STRONG LAWS FOR L- AND U-STATISTICS

J. AARONSON, R. BURTON, H. DEHLING, D. GILAT, T. HILL, B. WEISS

ABSTRACT. Strong laws of large numbers are given for L-statistics (linear combinations of order statistics) and for U-statistics (averages of kernels of random samples) for ergodic stationary processes, extending classical theorems of Hoeffding and of Helmers for iid sequences. Examples are given to show that strong and even weak convergence may fail if the given sufficient conditions are not satisfied, and an application is given to estimation of correlation dimension of invariant measures.

1. INTRODUCTION

One of the fundamental problems in statistics is the estimation of a parameter $\theta = \theta(F)$ of an unknown distribution F, based on functions of observations $X_1, X_2, \ldots$ from a statistical experiment (see e.g. [Le]). This article will consider the so called L-, and U-parameters (introduced in §2), which include certain of the following classical parameters:

Moments	$M_\alpha(F) = E(X^\alpha)$;
Central moments	$\sigma^\alpha(F) = E(|X - EX|^\alpha)$;
Generalized expected maxima	$P_\alpha(F) = m(F^\alpha)$;
Quantiles	$Q_\alpha(F) = F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}$;
Generalized Gini differences	$g_\alpha(F) = E(|X - \widehat{X}|^\alpha)$;

where X and $\widehat{X}$ are independent with distribution F; EX denotes the expected value of X; and $m(F)(= \int x dF(x))$ is the mean of the distribution F.

In the above notation, for example, $M_1(F) = P_1(F) =$ expected value of X, $\sigma^2(F) =$ variance of X, $P_2(F) = E(\max\{X, \widehat{X}\})$, $g_1(F) =$ Gini mean difference of X, and $Q_{\frac{1}{2}}(F) =$ smallest median of X.

Various functions (statistics) of the observations $X_1, X_2, \ldots$ of the underlying process can be used to estimate parameters, including the L- and U-statistics described below. If the sequence of observations is iid, much is known about the

limiting behavior of these statistics. On the other hand, iid realizations are some-times unrealistic, as is often the case when the observations come from real data which cannot be replicated in computer experiments (see §6).

It is the main purpose of this article to establish strong laws of large numbers for both L- and U-statistics for ergodic stationary processes (ESP).

Recall that a (real valued) *ergodic, stationary process* (ESP) with sample space $(\Omega, \mathcal{A}, P)$ is a stochastic sequence $(X_1, X_2, \dots)$ of form $X_k = f \circ T^k$ where T is an ergodic, probability-preserving transformation of the probability space $(\Omega, \mathcal{A}, P)$, and $f : \Omega \to \mathbb{R}$ is a measurable function. The *marginal* of the ESP is the distribution of X_1, and the ESP is called *integrable* if X_1 is integrable, and *bounded* if X_1 is (essentially) bounded.

The organization is as follows: §2 introduces L- and U-parameters and strong laws for their statistics; §3 and §4 establish the L-parameter and U-parameter strong laws of large numbers for ergodic stationary processes, respectively; §5 proves the strong law for U-statistics for weakly Bernoulli sequences; and §6 contains an application to dimension estimation.

2. L- AND U-PARAMETERS AND STATISTICS

Given a probability distribution function F on the real line $\mathbb{R}$, we denote by $F^{-1} : [0, 1] \to [-\infty, \infty]$ the *lower inverse* defined by $F^{-1}(0) = \operatorname{ess\,inf}(F)$, and for $u \in (0, 1]$,

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}.$$

Given a finite sequence $X_1, X_2, \dots, X_n$ of random variables, the *empirical distribution function* F_n of the random variables $X_1, X_2, \dots, X_n$ is the random probability measure determined by

$$F_n(x) := \frac{1}{n} \#\{i \leq n : X_i \leq x\},$$

and their *order statistics* $\{X_{n;i} : 1 \leq i \leq n\}$ are the values of the random variables in increasing order: $X_{n;1} \leq X_{n;2} \leq \cdots \leq X_{n;n}$.

Note that

$$F_n^{-1} = X_{n;1} 1_{\{0\}} + \sum_{i=1}^{n} X_{n;i} 1_{(\frac{i-1}{n}, \frac{i}{n}]},$$

where 1_A denotes the indicator function of the set A.

Definition 2.1. $\theta = \theta(F)$ is an *L-parameter* of F if there exists a *representing* (finite signed Borel) *measure* $\mu = \mu_\theta$ on $[0, 1]$ so that

$$\theta(F) = \theta_\mu(F) = \int_0^1 F^{-1} d\mu_\theta \text{ for all } F \text{ for which the integral is defined.}$$

Such representing measures are always unique, as can be established by evaluating θ for the distributions F of $\{0, 1\}$-valued random variables.

In case the representing measure μ is absolutely continuous (a.c.), $J = J_\mu$ will denote the Radon-Nikodym derivative $\frac{d\mu}{d\lambda}$ where λ (here and throughout) is Lebesgue measure. The class of all L parameters is denoted by $\mathcal{L}$.

Intuitively, an L-parameter is a parameter of a distribution which may be expressed as the a.s. limit of distribution-free linear combinations of the order statistics of the sample $X_1, X_2, \ldots, X_n$. Analogous definitions have been given in a variety of settings (see [Se] and references therein). Although technically $M_2(F)(= E(X^2))$ is not an L-parameter, it may easily be estimated using L-statistics based on the order statistics for $X_1^2, X_2^2, \ldots, X_n^2$ (see Example 2.2 below for the mean), and similar such straightforward extensions of the definition of L-parameter are left to the interested reader.

Example 2.2. For the classical parameters listed above, it is easily seen that the mean $M_1(F)$ and $P_1(F)$ are L-parameters with $J(u) \equiv 1$; $P_\alpha(F)$ is an L-parameter with $J(u) = \alpha u^{\alpha-1}$ for $\alpha \geq 1$; and the Gini mean-difference $g_1(F)$ is an L-parameter with $J(u) = 4u - 2$ (e.g. [Se, p.265]).

The main L-parameter result of this article is the next theorem, which extends the corresponding result for iid sequences (e.g. [He], [vZ]), to conclude that an L-parameter can be consistently estimated (in the a.s. sense) on the basis of linear combinations of order statistics of data (L-statistics) arising from ergodic stationary processes as well.

Definition 2.3. Given an L-parameter θ_μ, the *L-statistic* for θ_μ based on a sequence $X_1, \ldots, X_n$ is

$$L_\mu(X_1, \ldots, X_n) = \int_{[0,1]} F_n^{-1} d\mu = \mu(\{0\}) X_{n;1} + \sum_{i=1}^n \mu\left(\left(\frac{i-1}{n}, \frac{i}{n}\right]\right) X_{n;i}.$$

(The inclusion of interval endpoints is only relevant when μ has atoms.) The *L-parameter SLLN* is said to hold for $(X_k)_{k \in \mathbb{N}}$ and θ_μ if $L_\mu(X_1, \ldots, X_n) \to \theta_\mu$ P-a.s.

Theorem L (SLLN for L-statistics). *Let $(X_k)_{k \in \mathbb{N}}$ be an ergodic stationary process with marginal F, and let μ be an atomless finite signed Borel measure on $[0,1]$. If either:*
   (i) *$(X_k)_{k \in \mathbb{N}}$ is bounded; or*
   (ii) *$(X_k)_{k \in \mathbb{N}}$ is integrable, and μ is absolutely continuous with bounded density,*
then

$$(1) \qquad\qquad \lim_{n \to \infty} L_\mu(X_1, \ldots, X_n) = \theta_\mu(F) \qquad \text{P-a.s.}$$

The proof will be given in §3, along with examples to show the conclusions may fail without boundedness.

It is shown in [G-H, Example 3.1], that the L-parameter SLLN may fail even for iid sequences when the representing measure has atoms. As a complement to Theorem L(ii), there are L-parameters with a.c. representing measures for which the L-statistic SLLN fails for some integrable iid sequences (Example 3.2 below). Indeed, this failure is also of the corresponding weak law.

Next, U-parameters and their statistics will be introduced and the corresponding SLLN will be stated.

Definition 2.4. $\theta = \theta(F)$ is a *U-parameter* of F if there is a measurable function h, called the *kernel*, $h : \mathbb{R}^d \to \mathbb{R}$, so that

$$\theta(F) = \theta_h(F) = \int_{\mathbb{R}^d} h \, dF^{(d)} \qquad \text{for all F for which the integral is defined,}$$

where here (and throughout), $F^{(d)}$ is the product measure $F \times \cdots \times F$ on $\mathbb{R}^d$. The positive integer d is called the *order* of the kernel. Note that different kernels, with possibly different orders, may determine the same U-parameter. For example, if $h_1(x) = 2x$ and $h_2(x, y) = x + y$, then

$$\theta_{h_1}(F) = \theta_{h_2}(F) = 2 \int_{\mathbb{R}} x \, dF(x).$$

However, *symmetric* kernels of the same order which determine the same U-parameter coincide, which can be shown by evaluation of the parameters at those distributions supported on d (the order) points. The class of all U-parameters is denoted by $\mathcal{U}$.

