{\kappa}^{2i+1}(y)}{2}, \tag{7.2}$$

where $\bar{x} = x/2 + y/2$.

Then for each i, the large deviation principle holds for the sequence (ν_n^i), with rate function J_0, using the topology $\mathscr{T}_0$.

Proof. Let $\mathscr{I}$ denote the set of nonnegative integers i. Define the sequences (μ_n^i) by

$$\mu_n^i(A) = \nu_n^i(A \cap Y_0) \tag{7.3}$$

for all $A \in \mathscr{G}$. The conditions of Theorem 6.5 are easily seen to be satisfied, so the large deviation principle holds for each sequence (μ_n^i).

We can pass from the large deviation principle on Y to the large deviation principle on Y_0 just as in the proof of Theorem 5.1, so the corollary is proved.

We can apply this corollary to get a result for the empirical distributions of an independent sequence, much as in Corollary 5.4 above. Incidentally, it is a straightforward matter to show that in the independent case, if the pressure $\mathscr{P}_0$ exists for the sequence (ν_n^0), then the pressure $\mathscr{P}_i$ also exists for every sequence (ν_n^i), and has the same value for all i. Indeed, if ρ_j denotes the distribution of ξ_j, we have

$$\frac{1}{n(i+1)} \sum_{j=1}^{n(i+1)} \log \int e^f \, d\rho_j = \frac{i}{i+1} \frac{1}{ni} \sum_{j=1}^{ni} \log \int e^f \, d\rho_j + \frac{1}{i+1} \frac{1}{n} \sum_{k=1}^{n} \log \int e^f \, d\rho_{ni+k},$$

and the statement follows.

8. Empirical processes

We can also obtain a process level result from Theorem 6.5, which we now state.

COROLLARY 8.1. *Let $(X, \mathscr{B})$ be a measurable subset of a Polish space, together with its Borel σ-algebra. Let η_n, $n = 1, 2, \ldots,$ be an X-valued sequence of random variables defined on a sample space $(\Omega, \mathscr{F}, P)$. Define $\Psi = X \times X \times \ldots, \mathscr{D} = \mathscr{B} \times \mathscr{B} \times \ldots$. Let $\mathscr{D}_k$ be the σ-algebra generated by the first k coordinates on Ψ. For each positive integer n, let*

$$\xi_n = (\eta_n, \eta_{n+1}, \ldots).$$

For each positive integer n and each nonnegative integer i, let

$$S_n(i) = \delta_{\xi_{in+1}} + \cdots + \delta_{\xi_{(i+1)n}}.$$

For each positive integer k, let W_k be an algebra of bounded functions on Ψ such that $\sigma(W_k) = \mathscr{D}_k$ and W_k contains the constant functions. Suppose that $W_k \subset W_{k+1}$ and let W be the union of the spaces W_k. Let the notations of Setting 4.7 hold. Then $S_n(i)/n$ is a random variable taking values in the measurable space $Y_0 = \mathscr{M}_1(\mathscr{D})$, with σ-algebra $\mathscr{G}_0$. Let ν_n^i be the distribution of $S_n(i)/n$.

Assume that the pressure $\mathscr{P}$ exists for each sequence (ν_n^i) and is the same for all i. Define J by (3.4). Let J_0 be the restriction of J to Y_0.

Assume either that exponential tightness holds for the sequence (ν_n^0) *or that for each positive integer k*

$$\mathscr{P}(f_n) \searrow 0 \text{ for all } f_n \in W_k \text{ such that } f_n \searrow 0 \text{ pointwise}. \tag{8.2}$$

Then J_0 *is a good rate function on* Y_0.

Assume that for every nonnegative integer i, *and every* $x, y \in Y$, *condition* (7.2) *holds*.

Then for each i, *the large deviation principle holds for the sequence* (ν_n^i), *with rate function* J_0, *using the topology* $\mathscr{T}_0$.

Proof. It follows at once from the theorem that the large deviation principle holds for each sequence (μ_n^i), with rate J, where (μ_n^i) is defined by (7.3). If exponential tightness holds we use Proposition 4.6 as usual. Otherwise, we will follow the argument in [4], Section 4. Let $x \in Y$ such that $J(x)$ is finite. By Proposition 4.9 there exists a probability measure γ_k on $\mathscr{D}_k$ such that

$$\int f \, d\gamma_k = x(f)$$

for all $f \in W_k$. Kolmogorov's theorem then ensures that there is a single probability measure γ on $\mathscr{D}$ which agrees with all the γ_k. Thus $x = \gamma \in Y_0$. It follows that $J \equiv \infty$ on $Y - Y_0$, and the corollary follows from Proposition 4.1.

