OSCILLATIONS IN THE STRONG LAW OF LARGE NUMBERS

MORENIKEJI NERI

Department of Mathematics, Technische Universität Darmstadt
E-mail: neri@mathematik.tu-darmstadt.de

Abstract. It is well known that the sample means in the Strong Law of Large Numbers can converge arbitrarily slowly. The convergence of ergodic averages also suffers from such ineffectiveness, and in this situation a prominent way to study the quantitative nature of the convergence of such processes is through looking at the stochastic *fluctuations* or *oscillations*. As the Pointwise Ergodic Theorem generalises the Strong Law of Large Numbers when the random variables are independent identically distributed, the quantitative oscillatory results from ergodic theory carry over naturally to this setting.

We initiate the investigation of the oscillatory behaviour of the averages in the Strong Law of Large Numbers in the case where the random variables carry dependencies. In this paper, we consider the case where the random variables are pairwise independent and construct *learnable rates of uniform convergence*. This notion, introduced by the author and Powell (c.f. *Trans. Amer. Math. Soc. Series B*, 12:974–1019, 2025), strengthens Tao's notion of metastability and provides a quantitative bound on how far one must look to locate a region of local stability (that is, a region with no oscillations) with high probability.

Keywords: Proof mining; Fluctuations; Laws of Large Numbers
MSC2020 Classification: 60F15, 60G10, 03F10.

1. Introduction

Fix a dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, \tau)$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\tau : \Omega \to \Omega$ is a measurable, invertible, measure-preserving transformation. For $f \in L^1 := L^1(\Omega, \mathcal{F}, \mathbb{P})$, the *Birkhoff Pointwise Ergodic Theorem* implies that the ergodic averages

$$A_n f(\omega) := \frac{1}{n} \sum_{i=1}^{n} f(\tau^i \omega)$$

converge almost surely.

Because the averages in the Pointwise Ergodic Theorem can converge arbitrarily slowly, a prominent strategy for studying the quantitative behaviour of the theorem, and more generally the almost sure convergence of related ergodic averages, is to analyse the oscillatory behaviour of these processes.

For a sequence of real numbers $\{x_n\}$, we define $J_{N,\varepsilon}\{x_n\}$ to be the number of ε-fluctuations occurring in the initial segment $\{x_0, \ldots, x_{N-1}\}$. That is, $J_{N,\varepsilon}\{x_n\}$ is the maximal $k \in \mathbb{N}$ such that there exist indices

$$i_1 < j_1 \leqslant i_2 < j_2 \leqslant \cdots \leqslant i_k < j_k < N \quad \text{with} \quad |x_{i_\ell} - x_{j_\ell}| > \varepsilon$$

for all $\ell = 1, \ldots, k$.

Finally, we define the total number of ε-fluctuations in the sequence $\{x_n\}$ by

$$J_\varepsilon\{x_n\} := \lim_{N \to \infty} J_{N,\varepsilon}\{x_n\},$$

which may be infinite.

It is clear that the convergence of a sequence of real numbers $\{x_n\}$ is equivalent to it having finitely many fluctuations, that is, for all $\varepsilon > 0$

$$J_\varepsilon\{x_n\} < \infty.$$

Furthermore, we have the analogous situation in the stochastic setting where a sequence of random variables $\{X_n\}$ converging almost surely is equivalent to

$$\lim_{k\to\infty} \mathbb{P}\left(J_\varepsilon\{X_n\} \geqslant k\right) = 0$$

for all $\varepsilon > 0$, with a function $e : (0,1] \times (0,1] \to \mathbb{R}^+$ satisfying

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geqslant e(\lambda,\varepsilon)\right) \leqslant \lambda$$

for all $\lambda, \varepsilon \in (0,1]$ known as a modulus of finite fluctuations (c.f. Definition 4.5 of [22]).

In [9], Jones, Kaufman, Rosenblatt and Wierdl obtain the following quantitative strengthening of the Pointwise Ergodic Theorem:

Theorem 1.1 (c.f. Theorem 5.1 of [9])**.** *There exists a constant C such that, for any $f \in L^1$ and all $k, \varepsilon > 0$,*

$$\mathbb{P}\left(J_\varepsilon\{A_n f\} \geqslant k\right) \leqslant \frac{C\|f\|_1}{\varepsilon\sqrt{k}}.$$

It is clear that $e(\lambda,\varepsilon) = (C\|f\|_1/\varepsilon\lambda)^2$ is a modulus of finite fluctuations for $\{A_n f\}$ in Theorem 1.1.

Remark 1.2. Theorem 1.1 was originally conjectured by Ivanov [11], who, by entirely different methods, was only able to show that

$$\mathbb{P}\left(J_\varepsilon\{A_n f\} \geqslant k\right) \leqslant C\sqrt{\frac{\log k}{k}},$$

where C is a constant depending only on $\|f\|_1/\varepsilon$. In Example 5.4 of [22], it is shown that a slightly stronger bound than the one obtained by Ivanov (though still weaker than that in Theorem 1.1) follows from a general result stated as Theorem 5.1 in the same reference.