A U-parameter is often called an *estimable* parameter, indeed $\mathcal{U}$ is exactly the class of parameters that can be estimated in an unbiased fashion (see [Le]).

Definition 2.5. Given a U-parameter θ_h, the *U-statistic* for θ_h based on a sequence $X_1, \ldots, X_n$ is

$$U_h(X_1, \ldots, X_n) = \frac{(n-d)!}{n!} \sum \{ h(X_{i_1}, \ldots, X_{i_d}) : \{i_j\} \text{ distinct}, 1 \le i_j \le n \}.$$

Many authors (e.g. [Se, p. 172]) assume (without loss of generality) that h is symmetric, in which case the U-statistic is also given by

$$\frac{1}{\binom{n}{d}} \sum_{1 \le i_1 < i_2 < \cdots < i_d \le n} h(X_{i_1}, \ldots, X_{i_d}).$$

The *U-parameter SLLN* holds for $(X_k)_{k \in \mathbb{N}}$ and θ_h if $U_h(X_1, \ldots, X_n) \to \theta_h$ P-a.s. The closely related *V-statistic* (von Mises statistic) for θ_h and $(X_k)_{k \in \mathbb{N}}$ is

$$V_h(X_1, \ldots, X_n) = n^{-d} \sum \{ h(X_{i_1}, \ldots, X_{i_d}) : 1 \le i_j \le n \text{ for all } j \}.$$

Example 2.6. For the classical parameters, the mean $M_1(F)$ and $P_1(F)$ are U-parameters with kernel $h(x) = x$; for all integral $\alpha \ge 1$, $P_\alpha(F)$ is a U-parameter with kernel

$$h(x_1, \ldots, x_\alpha) = x_1 \vee x_2 \vee \cdots \vee x_\alpha$$

(and is *not* a U-parameter for non-integral α; see Proposition 2.9 below); and the generalized Gini difference $g_\alpha(F)$ is a U-parameter with kernel $h(x_1, x_2) = |x_1 - x_2|^\alpha$.

The first SLLN for U-parameters is due to Hoeffding ([Hoe], see also [Se, p 190]), who proved the SLLN for iid sequences with any integrable kernel.

The main U-parameter result of this article is Theorem U below, which extends Hoeffding's result to three large classes of nonindependent processes.

Definition 2.7. A *product function* on $\mathbb{R}^d$ is a function of the form

$$f_1 \otimes \cdots \otimes f_d(x_1, \ldots, x_d) = f_1(x_1) \ldots f_d(x_d)$$

where $f_1, \ldots, f_d : \mathbb{R} \to \mathbb{R}$. For a distribution F on $\mathbb{R}$, the *product* $f_1 \otimes \cdots \otimes f_d$ *is F-integrable* if each f_i is measurable and $\int |f_i| dF < \infty$. A measurable function $h : \mathbb{R}^d \to \mathbb{R}$ is *bounded by F-integrable products* if $|h| \leq f_1 \otimes \cdots \otimes f_d$ for some F-integrable product $f_1 \otimes \cdots \otimes f_d$. Note that this class includes all bounded measurable functions, and that if $|h| \leq f_1 \otimes \cdots \otimes f_d$, then $|h| \leq f \otimes \cdots \otimes f$ where $f = f_1 \vee \cdots \vee f_d$.

The following proposition shows that under the condition of bounded by integrable products, the strong law limiting behavior of U- and V-statistics for ESP's is identical. This will be used in the proof of Theorem U below, as well as in several examples and intermediate results.

Proposition 2.8. *Let $(X_k)_{k \in \mathbb{N}}$ be an ergodic stationary process, and let $h : \mathbb{R}^d \to \mathbb{R}$ be bounded by integrable products. Then*

$$\lim_{n \to \infty} |U_h(X_1, \ldots, X_n) - V_h(X_1, \ldots, X_n)| = 0 \quad a.s.$$

Proof. Since the conclusion of the Marcinkiewicz SLLN holds for ESP's (cf. [A]), if $(Y_k)_{k \in \mathbb{N}}$ is an ESP with $E|Y_1|^{1/d} < \infty$, then $n^{-d} \sum_{k=1}^n Y_k \to 0$ a.s. Thus for an h of order 2 bounded by an integrable product $f_1 \otimes f_2$, letting $f = \max\{f_1, f_2\}$ and $Y_k = f^2(X)$,

$$\lim_{n \to \infty} |U_h(X_1, \ldots, X_n) - V_h(X_1, \ldots, X_n)| \leq \lim_{n \to \infty} n^{-2} \sum_{k=1}^n |h(X_k, X_k)|$$

$$\leq \lim_{n \to \infty} n^{-2} \sum_{k=1}^n f^2(X_k)$$

$$= \lim_{n \to \infty} n^{-2} \sum_{k=1}^n Y_k = 0 \quad \text{a.s.}$$

The general case $d > 2$ follows similarly. $\qquad\square$

Theorem U (SLLN for U-statistics). *Let $(X_k)_{k \in \mathbb{N}}$ be a stationary ergodic process with marginal F, and let $h : \mathbb{R}^d \to \mathbb{R}$ be measurable, bounded by an F-integrable product. If any of the following three conditions hold:*
  (i) *F is discrete;*
 (ii) *h is continuous at $F^{(d)}$-almost every point;*
(iii) *$(X_k)_{k \in \mathbb{N}}$ is weakly Bernoulli;*
then

$$(2) \qquad\qquad \lim_{n \to \infty} U_h(X_1, \ldots, X_n) = \theta_h(F) \qquad P\text{-}a.s.$$

There are however ESP's and bounded kernels for which the corresponding U-statistic SLLN does not hold, as will be seen in §4. The proofs of (i) and (ii) will be given in §4 and that of (iii) in §5. By conclusion (ii) it follows that the kernel $h(x, y) = |x - y|^\alpha$ for generalized Gini's mean difference parameter satisfies the U-parameter SLLN whenever $\int_{\mathbb{R}} |x|^\alpha dF(x) < \infty$ because $|x - y|^\alpha \leq (1 + |x|)^\alpha (1 + |y|)^\alpha$. For the case $\alpha = 1$, since $|x - y| = 2(x \vee y) - (x + y)$, it follows from Proposition 2.9 below that h is also an L-parameter.

The final proposition in this section demonstrates that the set $\mathcal{L} \cap \mathcal{U}$, although nonempty, is a rather small subset of $\mathcal{L} \cup \mathcal{U}$. It is particularly noteworthy that any U-parameter whose kernel h is not homogeneous of order 1 (e.g. $h(x_1, x_2) = (x_1 - x_2)^2$) is *not* an L-parameter, and on the other hand any continuous non-polynomial J on $[0, 1]$ generates an L-parameter which is not a U-parameter. By way of introduction, for a distribution F with finite mean and for a positive integer k, consider the well-known identity

$$(3) \qquad k \int_0^1 u^{k-1} F^{-1}(u) du = \int_{\mathbb{R}} x F^k(dx) = E(\widehat{X}_1 \vee \cdots \vee \widehat{X}_k),$$

where $\widehat{X}_1, \ldots, \widehat{X}_k$ are independent F-distributed r.v.'s.

The extension of (3) to polynomials by linearity shows that the L-parameter determined by the polynomial $J(u) = \sum_{k=1}^d c_k k u^{k-1}$ is equal (for all F with finite mean) to the U-parameter determined by the kernel

$$(4) \qquad h(x_1, \ldots, x_d) = c_1 x_1 + c_2(x_1 \vee x_2) + \cdots + c_d(x_1 \vee \cdots \vee x_d).$$

The following proposition shows that the set $\mathcal{L} \cap \mathcal{U}$ consists precisely of these parameters.

Proposition 2.9. *The following are equivalent:*
  (i) *θ is both an L-parameter and a U-parameter;*
 (ii) *θ is an L-parameter with a.c. representing measure whose density is a polynomial;*
(iii) *θ is a U-parameter with kernel which is a linear combination of partial maxima (e.g. of form (4) above).*

Proof. The equivalence of (ii) and (iii), hence also the implication (ii) or (iii) $\Rightarrow$ (i), follows from the discussion preceding the statement of the proposition. It thus remains only to prove that (i) implies (ii). For $\theta \in \mathcal{L} \cap \mathcal{U}$ there is, by definition, a Borel measure μ on $[0, 1]$ and a measurable function h on $\mathbb{R}^d$ (for some d) such that

$$(5) \qquad \theta(F) = \int_0^1 F^{-1} d\mu = \int_{\mathbb{R}^d} h \, dF^{(d)}$$

for all F for which either of these integrals is finite. To prove that in this case μ is a.c. and $J = J_\mu$ is a polynomial, specialize the identity (5) to the one-parameter family $\{F_p\}_{0 \leq p \leq 1}$ of Bernoulli distributions, i.e. $F_p^{-1}(u) = 1$ for $1 - p < u \leq 1$ and 0 elsewhere. It is then easy to see that, whatever the function h, the right hand side of (5) is a polynomial in p; hence also $\theta(F_p) = \mu((1 - p, 1])$ must be a polynomial in p. Hence, μ is a.c. and J is a polynomial. $\qquad \square$

3. THE L-PARAMETER SLLN FOR ERGODIC STATIONARY PROCESSES

The main purpose of this section is to prove Theorem L. Note that it is sufficient (by the Hahn-Jordan decomposition theorem) to establish the L-parameter SLLN (1) for μ a probability, and therefore we assume without loss of generality throughout that μ is a probability.