9. The hypermixing case

Corollary 8.1 is phrased in a rather abstract form. To give a special case more explicitly, we may assume that the process is hypermixing, in the sense of the theorem of Chiyonobu and Kusuoka [5], but without assuming that it is stationary. We will follow the notation of [6] in what follows. A process will be said to be hypermixing if conditions (H-1) and (H-2) of Section 5.4 of [6] hold. We will recall these conditions shortly. First we state some notations. For any nonnegtive integers j, k with $j \le k$, we define the $\mathscr{F}[j, k]$ to be the σ-algebra generated by $\eta_j, \dots, \eta_k$, and we let $\mathscr{F}_n = \mathscr{F}[0, n]$.

For any positive integer l, and any functions $F_1, \dots, F_n$ on Ω, we will say that the functions $F_1, \dots, F_n$ are l-separated if there exist nonnegative integers $j(m), k(m), m = 1, \dots, n$, with $j(m) \le k(m)$ for $m = 1, \dots, n$, such that each of the intervals $[j(m), k(m)]$ is separated by a gap of at least l from the other intervals, and such that each function F_m is $\mathscr{F}[j(m), k(m)]$-measurable.

We assume that there is a nonnegative integer l_0, and three functions

$$\alpha, \beta\colon (l_0, \infty) \to [1, \infty) \quad \text{and} \quad \gamma\colon (l_0, \infty) \to [0, 1]$$

such that

$$\lim_{l\to\infty} \alpha(l) = 1, \tag{9.3}$$

$$\limsup_{l\to\infty} l(\beta(l) - 1) < \infty, \tag{9.4}$$

and

$$\lim_{l\to\infty} \gamma(l) = 0. \tag{9.5}$$

Condition (H-1) in [6] will now be stated. All L^p spaces will be with respect to $(\Omega, \mathscr{F}, P)$.

ASSUMPTION 9.6 (H-1). *For any $n \geq 2$, let $F_1, \dots, F_n$ be bounded l-separated functions on Ω. Then for all $l > l_0$,*

$$\|F_1 \dots F_n\|_1 \leq \prod_{m=1}^{n} \|F_m\|_{\alpha(l)}.$$

Condition (H-2) in [6] is as follows.

ASSUMPTION 9.7 (H-2). *For any bounded l-separated functions F and G on Ω,*

$$\left|\int (F - \bar{F}) G \, dP\right| \leq \gamma(l) \|F\|_{\beta(l)} \|G\|_{\beta(l)},$$

where $\bar{F}$ denotes $\int F \, dP$.

LEMMA 9.8. *Condition (H-2) implies* (7.2).

Proof. This is essentially the proof of Lemma 5.4.22 in [6], but for the convenience of the reader, we will sketch the argument. Let $U \in \mathscr{T} \cap \mathscr{G}$, such that $\bar{x} \in U$. Let $\varepsilon > 0$ and $f_1, \dots, f_s \in W$ be such that if

$$G = \{Z\colon z \in Y, |z(f_r) - \bar{x}(f_r)| < \varepsilon, r = 1, \dots, s\}$$

then $G \subset U$. We may assume that $|f_r| \leq 1$ everywhere, for $r = 1, \dots, s$. Fix i, and let

$$\bar{A}_n = \{S_n(i)/n \in G\}.$$

For each n, let $m = m(n)$ be the greatest integer less than or equal to $n/2$. It is easy to see that the quantity $f_r \circ \xi_{2mi+1} + \cdots + f_r \circ \xi_{2m(i+1)}$ can be obtained from the quantity $f \circ \xi_{ni+1} + \cdots + f \circ \xi_{n(i+1)}$ by adding at most i terms and subtracting at most $i + 1$ terms. Hence

$$|f_r \circ \xi_{2mi+1} + \cdots + f_r \circ \xi_{2m(i+1)} - f \circ \xi_{ni+1} - \cdots - f \circ \xi_{n(i+1)}| \leq 2i + 1.$$

It follows that if $n > (4i + 2)/\varepsilon$ we have $\bar{A}_n \supset \tilde{A}_n$, where

$$\tilde{A}_n = \{|f_r \circ \xi_{2mi+1} + \cdots + f_r \circ \xi_{2mi+2m} - mx(f_r) - my(f_r)| < m\varepsilon, r = 1, \ldots, s\},$$

and $m = m(n)$.