Lemma 3.1. *Suppose $(X_k)_{k \in \mathbb{N}}$ is an ergodic stationary process with marginal F. Then there is a countable set $\Gamma \subset [0,1]$ satisfying*

$$(6) \qquad \lim_{n \to \infty} F_n^{-1}(u) = F^{-1}(u) \quad \textit{a.s. for all } u \in [0,1] \backslash \Gamma.$$

Proof. It follows from the ergodic theorem that $F_n(x) \to F(x)$ a.s. for all $x \in \mathbb{R}$. Consequently $F_n \to F$ weakly a.s., and hence (e.g. [Bi, page 287]) there is a countable set Γ satisfying (6). □

Proof of Theorem L. To establish (i), note that P-almost surely,

$$F_n^{-1} \to F^{-1} \quad \mu - \text{a.e. on } [0,1]$$

by Lemma 3.1 since μ is atomless. Also

$$\|F_n^{-1}\|_{L^\infty([0,1])} \le \|F^{-1}\|_{L^\infty([0,1])} = \|X_1\|_{L^\infty(\Omega)} \text{ a.s.,}$$

so by Lebesgue's bounded convergence theorem,

$$\int_{[0,1]} F_n^{-1} d\mu \to \int_{[0,1]} F^{-1} d\mu \quad \text{a.s.,}$$

which proves (i).

Part (ii) of Theorem L will be established by an approximation argument using part (i). For $M > 0$, consider the continuous truncation function at M defined by

$$\tau_M(x) = \begin{cases} -M, & x < -M, \\ x, & |x| \le M, \\ M, & x > M. \end{cases}$$

Note that τ_M is odd, $\tau_M(x) \uparrow x$ as $M \to \infty$ for $x > 0$, and $|\tau_M(x)| = |x| \wedge M$. Also

$$x - \tau_M(x) = \text{sign}(x)(|x| - M)1_{[-M,M]^c}(x).$$

If G is the distribution function of $\tau_M(X)$, then clearly

$$G^{-1} = \tau_M \circ F^{-1}.$$

Since μ is continuous,

$$\int_{[0,1]} F_n^{-1} d\mu = \sum_{k=1}^{n} X_{n;k} \mu\left(\left[\frac{k-1}{n}, \frac{k}{n}\right]\right)$$

$$= \sum_{k=1}^{n} \tau_M(X_{n;k}) \mu\left(\left[\frac{k-1}{n}, \frac{k}{n}\right]\right) + \sum_{k=1}^{n} (X_{n;k} - \tau_M(X_{n;k})) \mu\left(\left[\frac{k-1}{n}, \frac{k}{n}\right]\right)$$

$$:= A_n + B_n.$$

Now,

$$A_n = \int_{[0,1]} \widehat{F}_n^{-1} d\mu$$

where $\widehat{F}_n$ is the empirical distribution of $(\tau_M(X_k))_{1 \leq k \leq n}$, and hence, by Theorem L(i),

$$A_n \to \int_{[0,1]} \tau_M \circ F^{-1} d\mu.$$

On the other hand,

$$|B_n| \leq \sum_{k=1}^{n} (|X_{n;k}| - M) 1_{[|X_{n;k}|>M]} \mu \left(\left[\frac{k-1}{n}, \frac{k}{n} \right] \right)$$

$$\leq \sum_{k=1}^{n} (|X_{n;k}| - M) 1_{[|X_{n;k}|>M]} \frac{\|J\|_\infty}{n}$$

$$:= \widehat{B}_n = \frac{\|J\|_\infty}{n} \sum_{k=1}^{n} (|X_k| - M) 1_{[|X_k|>M]} \to \|J\|_\infty E((|X| - M) 1_{[|X|>M]})$$

a.s. by the ergodic theorem.

By assumption of integrability,

$$E((|X| - M) 1_{[|X|>M]}) \to 0 \text{ as } M \to \infty, \text{ and } \int_{[0,1]} |F^{-1}| d\mu \leq \|J\|_\infty E(|X|) < \infty.$$

By Lebesgue's dominated convergence theorem

$$\int_{[0,1]} \tau_M \circ F^{-1} d\mu \to \int_{[0,1]} F^{-1} d\mu \text{ as } M \to \infty.$$

Accordingly, given $\epsilon > 0$, fix $M > 1$ such that

$$E((|X| - M) 1_{[|X|>M]}) < \frac{\epsilon}{\|J\|_\infty}, \text{ and } \left| \int_{[0,1]} \tau_M \circ F^{-1} d\mu - \int_{[0,1]} F^{-1} d\mu \right| < \epsilon$$

and obtain from the above that a.s.:

$$\left| \int_{[0,1]} F_n^{-1} d\mu - \int_{[0,1]} F^{-1} d\mu \right|$$

$$\leq \left| A_n - \int_{[0,1]} \tau_M \circ F^{-1} d\mu \right| + \widehat{B}_n + \left| \int_{[0,1]} \tau_M \circ F^{-1} d\mu - \int_{[0,1]} F^{-1} d\mu \right|$$

$$\underset{n \to \infty}{\to} E((|X| - M) 1_{[|X|>M]}) + \left| \int_{[0,1]} \tau_M \circ F^{-1} d\mu - \int_{[0,1]} F^{-1} d\mu \right|$$

$$< 2\epsilon$$

and so the L-statistic SLLN (1) follows.   □

The conclusion of this section is an example which shows that even the L-parameter *weak* law of large numbers may fail for L-parameters with a.c. representing measures with unbounded density, *even in the classical iid setting.* In particular, the example gives a distribution F of a random variable $X \geq 0$ with $EX < \infty$, an a.c. representing measure μ with $\int F^{-1} d\mu < \infty$, and a subsequence of positive integers $\{m_k\}$ satisfying $P\left(\int_{[0,1]} F_{m_k}^{-1} d\mu > k \right) \geq c > 0$ for all $k \in \mathbb{N}$.

Example 3.2. First, a simpler discrete version will be given. Let $n_0 = 2$, and for $k \in \mathbb{N}$ let $n_k = 2^{2^k}$, so $n_{k+1} = n_k^2$. Let X be a random variable with distribution $F(x) = 1 - n_{k+2}^{-1}$ for $x \in [n_k, n_{k+1})$, so $F^{-1}(1 - n_{k+2}^{-1}) = n_k$ and

$$EX = \sum_{k=1}^{\infty} n_k(n_{k+1}^{-1} - n_{k+2}^{-1}) \leq \sum_{k=1}^{\infty} n_k^{-1} < \infty.$$

Let μ be the purely atomic Borel measure on $[0,1]$ with $\mu(\{1 - n_{k+2}^{-1}\}) = (k+1)n_{k+1}^{-1} - (k+2)n_{k+2}^{-1}$, so $\mu([1 - n_{k+2}^{-1}, 1]) = (k+1)n_{k+1}^{-1}$, and

$$\int_{[0,1]} F^{-1}d\mu = \sum_{k=1}^{\infty} F^{-1}(1 - n_{k+2}^{-1})\mu(\{1 - n_{k+2}^{-1}\}) \leq \sum_{k=1}^{\infty} n_k(k+1)n_{k+1}^{-1}$$

$$= \sum_{k=1}^{\infty} (k+1)n_k^{-1} < \infty.$$

Note that $\int_{[0,1]} F_n^{-1}d\mu \geq X_{n;n}\mu\left(1 - \frac{1}{n}, 1\right]$, so for $m_k = n_{k+2} - 1$,

$$P\left(\int_{[0,1]} F_{m_k}^{-1}d\mu > k\right) \geq P\left(X_{m_k;m_k}\mu\left(1 - m_k^{-1}, 1\right] > k\right)$$

$$= P(X_{m_k;m_k}\mu(\{1 - n_{k+2}^{-1}\}) > k) = P(X_{m_k;m_k} > n_{k+1}k/(k-1))$$

$$= 1 - [1 - F(n_{k+1}k/(k-1))]^{m_k}$$

$$= 1 - (1 - n_{k+2}^{-1})^{m_k} \to 1 - e^{-1} \quad \text{as } k \to \infty.$$

To obtain an a.c. measure with this same property, simply replace the mass on $\{1 - n_k^{-1}\}$ with the same mass uniformly distributed on the interval $(1 - (n_k - 1)^{-1}, 1 - n_k^{-1})$ for each k. Likewise the discreteness of X is also not essential here, and a continuous analog can be found by convolving F with a $U(0,1)$ distribution, for example.

4. The U-Parameter SLLN for Ergodic Stationary Processes

The main purpose of this section is to prove Theorem U(i) and (ii), and give examples to indicate the significance of the kernel being bounded by an integrable product, and demonstrate the role played by continuity properties of the kernel. Let $(X_1, X_2, \dots)$ be an ESP with sample space $(\Omega, \mathcal{A}, P)$ and marginal distribution F, let $d \in \mathbb{N}$, and let h be a real-valued, measurable function on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} |h|dF^{(d)} < \infty$.

When $d = 1$, the U-parameter SLLN (2) is a consequence of the pointwise ergodic theorem. When $d \geq 2$, it is not, as the pointwise ergodic theorem establishes convergence a.e. on Ω^d with respect to the d-fold product measure $P \times \dots \times P$ rather than on Ω with respect to P (or on Ω^d with respect to the *diagonal* measure). The situation in (2) (when $d \geq 2$) is complicated by the fact that the convergence is demanded to be a.e. with respect to a measure which (when F is atomless) is singular with respect to the measure of integration in the limit. This is seen in the following example, which shows that the U-parameter SLLN (2) may even fail for bounded kernels.

Example 4.1. Consider the Lebesgue-measure-preserving and ergodic transformation $T : [0, 1) \to [0, 1)$ defined by $T\omega = 2\omega (\mathrm{mod}\ 1)$, and let $X_i = T^i\omega$. Denote by G the union of the graphs of T and all its iterates (G is sometimes called the T-orbit of the diagonal), and let $h = 1_G$. Since the pairs (X_i, X_j) all lie in G, $U_h(X_1, \ldots, X_n) = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j) = 1$ for all n, but $\int \int h(x, y) dx dy = 0$ because G clearly has (planar) Lebesgue-measure zero.

Let

$$H_F = H_F^{(d)} = \{h \in L^1(F^{(d)}) : (2) \text{ holds for all ESP with marginal distribution } F\}.$$

Lemma 4.2. *If $f_i \in L^1(F)$ $(1 \le i \le d)$ and $h = f_1 \otimes \cdots \otimes f_d$, then $h \in H_F$.*

Proof. We have

$$V_h(X_1, \ldots, X_n) = \prod_{i=1}^{d} \left(\frac{1}{n} \sum_{k=1}^{n} f_i(X_k) \right) \to \prod_{i=1}^{d} \left(\int_{\mathbb{R}} f_i dF \right) = \int_{\mathbb{R}^d} h dF^{(d)}$$

P-a.e. by the pointwise ergodic theorem and Fubini's theorem. Then (2) follows from Proposition 2.8. □

Also, H_F is clearly a linear space, and in particular, linear combinations of F-integrable product functions are in H_F.

Lemma 4.3 (Sandwich lemma). *Suppose that $h \in L^1(F^{(d)})$, and that for all $\epsilon > 0$, there are $u, v \in H_F$ such that $|u - h| \le v$ $F^{(d)}$-a.e., and $\theta_v(F) < \epsilon$. Then $h \in H_F$.*

Proof. Given $\epsilon > 0$, fix $u, v \in H_F$ satisfying the hypothesis. Then

$$|U_h(X_1, \ldots, X_n) - \theta_h(F)| \le |U_h(X_1, \ldots, X_n) - U_u(X_1, \ldots, X_n)|$$
$$+ |(U_u(X_1, \ldots, X_n) - \theta_u(F)| + |\theta_u(F) - \theta_h(F)|$$
$$\le U_{|h-u|}(X_1, \ldots, X_n) + |U_u(X_1, \ldots, X_n) - \theta_u(F)| + \theta_v(F).$$

The first term in the right hand side is $F^{(d)}$-a.e. bounded by $U_v(X_1, \ldots, X_n)$, which converges to $\theta_v(F)$ since $v \in H_F$, and the second term is $o(1)$ since $u \in H_F$. Since ϵ is arbitrary, this implies $U_h(X_1, \ldots, X_n) \to \theta_h(F)$ a.s. □

Proposition 4.4. *If $h : \mathbb{R}^d \to \mathbb{R}$ is bounded with compact support and continuous at $F^{(d)}$-a.e. point, then $h \in H_F$.*

Proof. Since h is Riemann-Stieltjes integrable with respect to $F^{(d)}$, for any $\epsilon > 0$, there are d-dimensional step functions u and v (i.e. linear combinations of products of indicators of intervals) satisfying the approximation condition of Lemma 4.3. □

Proof of Theorem U(i). Assume $d = 2$, the general argument being analogous. Let $|h| \le f \otimes f$ with f F-integrable, and let

$$\Gamma = \{x \in \mathbb{R} : F(\{x\}) > 0\} = \bigcup_{n=1}^{\infty} \Gamma_n$$

where $\#\Gamma_n < \infty$, and $\Gamma_n \subseteq \Gamma_{n+1}$.

Without loss of generality $h : \Gamma \times \Gamma \to \mathbb{R}$ and

$$h = \sum_{a,b \in \Gamma} h(a,b) 1_{\{a\} \times \{b\}}.$$

Set

$$u_N = \sum_{a,b \in \Gamma_N} h(a,b) 1_{\{a\} \times \{b\}},$$

and

$$v_N = (f 1_{\Gamma_N^c} \otimes f 1_{\Gamma_N}) + (f 1_{\Gamma_N} \otimes f 1_{\Gamma_N^c}) + (f 1_{\Gamma_N^c} \otimes f 1_{\Gamma_N^c}).$$

By Lemma 4.2 $u_N, v_N \in H_F$, $|h - u_N| \leq v_N$, and

$$\int_{\Gamma^2} v_N dF^{(2)} = \left(\int_\Gamma f dF \right)^2 - \left(\int_{\Gamma_N} f dF \right)^2 \to 0$$

as $N \to \infty$, and (2) follows from Lemma 4.3. $\qquad\square$

Proof of Theorem U(ii). Again suppose $d = 2$, the general argument being analogous. Suppose $|h| \leq f \otimes f$, where f is F-integrable. Fix $M > 0$ so that $P(|X| = M) = 0$ and $F^{(2)}([|h| = M]) = 0$, (i.e. M is a continuity point of the distributions of $|X|$ and of $|h(X, \widehat{X})|$, where $X, \widehat{X}$ are iid with distribution F).
   Define u_M by

$$u_M(x, y) = h(x, y) 1_{[|h| \leq M]}(x, y) 1_{[-M,M] \times [-M,M]}(x, y),$$

and

$$v_M = (f \otimes f) \left(1_{[-M,M]^c \times \mathbb{R}} + 1_{\mathbb{R} \times [-M,M]^c} + 1_{[f \geq \sqrt{M}] \times \mathbb{R}} + 1_{\mathbb{R} \times [f \geq \sqrt{M}]} \right).$$

Clearly, $v_M \in H_F^{(2)}$ as a sum of F-integrable products. Since h is $F^{(2)}$-a.e. continuous and $F^{(2)}([|h| = M]) = P(|X| = M) = 0$, u_M is bounded, of compact support and $F^{(2)}$-a.e. continuous. Therefore $u_M \in H_F^{(2)}$ by Proposition 4.4. To see that $|h - u_M| \leq v_M$, note that

$$|h - u_M| \leq |h| \left(1_{[|h| > M]} + 1_{([-M,M] \times [-M,M])^c} \right)$$

$$\leq (f \otimes f) \left(1_{[f \otimes f \geq M]} + 1_{\mathbb{R} \times [-M,M]^c} + 1_{([-M,M]^c \times \mathbb{R})} \right) \leq v_M \quad F^{(2)}\text{-a.e.},$$

since $a, b \geq 0$, $ab \geq M \Rightarrow a \vee b \geq \sqrt{M}$. Finally, by Fubini,

$$\int_{\mathbb{R}^2} v_M dF^{(2)} = 2E(f(X))E\left(f(X)(1_{[f(X) \geq \sqrt{M}]} + 1_{[|X| > M]}) \right) \to 0$$

as $M \to \infty$, and the conclusion follows from Lemma 4.3. $\qquad\square$

   The next example shows that one cannot omit entirely the condition of boundedness by integrable products in Theorem U(ii).