In this case $\tilde{A}_n \supset B_m \cap C_m$, where

$$B_m = \{|f_r \circ \xi_{2mi+1} + \cdots + f_r \circ \xi_{2mi+m} - mx(f_r)| < m\varepsilon/2, r = 1, \ldots, s\},$$
$$C_m = \{|f_r \circ \xi_{(2i+1)m+1} + \cdots + f_r \circ \xi_{2mi+2m} - my(f_r)| < m\varepsilon/2, r = 1, \ldots, s\}.$$

Let k be a positive integer such that every function f_r is $\mathscr{D}_k$-measurable. Suppose that n is large enough that $m(n)\varepsilon/32 > k$. Then we can choose, for each n, an integer $l = l(n)$ with

$$\frac{m(n)\varepsilon}{32} < l(n) < \frac{m(n)\varepsilon}{16}, \quad k + l(n) < \frac{m(n)\varepsilon}{8}.$$

For such n we have $B_m \supset \tilde{B}_m \supset \hat{B}_m$, where

$$\tilde{B}_m = \{|f_r \circ \xi_{2mi+1} + \cdots + f_r \circ \xi_{2mi+m-k-l} - mx(f_r)| < m\varepsilon/4, r = 1, \ldots, s\},$$
$$\hat{B}_m = \{|f_r \circ \xi_{2mi+1} + \cdots + f_r \circ \xi_{2mi+m} - mx(f_r)| < m\varepsilon/8, r = 1, \ldots, s\}.$$

Clearly

$$\liminf_{m\to\infty} \frac{1}{m} \log P(\hat{B}_m) \geq \underline{\kappa}^{2i}(x), \quad \liminf_{m\to\infty} \frac{1}{m} \log P(C_m) \geq \underline{\kappa}^{2i+1}(y). \quad (9.9)$$

If either $\underline{\kappa}^{2i}(x) = -\infty$ or $\underline{\kappa}^{2i+1}(y) = -\infty$ then (7.2) obviously holds. Thus we may assume without loss of generality that there is some real number R such that

$$P(\hat{B}_m) > e^{mR}, \quad P(C_m) > e^{mR} \tag{9.10}$$

for all m.

We consider n such that $m(n)\varepsilon/32 > \max(k, l_0)$ and $n > (4i + 2)/\varepsilon$, where l_0 is defined prior to (9.3). Then $l(n) > l_0$ and we can use Assumption 9.7 to conclude that

$$P(\tilde{B}_m \cap C_m) \geq P(\tilde{B}_m)P(C_m) - \gamma(l)\left(P(\tilde{B}_m)P(C_m)\right)^{1/\beta(l)}.$$

Thus

$$P(\tilde{B}_m \cap C_m) \geq P(\tilde{B}_m)P(C_m)(1 - \gamma(l)t),$$

where

$$t = \left(P(\tilde{B}_m)P(C_m)\right)^{-1+1/\beta(l)}.$$

Because of (9.10), (9.4), and the fact that $l > m\varepsilon/32$, we see easily that t is bounded in n. It follows that for large n,

$$P(\tilde{B}_m \cap C_m) \geq \tfrac{1}{2}P(\tilde{B}_m)P(C_m).$$

We then can use the definition of $\underline{K}(U)$ and (9.9) to conclude that

$$\underline{K}(U) \geq \tfrac{1}{2}\underline{\kappa}^{2i}(x) + \tfrac{1}{2}\underline{\kappa}^{2i+1}(y).$$

Taking the infimum on U proves the lemma.

If the process is stationary, Condition (H-1) easily implies that the pressure exists (cf. the proof of Lemma 5.4.13 in [6]). However since we do not assume that the process is stationary, we will make a separate assumption that the pressure exists.

ASSUMPTION 9.11. *The pressure $\mathscr{P}_0$ exists for the single sequence (ν_n^0).*

LEMMA 9.12. *Assumption* (9.11), *together with Assumption* (*H*-2), *imply that the pressure $\mathscr{P}_i$ exists for every sequence (ν_n^i), and that $\mathscr{P}_i$ is the same for all i.*

Proof. This is a messier version of the argument sketched in Section 7. Consider a function $f \in W_k$. Suppose that $|f| \leq c$ everywhere. Let $\varepsilon > 0$ be given. For each n such that $n\varepsilon > 1$ and $n > 2k$, choose $l = l(n)$ such that

$$n\varepsilon < l(n) < 2n\varepsilon, \quad l(n) + k < n.$$

Consider n such that $n\varepsilon > 1$, $n > 2k$, and $n\varepsilon > l_0$, where l_0 is defined prior to (9.3). Let $f(j) = f \circ \xi_j$. We have

$$\int e^{f(1)+\cdots+f(n(i+1))}\,dP \geq \int e^{f(1)+\cdots+f(ni-k-l)-(k+l)c+f(ni+1)+\cdots+f(n(i+1))}\,dP.$$

Also, by (H-2),

$$\int e^{f(1)+\cdots+f(ni-k-l)+f(ni+1)+\cdots+f(n(i+1))}\,dP$$
$$\geq \int e^{f(1)+\cdots+f(ni-k-l)}\,dP \int e^{f(ni+1)+\cdots+f(n(i+1))}\,dP - \gamma(l)v,$$

where

$$v = \left(\int e^{\beta(l)(f(1)+\cdots+f(ni-k-l))}\,dP \int e^{\beta(l)(f(ni+1)+\cdots+f(n(i+1)))}\,dP\right)^{1/\beta(l)}.$$