Example 4.5. Let $Z_0, Z_1, \ldots$ be independent, $\{0,1\}$-valued, symmetrically distributed random variables. Let $\{Y_n : n \geq 0\}$ be iid rv's uniform on $[0,1]$, and independent of $\{Z_n : n \geq 0\}$. Define $\{X_n : n \geq 0\}$ by $X_0 = Y_0$, and

$$X_{n+1} = \begin{cases} X_n & \text{if } Z_n = 1, \\ Y_{n+1} & \text{if } Z_n = 0. \end{cases}$$

Since $\{X_n\}$ is stationary, and Lebesgue measure is the unique invariant measure (in fact Lebesgue measure attracts every initial distribution), $\{X_n : n \geq 0\}$ is an ESP and X_n is uniform on $[0,1]$. Now choose $h : [0,1] \times [0,1] \to \mathbb{R}_+$, continuous on $[0,1] \times [0,1] \setminus \{(0,0)\}$, and such that $\int_{[0,1] \times [0,1]} h d(\lambda \times \lambda) = 1$ and $h(x,x) = \frac{1}{x^3}$. It will now be shown that the U-SLLN fails for h and $\{X_n\}$. Define $i_0 < i_1 < \ldots$ inductively by $i_0 = 0$ and $i_{k+1} = \min\{i > i_k : Z_i = 0 \text{ and } Z_{i+1} = 1\}$, so $\{X_{i_n}\}$ are conditionally iid $U(0,1)$ given $\{Z_n : n \geq 0\}$, and $X_{i_n+1} = X_{i_n}$. Since $\lim_{n \to \infty} i_n/n = 4$ a.s. (by the ergodic theorem),

$$\limsup_{n \to \infty} U_h(X_1, \ldots, X_n) \geq \limsup_{n \to \infty} \frac{1}{n^2} h(X_{i_n}, X_{i_n+1}) = \limsup_{n \to \infty} \frac{1}{n^2} \frac{1}{X_{i_n}^3} = \infty \text{ a.s.,}$$

by the Borel-Cantelli Lemma, the conditional independence of $\{X_{i_n}\}$ and the fact that $P(X_{i_n} \leq \frac{1}{n} \mid Z_n : n \geq 0) = \frac{1}{n}$.

It is not clear whether the kernel of Example 4.5 violates the U-statistic weak law. The kernel in the next example indeed does this.

Example 4.6. Let $\{Y_k\}$ be iid $U[0,1]$, and let $g : (0,1] \to \mathbb{R}_+$ be a non-negative, decreasing continuous function such that $g(Y_0)$ has a positive stable law of index $\frac{1}{4}$. If $S_n = \sum_{k=1}^n g(Y_k)$; then $E(e^{-tS_n}) = e^{-cnt^{1/4}}$ where $c > 0$. Fix $M > 0$; then for all $t > 0$, by Markov's inequality,

$$P([S_n < Mn^2]) = P([e^{-tS_n} \geq e^{-Mtn^2}]) \leq e^{Mtn^2 - cnt^{1/4}},$$

and choosing $t > 0$ which minimizes this yields

$$P([S_n \leq Mn^2]) \leq e^{-c'n^{2/3}}$$

where $c' = c'(M) > 0$.

It follows from Borel-Cantelli that

$$\frac{1}{n^2} \sum_{k=1}^n g(Y_k) \to \infty \quad \text{a.e.}$$

Now choose $h : [0,1] \times [0,1] \to \mathbb{R}_+$, continuous on $[0,1] \times [0,1] \setminus \{(0,0)\}$, and such that

$$\int_{[0,1] \times [0,1]} h d(\lambda \times \lambda) = 1$$

and

$$h(x,x) = g(x).$$

It will now be shown that the U-statistic WLLN fails for h and the $\{X_n\}$ as in Example 4.5. Let $i_1 < i_2 < \dots$ be as in that example; setting

$$t_n = \max\{k : i_k \leq n\} \sim \frac{n}{4} \text{ a.s.,}$$

it follows that

$$U_h(X_1, \dots, X_n) \geq \frac{1}{n(n-1)} \sum_{1 \leq j \leq t_n} h(X_{i_j}, X_{i_j+1})$$

$$= \frac{1}{n(n-1)} \sum_{1 \leq j \leq t_n} g(X_{i_j}) \to \infty \quad \text{a.s.}$$

To obtain a discrete version of this example, simply replace g by a function $f \geq g$ defined by $f(y) = n$ on the set $\{y : n - 1 < g(y) \leq n\}$, $n = 1, 2, \dots$.

The conclusion of this section gives a sufficient condition (Proposition 4.9) for the indicator function of a countable union of product sets to be in $H_F^{(d)}$. The method works in the absence of continuity and uses approximation with error estimated by the maximal function of the U-statistic. Although all indicator functions of finite unions of product sets are in H_F (Lemma 4.2), this is not true for countable unions, as can be seen by looking at such a union of less than full measure which contains the T-orbit G in Example 4.1.

For $h : \mathbb{R}^d \to \mathbb{R}$ measurable and an ESP $(X_k)_{k \in \mathbb{N}}$, let

$$M(h) = M(h)(X_1, X_2, \dots) = \sup_{n \geq 1} |U_h(X_1, \dots, X_n)|.$$

Lemma 4.7. *Suppose $h \in L^1(F^{(d)})$, and that for all $\epsilon > 0$ there exists $u(\epsilon) \in H_F^{(d)}$ such that $E[M(|h - u(\epsilon)|)] < \epsilon$ for all ESP with marginal F, and that $\int_{\mathbb{R}^d} |h - u(\epsilon)| dF^{(d)} < \epsilon$. Then $h \in H_F^{(d)}$.*

Proof. For $\epsilon > 0$, let $u(\epsilon) \in H_F^{(d)}$ satisfy the hypotheses. Then

$$U_h(X_1, \dots, X_n) = U_{u(\epsilon)}(X_1, \dots, X_n) + U_{h-u(\epsilon)}(X_1, \dots, X_n).$$

Since $u(\epsilon) \in H_F^{(d)}$, $U_{u(\epsilon)}(X_1, \dots, X_n) \to \int_{\mathbb{R}^d} u(\epsilon) dF^{(d)}$ a.s. Also,

$$|U_{h-u(\epsilon)}(X_1, \dots, X_n)| \leq M(|h - u(\epsilon)|),$$

so for all $\epsilon > 0$,

$$G(\epsilon) := \limsup_{n \to \infty} |U_h(X_1, \dots, X_n) - \int_{\mathbb{R}^d} h \, dF^{(d)}|$$

$$\leq \int_{\mathbb{R}^d} |h - u(\epsilon)| dF^{(d)} + M(|h - u(\epsilon)|) \leq \epsilon + M(|h - u(\epsilon)|),$$

since $\int_{\mathbb{R}^d} |h - u(\epsilon)| dF^{(d)} \leq EM(|h - u(\epsilon)|) < \epsilon$. Thus $P([G(\epsilon) \geq \epsilon + \sqrt{\epsilon}]) \leq P([M(|h - u(\epsilon)|) \geq \sqrt{\epsilon}]) \leq \sqrt{\epsilon}$, so $G(n^{-4}) \to 0$ a.s. by Borel-Cantelli. $\square$

Given $a_1, \ldots, a_d \in [0, 1]$, set

$$m_d(a_1, \ldots, a_d) = \min_{(x_1, \ldots, x_d) \in [0,1]^d, \sum_{i=1}^d x_i = 1} \prod_{i=1}^d \frac{a_i^{x_i}}{1 - x_i}.$$

It is not hard to show that

$$m_2(a, b) = \left(2 + \sqrt{A^2 + 4}\right) e^{\frac{A + 2 - \sqrt{A^2 + 4}}{2}} (a \wedge b)$$

where

$$A = \log\left(\frac{a \vee b}{a \wedge b}\right).$$

Note that

(7) $\qquad m_2(a, a) = 4a$ and $m_2(a, b) \sim (a \wedge b)A$ as $A \to \infty$.

Also, there are constants $\alpha_d > 0$ $(d \geq 3)$ such that

(8) $\quad m_2(a_{d;1}, a_{d;2}) \geq m_d(a_1, \ldots, a_d) \geq \alpha_d m_2(a_{d;1}, a_{d;2}) \; \forall \, a_1, \ldots, a_n \in [0, 1]$

where $a_{d;1} \leq a_{d;2} \leq \cdots \leq a_{d;d}$ are the order statistics of the constants $a_1, \ldots, a_d$. The right hand inequality is not used in the sequel and is included for the interested reader.

Lemma 4.8. *If $A_1, \ldots, A_d \in \mathcal{B}(\mathbb{R})$ and $(X_k)_{k \in \mathbb{N}}$ is an ESP with marginal F, then*

$$E\left(M_d(1_{A_1 \times \cdots \times A_d})\right) \leq m_d(F(A_1), \ldots, F(A_d)).$$

Proof. Note first that $M_d(1_{A_1 \times \cdots \times A_d}) = M_d(1_{A_1} \otimes \cdots \otimes 1_{A_d}) \leq \prod_{i=1}^d M_1(1_{A_i})$, so for all $(x_1, \ldots, x_d) \in [0, 1]^d$ with $\sum_{i=1}^d x_i = 1$,

$$E\left(M_d(1_{A_1 \times \cdots \times A_d})\right) \leq E\left(\prod_{i=1}^d M_1(1_{A_i})\right) \leq \prod_{i=1}^d \|M_1(1_{A_i})\|_{L^{1/x_i}(P)}$$

$$\leq \prod_{i=1}^d \frac{1}{1 - x_i} \|1_{A_i}\|_{L^{1/x_i}(P)} = \prod_{i=1}^d \frac{F(A_i)^{x_i}}{1 - x_i},$$

where the second inequality follows by Hölder's inequality, and the third inequality by the maximal inequality (cf. [Ga], Theorem 2.2.3, p. 25). Minimizing this over $x_1, \ldots, x_d$ establishes the desired inequality. $\qquad\square$

Proposition 4.9. *Suppose that F is a probability distribution on $\mathbb{R}$ and that $A = \bigcup_{n=1}^\infty A_1^{(n)} \times \cdots \times A_d^{(n)}$, where $A_i^{(n)} \in \mathcal{B}(\mathbb{R})$.*
If $\sum_{n=1}^\infty \prod_{k=1}^d F(A_k^{(n)}) < \infty$, and $\sum_{n=1}^\infty m_d(F(A_1^{(n)}), \ldots, F(A_d^{(n)})) < \infty$, then $1_A \in H_F^{(d)}$.