Since

$$\int e^{\beta(l)(f(1)+\cdots+f(ni-k-l))}\,dP \leq \int e^{f(1)+\cdots+f(ni-k-l)}\,dP\, e^{nic(\beta(l)-1)},$$
$$\int e^{\beta(l)(f(ni+1)+\cdots+f(n(i+1)))}\,dP \leq \int e^{f(ni+1)+\cdots+f(n(i+1))}\,dP\, e^{nc(\beta(l)-1)},$$

we see that we have

$$\int e^{f(1)+\cdots+f(n(i+1))}\,dP$$
$$\geq e^{-(k+l)c} \int e^{f(1)+\cdots+f(ni-k-l)}\,dP \int e^{f(ni+1)+\cdots+f(n(i+1))}\,dP(1 - \gamma(l)t),$$

where the quantity t is seen to be bounded in n because of (9.4) and the fact that $l > n\varepsilon/8$. Thus for large n we have

$$\int e^{f(1)+\cdots+f(n(i+1))}\,dP \geq \frac{e^{-2(k+l)c}}{2} \int e^{f(1)+\cdots+f(ni)}\,dP \int e^{f(ni+1)+\cdots+f(n(i+1))}\,dP.$$

The definition of $\mathscr{P}_0$ then shows that

$$\mathscr{P}_0(f) \geq -4\varepsilon c + \limsup_{n\to\infty} \frac{1}{n} \log \int e^{f(ni+1)+ \cdots +f(n(i+1))}\, dP.$$

A very similar argument shows that

$$P_0(f) \leq 4\varepsilon c + \liminf_{n\to\infty} \frac{1}{n} \log \int e^{f(ni+1)+ \cdots +f(n(i+1))}\, dP,$$

so we see that $\mathscr{P}_i(f)$ exists and is equal to $\mathscr{P}_0(f)$, proving the lemma.

THEOREM 9.13. (i) *Suppose that Assumptions* 9.7 *and* 9.11 *hold. Let* $\mathscr{P} = \mathscr{P}_0$, *and let J be defined by* (3.4), *and let J_0 be the restriction of J to Y_0. If exponential tightness holds for the sequence* (ν_n^0), *using the topology* $\mathscr{T}_0$, *then the large deviation principle holds for every sequence* (ν_n^i), *with rate function* J_0, *using the topology* $\mathscr{T}_0$.

(ii) *If X is a compact metric space, and if the functions in W are continuous with respect to the product topology on* Ψ, *then* $Y = Y_0$, *and in particular exponential tightness holds.*

(iii) *If Assumptions* 9.7, 9.11 *and* 8.2 *hold, then the conclusion of* (i) *holds.*

(iv) *Assumption* 8.2 *will hold if Assumption* 9.6 *holds and if for each k there exists a bounded measure β_k on $\mathscr{D}_k$, such that for every $f \in W_k$,*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \int f \circ \xi_j \, dP \leq \int |f|\, d\beta_k. \tag{9.14}$$

Proof. Most of the statements in this theorem follow easily from Corollary 8.1, or from the definitions. Only the final statement needs to be checked. We can easily see from Assumption 9.6 and (9.14) that for any positive integers k and l, with $l > l_0$, and for any function $f \in W_k$,

$$\mathscr{P}(f) \leq \frac{1}{\alpha(l)(k+l)} \log \int e^{\alpha(l)(k+l)f}\, d\beta_k.$$

(This is essentially Lemma 5.4.13 in [6].) Condition (8.2) then follows at once.

10. Independent sequences

We will next take a closer look at a particular case of hypermixing processes, namely independent sequences. We will use the notation of Sections 8 and 9 and apply the results of those sections. This gives a new proof of a large deviation theorem obtained in [10]. We shall then identify the rate function as a generalized specific entropy, using the argument in [10]. It is convenient to consider the case in which the sample space Ω is equal to the canonical product sample space Ψ, and so the product σ-algebras $\mathscr{D}_k$ and $\mathscr{F}_k$ coincide. We will use $\mathscr{F}$ exclusively here. For the sake of clarity we take X to be a Polish space, rather than just a Borel subset of a Polish space. Then Ψ is also a Polish space with the usual product metric. The second-countability of the topology guarantees that $\mathscr{F}$ coincides with the Borel σ-algebra of Ψ. Let η_j be the jth coordinate variable, so that $\mathscr{F}_n$ is generated by $\eta_1, \ldots, \eta_n$. Let θ denote the shift on the sequence space Ψ, so that

$$\eta_j \circ \theta = \eta_{j+1}.$$

Let $\rho = (\rho_j)$ be a sequence of Borel probability measures on X. Define the product measure P^ρ on the product space Ψ by

$$P^\rho = \rho_1 \otimes \rho_2 \otimes \rho_2 \otimes \cdots$$

Then (η_n) is a sequence of independent, X-valued random variables with distributions ρ_n, defined on the sample space $(\Psi, \mathscr{F}, P^\rho)$. Define ξ_n as in Section 8 by

$$\xi_n = (\eta_n, \eta_{n+1}, \ldots).$$

In the present case ξ_1 is the identity map, and $\xi_n = \theta^{n-1}$ for $n > 1$. The aim of this section is to describe the position-level and process-level large deviations of the process (η_n) under P^ρ.

The space $\mathscr{M}_1(\mathscr{B})$ of Borel probability measures on X is also a Polish space under the weak topology generated by the bounded continuous functions on X. Let us write $\Gamma = \mathscr{M}_1(\mathscr{B}) \times \mathscr{M}_1(\mathscr{B}) \times \cdots$ for the corresponding sequence space. Let $\mathscr{H}$ be the Borel σ-algebra of Γ. The shift map on Γ is denoted by T and defined by

$$T(\rho_1, \rho_2, \rho_3, \ldots) = (\rho_2, \rho_3, \rho_4, \ldots).$$

Clearly

$$\int f \circ \theta \, dP^\rho = \int f dP^{T\rho},$$

for any bounded measurable function f on Ψ.

It will be useful to consider the empirical processes on Γ, which are maps R_n from Γ into $\mathscr{M}_1(\mathscr{H})$ defined by

$$R_n(\rho) = \frac{1}{n}\sum_{k=0}^{n-1} \delta_{T^k\rho}.$$

Following the terminology of [9], let us call an element ρ of Γ quasi-regular if $R_n(\rho)$ converges in the weak topology of $\mathscr{M}_1(\mathscr{H})$. Denote the limit measure by π_ρ. The set of quasi-regular elements is a shift-invariant Borel subset of Γ, and on this set, $\rho \mapsto \pi_\rho$ is a shift-invariant, Borel measurable map with values in the set of shift-invariant probability measures on Γ. The large deviation behavior of sequences (η_n) under P^ρ turns out to be especially clear when P^ρ comes from a quasi-regular ρ. This provides a large class of examples of nonstationary sequences for which the large deviation principle holds.

For the process level result, let $S_n(i)$ be defined as in Corollary 8.1. We choose W_k to be the collection of bounded, continuous, $\mathscr{F}_k$-measurable functions on Ψ. The union W of the W_k generates the weak topology on $\mathscr{M}_1(\mathscr{F})$.

THEOREM 10.1. *Let ρ be a quasi-regular sequence and $f \in W$. Then the pressure $\mathscr{P}_i^\rho$ defined by*

$$\mathscr{P}_i^\rho(f) = \lim_{n\to\infty} \frac{1}{n} \log\int e^{f\circ\xi_{ni+1}+\cdots+f\circ\xi_{n(i+1)}}\,dP^\rho$$

exists for all i, and has a common value, independent of i, which we denote by $\mathscr{P}^\rho(f)$. The limit

$$\lim_{n\to\infty}\frac{1}{n}\int \log\int e^{f\circ\xi_1+\cdots+f\circ\xi_n}\,dP^\gamma\,\pi_\rho(d\gamma) \tag{10.2}$$

exists and is equal to $\mathscr{P}^\rho(f)$. Define $J^\rho(Q)$ for any $Q \in \mathscr{M}_1(\mathscr{F})$ by

$$J^\rho(Q) = \sup\left\{\int f\,dQ - \mathscr{P}^\rho(f) : f \in W\right\}.$$

Then the large deviation principle holds for the sequence

$$\left(P^\rho\left\{\frac{S_n(i)}{n} \in \cdot\right\}\right)$$

of distributions on $\mathscr{M}_1(\mathscr{F})$, with rate J^ρ, using the weak topology on $\mathscr{M}_1(\mathscr{F})$.

Proof. By independence, the hypermixing assumptions (H-1) and (H-2) of Section 9 are trivially satisfied. In view of parts (i), (iii), and (iv) of Theorem 9.13, we only need to verify the existence of the pressure $\mathscr{P}_0^\rho(f)$ for $f \in W$, and Condition (9.14).

Condition (9.14) is immediate, for by the quasi-regularity of ρ,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\int_\Psi f\circ\xi_j\,dP^\rho = \lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\int_\Psi f\,dP^{T^j\rho} = \int_\Gamma\left[\int_\Psi f\,dP^\gamma\right]\pi_\rho(d\gamma). \tag{10.3}$$

Here we used the fact that, for $f \in W$, the function

$$\rho \mapsto \int_\Psi f\,dP^\rho$$

is bounded and continuous on Γ.