Proof. Let $\epsilon > 0$. By the assumptions, there exists $N = N(\epsilon) \geq 1$ such that

$$\sum_{n=N}^\infty \prod_{k=1}^d F(A_k^{(n)}) < \epsilon, \text{ and } \sum_{n=N}^\infty m_d(F(A_1^{(n)}), \ldots, F(A_d^{(n)})) < \epsilon.$$

Set $u = u(\epsilon) = 1_B$, where

$$B = B(\epsilon) = \bigcup_{n=1}^{N} A_1^{(n)} \times \cdots \times A_d^{(n)}.$$

Since B can also be written as a disjoint union of product sets, it follows by Lemma 4.2 that $u \in H_F^{(d)}$ as a sum of product functions.

Since $0 \le 1_A - u \le \sum_{n=N}^{\infty} 1_{A_1^{(n)}} \otimes \cdots \otimes 1_{A_d^{(n)}}$, it follows that

$$\int_{\mathbb{R}^d} |1_A - u| dF^{(d)} \le \sum_{n=N}^{\infty} \prod_{k=1}^{d} F(A_k^{(n)}) < \epsilon;$$

and

$$E[M(|1_A - u|)] \le \sum_{n=N}^{\infty} E[M(1_{A_1^{(n)}} \otimes \cdots \otimes 1_{A_d^{(n)}})]$$

$$\le \sum_{n=N}^{\infty} \mu_d(F(A_1^{(n)}), \ldots, F(A_d^{(n)})) < \epsilon.$$

Thus the conditions of Lemma 4.7 are satisfied, and so $1_A \in H_F^{(d)}$. $\qquad\square$

Example 4.10. Let F be uniform on $[0,1]$, let $\{q_n : n \ge 1\}$ denote the set of points in $[0,1]^2$ with rational coordinates, and let $A = \bigcup_{n=1}^{\infty} S(q_n, \frac{1}{4^n})$ where

$$S((r,s),\delta) := \{(x,y) \in [0,1]^2 : |x - r|, |y - s| < \delta\}.$$

The set A is dense and open in $[0,1]^2$, but not of full measure, so 1_A is not continuous at $F^{(2)}$-a.e. point, and $1_A \in H_F^{(2)}$ cannot be deduced from Theorem U(ii).

To see that in fact 1_A is in $H_F^{(2)}$, note that $S(q_n, 1/4^n) = I_n \times J_n$ where $F(I_n) = F(J_n) = 2/4^n$, so since $m_2(a,a) = 4a$,

$$\sum_{n=1}^{\infty} m_2(F(I_n), F(J_n)) = \sum_{n=1}^{\infty} \frac{8}{4^n} < \infty, \quad \text{and} \quad \sum_{n=1}^{\infty} F(I_n)F(J_n) < \infty.$$

By Proposition 4.9, $1_A \in H_F^{(2)}$.

Higher order examples can be constructed using the following result.

Corollary 4.11. *Let F be a probability distribution on $\mathbb{R}$, let $d \ge 1$, and let $A = \bigcup_{n=1}^{\infty} A_1^{(n)} \times \cdots \times A_d^{(n)}$, where $A_i^{(n)} \in \mathcal{B}(\mathbb{R})$. If $\epsilon_n := \min_{1 \le k \le d} F(A_k^{(n)})$ satisfies $\limsup_{n \to \infty} \epsilon_n < 1$ and $\sum_{n=1}^{\infty} \epsilon_n \log(1/\epsilon_n) < \infty$, then $1_A \in H_F^{(d)}$.*

Proof. The assumptions imply

$$\sum_{n=1}^{\infty} \prod_{k=1}^{d} F(A_k^{(n)}) \le \sum_{n=1}^{\infty} \epsilon_n < \infty;$$

and

$$\sum_{n=1}^{\infty} m_d(F(A_1^{(n)}), \ldots, F(A_d^{(n)})) \le \sum_{n=1}^{\infty} m_2(\epsilon_n, 1) < \infty$$

by (7) and (8).

By Proposition 4.9, $1_A \in H_F^{(d)}$. $\qquad\square$

5. Weakly Bernoulli Sequences

Example 4.1 shows that the U-statistic SLLN may fail for bounded measurable kernels whose discontinuity set is large. On the other hand, Hoeffding [Hoe] proved that the U-statistic SLLN holds for iid random variables and *any* bounded measurable kernel. The main purpose of this section is to extend Hoeffding's result to weakly Bernoulli ESP, proving Theorem U(iii). Actually, a somewhat stronger result will be proved.

Definition 5.1. A process $(X_k)_{k\in\mathbb{N}}$ is called *F-regular* if for every $\epsilon > 0$ there is an integer $m \geq 1$ such that for every $N \geq 1$, there exists (enlarging the probability space if necessary) an iid sequence of N-dimensional random vectors $\xi_1', \xi_2', \ldots$ whose coordinate marginal distributions are F and which satisfy

$$\lim_{K\to\infty} K^{-1}\#\{k \leq K : \xi_k \neq \xi_k'\} \leq \epsilon \quad \text{a.s.,}$$

where $\xi_k = (X_{(k-1)(N+m)+1}, \ldots, X_{kN+(k-1)m})$, $k = 1, 2, \ldots$.

F-regularity of a sequence says that it is "almost iid for SLLN purposes," in the sense that periodic blocks of arbitrarily long sequences differ from those of an iid sequence only over a set of indices of arbitrarily small density. The next theorem says that the U-statistic SLLN holds for F-regular sequences and kernels bounded by integrable products.

Theorem 5.2. *Let $(X_k)_{k\in\mathbb{N}}$ be a F-regular process and let $h : \mathbb{R}^d \to \mathbb{R}$ be measurable and bounded by an F-integrable product. Then*

$$U_h(X_1, \ldots, X_n) \to \theta_h(F) \quad a.s.$$

Proof. In the interest of simplicity, the case $d = 2$ is presented; the general argument is similar. Using the truncation argument in the proof of Theorem U(ii), reduce to the case where h is bounded, say $|h| \leq 1$. Let $\epsilon > 0$, fix $m = m(\epsilon)$ as in Definition 5.1 and fix an integer N so $\frac{m}{m+N} < \epsilon$. The idea is to split the integers up into consecutive blocks of length N (the big blocks) and length m (the small blocks), respectively and then essentially discard the small blocks and approximate the sequence of large blocks by an iid sequence. Let $n_k = (k-1)(m+N)$ and define the block vector

$$\xi_k = (X_{n_k+1}, \ldots, X_{n_k+N}).$$

Define the kernel $\widehat{h} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ by

$$\widehat{h}(\xi, \eta) = \frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} h(X_i, Y_j)$$

where $\xi = (X_1, \ldots, X_N)$ and $\eta = (Y_1, \ldots, Y_N)$. Note that for independent ξ and η (each with F-distributed individual coordinates but otherwise with any joint distribution on $\mathbb{R}^N$), $E\widehat{h}(\xi, \eta) = \int_{\mathbb{R}^2} h dF^{(2)}$. If n is the sample size, the index of the last block fully contained in $\{1, 2, \ldots, n\}$ is given by $p := [\frac{n}{N+m}]$. Then

$$\left| U_h(X_1, \ldots, X_n) - \frac{1}{n(n-1)} \sum_{1 \leq k \neq \ell \leq p} \sum_{i=n_k+1}^{n_k+N} \sum_{j=n_\ell+1}^{n_\ell+N} h(X_i, X_j) \right| \leq \frac{2mp}{n} + O\left(\frac{1}{p}\right),$$

so

$$\left| U_h(X_1, \ldots, X_n) - \frac{1}{n(n-1)} \sum_{1 \leq k \neq \ell \leq p} N(N-1)\widehat{h}(\xi_k, \xi_\ell) \right| \leq 3\epsilon \quad \text{for } p \text{ large.}$$