To show the existence of the pressure, fix a function $f \in W$ and an integer r such that f is $\mathscr{F}_r$-measurable. Let c be the supremum norm of f. Let m, n be any positive integers with $m > r$ and $n \geq 2m$. For fixed m, let $k = k(n)$ be the largest positive integer k such that $(k+1)m \leq n$. Let $f(j)$ denote $f\circ\xi_j$. Let $g = f(1) + \cdots + f(m-r)$. Let $g(l) = g\circ\xi_l$. Fix a t with $1 \leq t \leq m$. It is easy to see that the quantity $g(t) + g(t+m) + \cdots + g(t+(k-1)m)$ is obtained from the quantity $f(1) + \cdots + f(n)$ by subtracting at most $kr + 2m$ terms. Hence

$$\begin{aligned} f(1) + \cdots + f(n) \geq -krc - 2mc + g(t) + g(t+m) \\ + \cdots + g(t+(k-1)m). \end{aligned}$$

For fixed t, the functions $g(t+lm)$ depend on disjoint sets of coordinates for distinct values of l, and hence they are independent under P^ρ. After taking exponentials, integrating, and then taking logs, we have

$$\log\int e^{f(1)+\cdots+f(n)}\,dP^\rho \geq -k(n)rc - 2mc + \sum_{l=0}^{k(n)-1}\log\int e^{g(t+lm)}\,dP^\rho.$$

Hence, averaging over $t = 1, \ldots, m$,

$$\log\int e^{f(1)+\cdots+f(n)}\,dP^\rho \geq -k(n)rc - 2mc + \frac{1}{m}\sum_{j=1}^{k(n)m}\log\int e^{g(j)}\,dP^\rho.$$

Dividing by n, noting that $k(n)/n \to 1/m$ as $n \to \infty$, and using the quasi-regularity of ρ, we obtain

$$\underline{\mathscr{P}}_0^\rho(f) \geq -\frac{rc}{m} + \frac{1}{m}\int F_{m-r}\, d\pi_\rho,$$

where, for any $\gamma \in \Gamma$ and any positive integer m,

$$F_m(\gamma) = \log\int e^{f\circ\xi_1 + \cdots + f\circ\xi_m}\, dP^\gamma. \tag{10.4}$$

In a very similar way we also have

$$\overline{\mathscr{P}}_0^\rho(f) \leq \frac{rc}{m} + \frac{1}{m}\int F_{m-r}\, d\pi_\rho,$$

and the theorem follows.

The next question to address is the identification of the rate function J^ρ. Fix a quasi-regular element ρ of Γ, and a shift-invariant probability Q on Ψ. For positive integers n, define

$$K_n^\rho(Q) = \sup\left\{\int f\,dQ - \int_\Gamma\left[\log\int_\Psi e^f\,dP^\gamma\right]\pi_\rho(d\gamma) : f \in W_n\right\},$$

and then let

$$k^\rho(Q) = \sup_n \frac{1}{n}K_n^\rho(Q).$$

LEMMA 10.5. *With ρ and Q as above, we have*

$$k^\rho(Q) = \lim_{n\to\infty}\frac{1}{n}K_n^\rho(Q).$$

Proof. Let $f \in W_m$ and $g \in W_n$. Then f and $g\circ\theta^m$ are independent under every measure P^γ for $\gamma \in \Gamma$, and $f + g\circ\theta^m$ is an element of W_{m+n}. Using the shift-invariance of Q and π_ρ, we get

$$\begin{aligned}K_{m+n}^\rho(Q) &\geq \int(f + g\circ\theta^m)\,dQ - \int_\Gamma\left[\log\int_\Psi e^{f+g\circ\theta^m}\,dP^\gamma\right]\pi_\rho(d\gamma)\\ &= \left\{\int f\,dQ - \int_\Gamma\left[\log\int_\Psi e^f\,dP^\gamma\right]\pi_\rho(d\gamma)\right\}\\ &\quad + \left\{\int g\,dQ - \int_\Gamma\left[\log\int_\Psi e^g\,dP^\gamma\right]\pi_\rho(d\gamma)\right\}.\end{aligned}$$

Since f and g were arbitrary, we get the superadditivity property

$$K^{\rho}_{m+n}(Q) \geq K^{\rho}_{m}(Q) + K^{\rho}_{n}(Q),$$

which implies the conclusion of the lemma.