Let K denote the set of k for which $\xi_k = \xi'_k$. Then $|K^c \cap [1, p]| < \epsilon p$ for all p large, so, a.s.,

$$\left| \sum_{1 \leq k \neq \ell \leq p} \widehat{h}(\xi_k, \xi_\ell) - \sum_{1 \leq k \neq \ell \leq p} \widehat{h}(\xi'_k, \xi'_\ell) \right| \leq \sum_{1 \leq k \neq \ell \leq p, k \notin K \text{ or } \ell \notin K} |\widehat{h}(\xi_k, \xi_\ell) - \widehat{h}(\xi'_k, \xi'_\ell)|$$

$$\leq 4p|K^c \cap [1, p]| < 4p^2\epsilon \quad \text{for all } p \text{ large.}$$

By Hoeffding's Theorem ([Hoe], the U-statistic SLLN for iid rv's),

$$\lim_{p \to \infty} \frac{1}{p(p-1)} \sum_{1 \leq k \neq \ell \leq p} \widehat{h}(\xi'_k, \xi'_\ell) = E(\widehat{h}(\xi'_1, \xi'_2)) = \int_{\mathbb{R}^2} h \, dF^{(2)} \quad \text{a.s.}$$

These estimations imply that

$$\limsup_{n \to \infty} |U_h(X_1, \ldots, X_n) - \int_{\mathbb{R}^2} h \, dF^{(2)}| < \epsilon + 4\epsilon^3. \qquad \square$$

The next basic theorem gives the link between F-regularity and weak Bernoulli; as no reference is known to the authors, the proof is given for completeness. Together with Theorem 5.2, this will complete the proof of Theorem U(iii). Note that the converse of Theorem 5.3 is not true, since F-regularity does not imply stationarity (e.g., the deterministic sequence $X_k = 0$ if $k \neq 2n$ and $= 1$ if $k = 2n$ is F-regular with $F = \delta_0$, but is not stationary). Stationarity was not needed in Theorem 5.2, but is crucial in Theorem 5.3.

Recall that the stationary sequence $(X_k)_{k \in \mathbb{N}}$ is called *weakly Bernoulli* (WB) (also known as absolutely regular) if $d(m; k) \to 0$ uniformly in k as $m \to \infty$, where $d(m; k)$ is the supremum of $\sum_{i=1}^{n} |P(A_i \cap B_i) - P(A_i)P(B_i)|$ over all families of disjoint sets $A_i \cap B_i$, $i = 1, 2, \ldots, n$, where $A_i \in \sigma(X_1, \ldots, X_k)$ and $B_i \in \sigma(X_{k+m}, \ldots)$.

Theorem 5.3. *If $(X_k)_{k \in \mathbb{N}}$ is weakly Bernoulli with marginal F, then it is F-regular.*

The following coupling lemma of Berbee is one of the key tools in the proof. Here

$$\perp (X, Y) = \frac{1}{2} \|P_{(X,Y)} - P_X \times P_Y\|$$

is the *dependence* between random vectors X and Y, where $\| \cdot \|$ denotes the variational norm on measures, $P_X, P_Y, P_{(X,Y)}$ are the distributions of X, Y and (X, Y) respectively, and $P_X \times P_Y$ is the product measure. Note that $\perp (X, Y) = 0$ iff X and Y are independent.

Lemma 5.4 ([Ber, Corollary 4.2.5]). *Suppose that X, Y are random variables defined on a probability space (Ω, P). Then there is a random variable Y' defined on $(\Omega \times [0,1], P')$, where $P' = P \times \lambda$, such that:*

(i) *Y and Y' have the same distribution;*

(ii) *$\perp (\widehat{X}, Y') = 0$;*

(iii) *$P'(\widehat{Y} \neq Y') = \perp (X, Y)$;*

(iv) *$P_{Z|(X,Y,W)} = P'_{\widehat{Z}|(\widehat{X}, \widehat{Y}, \widehat{W}, Y')}$ for all rv's Z and W on Ω,*

where for rv's Z and W on Ω, $\widehat{Z}$ is defined on $\Omega \times [0,1]$ by $\widehat{Z}(\omega, t) = Z(\omega)$, and $P_{Z|W}$ denotes the P-conditional distribution of Z given W.

Proof of Theorem 5.3. Choose $m \geq 1$ so that $d(m) := \sup_k d(m; k) < \epsilon$ and for fixed N define $\xi_k = (X_{(k-1)(N+m)+1}, \ldots, X_{kN+(k-1)m})$, and set $\xi_1' = \xi_1$. Without loss of generality, take the underlying measure space to be $(\mathbb{R}^N \times \mathbb{R}^N)^{\mathbb{N}}$, which is a complete separable metric space. In Lemma 5.4 take $X = (\xi_1, \xi_1')$, $Y = \xi_2$ and denote the resulting Y' by ξ_2'. Clearly $P(\xi_2 \neq \xi_2') = \perp (\xi_2, \xi_1)$. Note that for all $k \geq 3$,

$$P_{\xi_k|(\xi_1, \ldots, \xi_{k-1}, \xi_1', \xi_2')} = P_{\xi_k|(\xi_1, \ldots, \xi_{k-1})},$$

and thus by a straightforward calculation (cf., [Ber, Prop. 4.1.1])

$$\perp (\xi_k, \{\xi_1, \ldots, \xi_{k-1}, \xi_1', \xi_2'\}) = \perp (\xi_k, \{\xi_1, \ldots, \xi_{k-1}\}).$$

Apply Lemma 5.4 again with $X = (\xi_1, \xi_2, \xi_1', \xi_2')$ and $Y = \xi_3$ to find Y', now denoted by ξ_3', so that

$$\perp (\xi_3', \{\xi_1, \xi_2, \xi_1', \xi_2'\}) = 0$$

and

$$P(\xi_3' \neq \xi_3) = \perp (\xi_3, \{\xi_1, \xi_2\}).$$

This procedure when iterated yields a measure μ on $\Omega \times \Omega$ with the following properties:

$$(9) \qquad \mu \circ \pi_1^{-1} = \mu_1 \text{ has the distribution of the original } \{\xi_k\} \text{ sequence;}$$

$$(10) \qquad \mu \circ \pi_2^{-1} = \mu_2 \text{ has iid coordinates with marginal that of } \xi_1;$$

$$(11) \qquad \mu\{(\omega_1, \omega_2) : \omega_1(k) \neq \omega_2(k)\} \leq d(m) \text{ for all } k,$$

where π_i is the projection onto the i-th coordinate, and $\omega_i(k)$ is the k-th coordinate of $\omega_i \in (\mathbb{R}^N)^{\mathbb{N}}$, i.e., an element of $\mathbb{R}^N$.

Claim 1. The collection of μ's satisfying (9)–(11) is convex and weakly closed (against bounded continuous functions).

The convexity is obvious, while for weak closure note that for fixed k the set in (11), call it S_k, is open. If f is a continuous function between 0 and 1 with support in S_k, and $\mu^{\alpha} \to \mu$ weakly as $\alpha \to \infty$, with μ^{α} satisfying (9)–(11), then

since $0 \le f \le 1_{S_k}$,

$$d(m) \ge \int 1_{S_k} d\mu^\alpha \ge \int f d\mu^\alpha \to \int f d\mu.$$

Thus for all such f, $\int f d\mu \le d(m)$, and since

$$\int 1_{S_k} d\mu = \sup \left\{ \int f d\mu : 0 \le f \le 1_{S_k} \right\},$$

it follows that $\int 1_{S_k} d\mu \le d(m)$, which establishes Claim 1.

Let σ denote the shift on $\Omega \times \Omega$ with σ_1, σ_2 the shift on the first and second coordinates. Note that $\mu_1 \circ \sigma_1^{-1} = \mu_1$, $\mu_2 \circ \sigma_2^{-1} = \mu_2$. Form the sequence

$$(12) \qquad\qquad \frac{1}{L} \sum_{\ell=1}^{L} \sigma^\ell \cdot \mu = \mu^L.$$

Note that μ^L continues to satisfy (9)–(11). Take a limit point $\widehat{\mu}$ which exists by tightness (if μ_1 and μ_2 are two fixed regular probability measures on Ω, then the family of all μ on $\Omega \times \Omega$ which project onto μ_1, μ_2, respectively, is tight). Since in *variation* $\|\sigma \cdot \mu^L - \mu^L\| \le 2/L$, it is clear that $\sigma \cdot \widehat{\mu} = \widehat{\mu}$. That is, $\widehat{\mu}$ is a *stationary* measure under σ, satisfying (9)–(11), so (cf. [vN]) $\widehat{\mu}$ can be decomposed as

$$\widehat{\mu} = \int_0^1 \widehat{\mu}_t d\nu(t),$$

where ν is a Borel probability measure on $[0,1]$, and $\widehat{\mu}_t$ are stationary *ergodic* measures on $\Omega \times \Omega$. Since both μ_1 and μ_2 were ergodic under σ_1, σ_2 respectively, it follows that for ν-a.e. t, $\pi_i(\widehat{\mu}_t) = \mu_i$ for $i = 1, 2$, since $\pi_i \cdot \widehat{\mu} = \int_0^1 \pi_i \cdot \widehat{\mu}_t d\nu(t)$.