THEOREM 10.6. *Let ρ be a quasi-regular element of Γ. The rate J^{ρ} is given by*

$$J^{\rho}(Q) = \begin{cases} k^{\rho}(Q) & \text{if } Q \text{ is shift-invariant;} \\ \infty & \text{otherwise.} \end{cases}$$

Proof. That the rate is identically infinite off the shift-invariant measures follows easily from the observation that $\mathscr{P}^{\rho}(g - g \circ \theta) = 0$ for any $g \in W$. The precise argument can be found in [4] or [6].

For the remainder of the proof, fix a shift-invariant probability Q on Ψ. Let $f \in W_r$, so that the function $f \circ \xi_1 + \cdots + f \circ \xi_m$ is an element of W_{m+r}. Recalling the definition (10.4) of F_m, and using the shift-invariance of Q, we have

$$K^{\rho}_{m+r}(Q) \geq m\int f\,dQ - \int F_m(\gamma)\pi_{\rho}(d\gamma).$$

Dividing by m and letting $m \uparrow \infty$ gives, by Theorem 10.1 and Lemma 10.5, that

$$k^{\rho}(Q) \geq \int f\,dQ - \mathscr{P}^{\rho}(f).$$

Since r and $f \in W_r$ were arbitrary, this shows that $k^{\rho}(Q) \geq J^{\rho}(Q)$.

Let us again consider any $f \in W_r$. By Hölder's inequality and the independence structure of P^{ρ}, we have

$$\frac{1}{mr}\log\int \exp\left(\sum_{k=1}^{mr} f \circ \xi_k\right) dP^{\rho} \leq \frac{1}{mr^2}\sum_{k=1}^{mr} \log\int e^{rf}\,dP^{T_k\rho},$$

from which, by quasi-regularity, we see that

$$\mathscr{P}^{\rho}(f) \leq \frac{1}{r}\int\left[\log\int e^{rf}\,dP^{\gamma}\right]\pi_{\rho}(d\gamma).$$

From the definition of J we then get

$$J^\rho(Q) \geq \int f\,dQ - \frac{1}{r}\int\left[\log\int e^{rf}\,dP^\gamma\right]\pi_\rho(d\gamma)$$
$$= \frac{1}{r}\left\{\int rf\,dQ - \int\left[\log\int e^{rf}\,dP^\gamma\right]\pi_\rho(d\gamma)\right\}.$$

Since $f \in W_r$ was arbitrary, we get

$$J^\rho(Q) \geq \frac{1}{r}K_r^\rho(Q),$$

and letting r vary gives $J^\rho(Q) \geq k^\rho(Q)$. This completes the proof of the theorem.

Example 10.7. Let p be a probability measure on X, and let $\sigma = (p, p, p, \dots)$ be the corresponding constant sequence. It is clear that $k^\rho(Q)$ reduces to the familiar specific relative entropy $h(Q|P^\sigma)$ against the i.i.d. measure

$$P^\sigma \equiv p \otimes p \otimes p \otimes \cdots$$

in case $\pi_\rho = \delta_\sigma$. Any $\rho = (\rho_1, \rho_2, \rho_3, \dots)$ satisfying

$$\lim_{j \to \infty} \rho_j = p,$$

in the weak topology of $\mathscr{M}_1(\mathscr{B})$, is an example of a quasi-regular ρ such that $\pi_\rho = \delta_\sigma$. (For a detailed account of specific relative entropy, see [6].)

Consider a model where the sequence (ρ_j) of distributions is generated by a stationary, $\mathscr{M}_1(\mathscr{B})$-valued stochastic process. Let V be the distribution of this process on Γ, so that we can think of the process as given by the coordinate projections $\rho \mapsto \rho_j$ under the shift-invariant probability V. By the ergodic theorem, the set of quasi-regular elements is of V-measure 1, and in fact the stochastic kernel π gives the ergodic decomposition of V: For any Borel subset A of Γ,

$$V(A) = \int \pi_\rho(A)V(d\rho).$$

We may then conclude that the large deviation principle holds for V-almost every ρ, with rate J^ρ.

A fundamental property of the ergodic decomposition kernel is that a shift-invariant probability measure V is ergodic if and only if

$$V\{\rho \in \Gamma: \pi_\rho = V\} = 1.$$

(For a proof, see [9] or [7].) For an ergodic V, the set $\{\rho \in \Gamma: \pi_\rho = V\}$ is called the quasi-ergodic set of V. Pick any ρ in this set, and then define

$$J^V(Q) = \begin{cases} k^\rho(Q) & \text{if } Q \text{ is shift-invariant;} \\ \infty & \text{otherwise.} \end{cases}$$

These remarks suffice to make the following theorem an immediate corollary of Theorem 10.6.

THEOREM 10.8. *Let V be a shift-invariant, ergodic probability measure on Γ. For V-almost all $\rho \in \Gamma$, the large deviation principle holds for the sequence*

$$\left(P^\rho\left\{\frac{S_n(i)}{n} \in \cdot\right\}\right)$$

of distributions on $\mathscr{M}_1(\mathscr{F})$, with rate J^V.