Finally, since $d(m) \ge \widehat{\mu}(S_1) = \int_0^1 \widehat{\mu}_t(S_1) d\nu(t)$, there must be a set of t values of positive measure where $\widehat{\mu}_t(S_1) \le d(m)$. Choose *any* one, call it t_0, and observe that $\widehat{\mu}_{t_0}$ is an ergodic stationary measure satisfying (9)–(11). Note that by *stationarity* $\widehat{\mu}_t(S_k) = \widehat{\mu}_t(S_1)$ for all $k \ge 1$. Now the ergodic theorem applied to 1_{S_1} yields

$$\lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} 1_{S_1}(\sigma^k(\omega_1, \omega_2)) = \int 1_{S_1}(\omega_1, \omega_2) d\widehat{\mu}_{t_0}(\omega_1, \omega_2)$$

$$= \widehat{\mu}_{t_0}(S_1) \le d(m) \quad \widehat{\mu}_{t_0}\text{-a.e.}$$

and $\sum_{k=1}^{K} 1_{S_1}(\sigma^k(\omega_1, \omega_2)) = \#\{h \le K : \xi_k \ne \xi_k'\}$, where the (ξ_k, ξ_k') are now the desired rv's, with probability measure given by $\widehat{\mu}_{t_0}$. $\qquad \square$

Together, Theorems 5.2 and 5.3 prove the U-statistic SLLN for weak Bernoulli sequences with *all* kernels which are bounded by an integrable product (Theorem U(iii)). The ESP in Example 4.5 is weakly Bernoulli, the kernel there is integrable (and continuous except at one point), but nevertheless the U-statistic SLLN fails. This shows that even when the ESP is weakly Bernoulli, one cannot omit entirely the condition of boundedness by integrable products in Theorem U.

6. An Application

Suppose that (Ω, ρ) is a metric space, and that $T : \Omega \to \Omega$ is a measurable map with invariant measure μ. In many examples it turns out that there exists a constant δ such that

$$D(\epsilon) = \mu \times \mu(\{(x, y) : \rho(x, y) \leq \epsilon\}) \sim C\epsilon^\delta$$

as $\epsilon \to 0$. The exponent δ is called the *correlation dimension* of μ. For example, if $\Omega \subset \mathbb{R}^m$ and μ is absolutely continuous with bounded density, then δ is the topological dimension m.

One possible estimation procedure for δ (suggested in [G-P]) is to estimate $D(\epsilon)$ by its empirical analog

$$D_n(\epsilon) := \frac{1}{n(n-1)} \#\{1 \leq i \neq j \leq n : \rho(X_i, X_j) \leq \epsilon\} = U_h(X_1, \ldots, X_n)$$

where $h : \mathbb{R}^2 \to \mathbb{R}$ is $h(x, y) = 1_{\{\rho(x,y) \leq \epsilon\}}$. A regression procedure based on $\log D(\epsilon) \approx \log C + d \log \epsilon$ is then used to estimate δ.

Note that this kernel is covered by Theorem U(ii), and $D_n(\epsilon)$ converges a.s. in the case

$$F \times F(\{(x, y) : \rho(x, y) = \epsilon\}) = 0$$

where F is the distribution of X. This convergence is also established (by different methods) in [Pe, Theorem 1].

For $\Omega \subset \mathbb{R}^m$ (and $\rho(x, y) = |x - y|$), an alternative procedure (presented in [Ta]) is first to generate iid observations $R_i = |W_i - Y_i|$ where dist.$(W_i, Y_i) = \mu \times \mu$. Assuming that actually for some $\epsilon_0 > 0$,

$$D(\epsilon) = C \cdot \epsilon^\delta \quad \text{for all } \epsilon \leq \epsilon_0,$$

the conditional distribution of $Z_i = R_i/\epsilon_0$ given $R_i \leq \epsilon_0$ is

$$P(Z_i \leq t | Z_i \leq 1) = t^\delta, \qquad 0 \leq t \leq 1.$$

Deleting the observations Z_i that exceed 1, it is then possible to estimate δ by standard methods such as maximum likelihood or UMVU. Note that the maximum likelihood estimate of δ is the reciprocal of

$$\frac{1}{n} \sum_{i=1, Z_i \leq 1}^{n} -\log Z_i = \frac{1}{n} \sum_{i=1, |W_i - Y_i| \leq \epsilon_0}^{n} -\log \left(\frac{|W_i - Y_i|}{\epsilon_0} \right),$$

while the UMVU estimator is $n^{-1}(n-1)$ times this.

The problem with this procedure is that it is not clear how to generate iid observations of $|W_i - Y_i|$ based on the non-iid $X_i(\omega) = T^i\omega$. A natural idea to remedy this would be to study the average of all $\log |X_i - X_j|$, $1 \leq i \neq j \leq n$:

$$U_{\widehat{h}}(X_1, \ldots, X_n) = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} -\log|X_i - X_j|,$$

where $\widehat{h} : \mathbb{R}^2 \to \mathbb{R}$ is $\widehat{h}(x, y) = -\log|x - y|$.

Unfortunately, this cannot work. This kernel $\log|x - y|$ does not satisfy the criteria of Theorem U, and moreover the following example shows that an SLLN for *U*-statistics based on it cannot be expected (even when the underlying ESP is WB).

Example 6.1. Let $W_1, W_2, \ldots$ be iid with a continuous distribution F such that

$$E\left(|\log|W_1 - W_2||\right) < \infty,$$

and let $Y_1, Y_2, \ldots$ be iid Bernoulli with $P(Y_i = 1) = p$, $0 < p < 1$, independent of $(W_1, \ldots)$. Define a stationary, weakly Bernoulli process with invariant distribution F by $X_1 = W_1$ and $X_n = W_n(1 - Y_n) + X_{n-1}Y_n$ for $n > 1$. Now (with probability one) there are infinitely many n with $X_n = X_{n+1}$ so the *U*-statistic with kernel $\widehat{h}(x, y) = -\log|x - y|$ does not satisfy the SLLN, diverging to ∞.

Acknowledgment

References

[A]   Aaronson, J., *An ergodic theorem with large normalising constants*, Israel J. Math. **38** (1981), 182–188. MR **83f:**28014

[Ber]   Berbee, H., *Random walks with stationary increments and renewal theory*, CWI Tracts, no. 112, Amsterdam, 1980. MR **81e:**60093

[Bi]   Billingsley, P., *Probability and measure*, Wiley, New York, 1979. MR **80h:**60001

[Ga]   Garsia, A., *Topics in Almost Everywhere Convergence*, Markham Publishing, Chicago, 1970. MR **41:**5869

[G-H]   Gilat, D. and T. Hill, *One-sided refinements of the strong law of large numbers and the Glivenko-Cantelli Theorem*, Ann. Probab. **20** (1992), 1213-1221. MR **93g:**60064

[G-P]   Grassberger, P. and I. Procaccia, *Characterization of strange attractors*, Phys. Rev. Letters **50** (1983), 346-349. MR **84k:**58141

[He]   Helmers, R., *A strong law of large numbers for linear combinations of order statistics*, Technical Report SW 50/77 Math. Centrum, Amsterdam (1977).

[Hoe]   Hoeffding, W., *The strong law of large numbers for U-statistics*, Univ. N. Carolina, Institute of Stat. Mimeo series **302** (1961).

[Le]   Lehmann, E.L., *Theory of point estimation*, Wiley, New York, 1983. MR **85a:**62001

[Pe]   Pesin, Y., *On rigorous mathematical definitions of correlation dimension and generalized spectrum for dimensions*, J. Stat. Physics **71** (1993), 529–547. MR **94d:**28008

[Se]   Serfling, R., *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980. MR **82a:**62003

[Ta]   Takens, F., *On the numerical determination of the dimension of an attractor*, Dynamical Systems and Bifurcations, Lecture Notes in Mathematics, vol. 1125, Springer, Berlin, 1985, pp. 99-106. MR **87c:**58063

[vN]   von Neumann, J., *Zur Operatorenmethode in der klassischen Mechanik*, Ann. Math **33** (1932), 587–642.

[vZ]   van Zwet, W., *A strong law for linear functions of order statistics*, Ann. Probab. **8** (1980), 986-990. MR **81j:**60043

School of Mathematical Sciences, Tel Aviv University, 69978 Tel Aviv, Israel
E-mail address: `aaro@math.tau.ac.il`

Department of Mathematics, Oregon State University, Corvallis, Oregon 97331-4605, USA
E-mail address: `burton@math.orst.edu`

Department of Mathematics, University of Groningen, Groningen, Netherlands
E-mail address: `dehling@math.rug.nl`

School of Mathematical Sciences, Tel Aviv University, 69978 Tel Aviv, Israel
E-mail address: `gilat@math.tau.ac.il`

School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332-0160, USA
E-mail address: `hill@math.gatech.edu`

Institute of Mathematics, Hebrew University of Jerusalem, Jerusalem, Israel
E-mail address: `weiss@math.huji.ac.il`