This and other related results for processes with nonstationary or randomly generated distributions can be found in [10] and [11]. Once the existence of the pressure has been established, there are a number of ways to complete the proof of a large deviation principle. Instead of using convexity and compactness as is done here, [10] proves the upper bound for Theorem 10.8 by an explicit verification of exponential tightness, and the lower bound by a Shannon-McMillan-type argument. These results extend also to independent variables indexed by an arbitrary lattice Z^d, and to Markov chains with randomly generated transition probabilities.

Let us now proceed to the corresponding position-level result. The random measures of interest are now the $\mathscr{M}_1(\mathscr{B})$-valued empirical distributions

$$L_n(i) = \frac{1}{n}\sum_{j=1}^{n} \delta_{\eta_{in+j}}.$$

Let W be the space of bounded continuous functions on X.

THEOREM 10.9. *Let $\rho \in \Gamma$ be such that the empirical distributions*

$$\frac{1}{n}\sum_{j=1}^{n} \delta_{\rho_j}$$

converge in the weak topology of $\mathscr{M}_1(\mathscr{M}_1(\mathscr{B}))$, *as* $n \to \infty$, *to some Borel probability measure* π_ρ° *on* $\mathscr{M}_1(\mathscr{B})$. *Define* $K^\rho: \mathscr{M}_1(\mathscr{B}) \to [0, \infty]$ *by*

$$K^\rho(q) = \sup\left\{\int f\,dq - \int_{\mathscr{M}_1(\mathscr{B})}\left[\log\int_X e^f\,dr\right]\pi_\rho^\circ(dr) : f \in W\right\}.$$

Then, for all i, the large deviation principle holds for the sequence

$$\left(P^\rho\{L_n(i) \in \cdot\}\right)$$

of distributions on $\mathscr{M}_1(\mathscr{B})$, *with rate* K^ρ.

The proof is similar but easier than the one given above for the process level result, so we shall omit it. The requirement that the occupation measures

$$\frac{1}{n}\sum_{j=1}^{n}\delta_{\rho_j}$$

converge is genuinely weaker than quasi-regularity. To give an easy example, construct ρ by repeating two sequences $(\alpha, \alpha, \beta, \beta)$ and $(\alpha, \beta, \alpha, \beta)$ in such a way that the relative frequency of the pair (α, α) does not converge. This violates quasi-regularity, but the relative frequencies of α and β each converge to $1/2$, so

$$\pi_\rho^\circ = \tfrac{1}{2}(\delta_\alpha + \delta_\beta).$$

Since Theorem 10.9 thus holds for a wider class of ρ's than Theorem 10.1, we cannot deduce the former from the latter by the usual contraction mapping, or push-forward, principle.

In case π_ρ° is a point mass δ_p, the functional K^ρ reduces to relative entropy against p. The above remarks and Theorem 10.8 about randomly generated sequences extend in an obvious way to position level. It is interesting to note that the position-level rate of the randomly generated sequence depends on the background measure V only through the marginal distribution of ρ_1.

References

[1] A. de Acosta, *Large deviations for empirical measures of Markov chains*, J. Theoretical Probability, vol. 3 (1990), pp. 395–431.

[2] R.R. Bahadur and S.L. Zabell, *Large deviations of the sample mean in general spaces*, Ann. Probab. vol. 7 (1979), pp. 587–621.

[3] J.R. Baxter and N.C. Jain, *A comparison theorem for large deviations*, Proc. Amer. Math. Soc., vol. 103 (1988), pp. 1235–1240.

[4] ______, *Convexity and compactness in large deviation theory*, preprint.

[5] C. Chiyonobu and S. Kusuoka, *The large deviation principle for hypermixing processes*, Probability Theory and Related Fields, vol. 78 (1988), pp. 627–649.

[6] J.-D. Deuschel and D. Stroock, *Large deviations*, Academic Press, San Diego, 1989.

[7] E.B. Dynkin, *The initial and final behaviour of the trajectories of Markov processes*, Russian Math. Surveys, vol. 26 (1971), pp. 165–185.

[8] Y. Kifer, *Large deviations in dynamical systems and stochastic processes*, Amer. Math. Soc., vol. 321 (1990), pp. 505–524.

[9] J.C. Oxtoby, *Ergodic sets*, Bull. Amer. Math. Soc., vol. 58 (1952), pp. 116–136.

[10] T.O. Seppäläinen, *Large deviations for processes with stationarily random distributions*, Ph.D. Thesis, University of Minnesota, 1991.

[11] ______, *Large deviations for conditioned and nonstationary lattice systems, I, II*, to appear in Probability Theory and Related Fields.

University of Minnesota
Minneapolis, Minnesota