A classic application of the Pointwise Ergodic Theorem is the Strong Law of Large Numbers, which states that for a sequence of independent identically distributed (iid) random variables $\{X_n\}$ with $\mathbb{E}(X_1) = 0$ and $\mathbb{E}(|X_1|) < \infty$, we have

$$\frac{1}{n}\sum_{k=1}^{n} X_k \to 0 \quad \text{almost surely.}$$

In particular, it is well known that one can construct a dynamical system together with an integrable function f such that the almost sure convergence of the ergodic averages of f is equivalent to the almost sure convergence of the sample means of the corresponding stochastic process. This construction also shows directly that Theorem 1.1 (and other oscillation results concerning the Pointwise Ergodic Theorem) applies to the averages in the Strong Law of Large Numbers.

Despite this connection, ergodic theory, or more precisely the study of almost everywhere convergence of ergodic averages, does not fully subsume the study of almost sure convergence of sample means for general stochastic processes, particularly when they carry dependencies. Different techniques are typically required to address the latter. It is therefore natural to ask whether one can also investigate the oscillatory behaviour of the sample means in such settings.

We finish here by noting that studying the oscillations of classes of stochastic processes with ineffective convergence speeds is common outside ergodic theory. For example, Doob's well-known upcrossing inequality [6] for L_1 sub- and supermartingales, along with other related results on the oscillations of martingales, for example [9, 10, 11, 21, 22]. In addition, we highlight the work of Hochman [8], who obtained upcrossing inequalities for processes associated with information theory using arguments similar to those for ergodic averages in [12].

1.1. **Main results.** In this paper we focus on the oscillations of the sample means of pairwise iid random variables in the Strong Law of Large Numbers.

Here and throughout the rest of the article, for $a, b \in \mathbb{N}$, we write $[a; b] := [a, b] \cap \mathbb{N}$. Our main result is the following:

Theorem 1.3. *Let $\{X_n\}$ be a sequence of pairwise independent and identically distributed random variables with $\mathbb{E}(X_n) = 0$ and $\mathbb{E}(|X_n|) =: \mu$. Define $S_n := \sum_{k=1}^{n} X_k$ and $A_n := S_n/n$.*

For every $\varepsilon, \lambda \in (0, 1]$, $R > 1$, and all sequences of natural numbers $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \dots$ satisfying $b_n/a_n > R$, we have

$$\mathbb{P}\left(\forall m \leqslant \frac{\kappa\rho^4}{\varepsilon^5\lambda^2}\, \exists i, j \in [a_m; b_m] \left(\left|A_i - A_j\right| > \varepsilon\right)\right) \leqslant \lambda,$$

where $\kappa > 0$ is a constant and

$$\rho := \max\left\{1, \frac{1}{3(R-1)}, \frac{\mu}{2}\right\}.$$

We can give the very crude bound $\kappa < 10^{11}$.

Remark 1.4. A function $e\colon (0, 1] \times (0, 1] \to \mathbb{R}^+$ is said to be a *learnable rate of pointwise convergence* (c.f. Definition 4.15 of [22]) for a sequence of random variables $\{X_n\}$ if for all $\varepsilon, \lambda \in (0, 1]$ and $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \dots$,

$$\mathbb{P}\left(\forall i \leqslant e(\lambda, \varepsilon)\, \exists k, l \in [a_i; b_i] \left(|X_k - X_l| > \varepsilon\right)\right) \leqslant \lambda.$$

One can easily verify that moduli of finite fluctuations are themselves learnable rates of pointwise convergence (see, for example, Theorem 4.16 of [22]), and so Theorem 1.1 implies that $e(\lambda, \varepsilon) = (C\|f\|_1/(\varepsilon\lambda))^2$ is a learnable rate of pointwise convergence for $\{A_n f\}$. In particular, this yields such a rate in the case where the random variables are iid, or more generally when they form a stationary sequence.

One can interpret Theorem 1.3 as a learnable rate of pointwise convergence for *lacunary* sequences (see Section 5 for a discussion of how similar lacunary conditions have arisen in ergodic theory) and we also note that the bound in Theorem 1.3 is asymptotically as good in the variable λ as the bound obtained from Theorem 1.1, a bound which is shown to be optimal in Theorem 3.9 of [10].

In fact, we prove a stronger result than Theorem 1.3:

Theorem 1.5. *Let $\{X_n\}$ be as Theorem 1.5. For every $\varepsilon, \lambda \in (0, 1]$, $R > 1$, and all sequences of natural numbers $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \dots$ satisfying $b_n/a_n > R$, there exists*

$$m \leqslant \frac{\kappa\rho^4}{\varepsilon^5\lambda^2}$$

satisfying

$$\mathbb{P}\left(\exists i, j \in [a_m; b_m] \left(|A_i - A_j| > \varepsilon\right)\right) \leqslant \lambda,$$

where $\kappa > 0$ is a constant and

$$\rho := \max\left\{1, \frac{1}{3(R-1)}, \frac{\mu}{2}\right\}.$$

Remark 1.6. This strengthening is related to the notion of a *learnable rate of uniform convergence* (c.f. [21, 22]), which is a function $e\colon (0,1] \times (0,1] \to \mathbb{R}^+$ satisfying that for all $\varepsilon, \lambda \in (0,1]$ and $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$, there exists $i \leqslant e(\lambda, \varepsilon)$ such that

$$\mathbb{P}\left(\exists k, l \in [a_i; b_i]\left(|X_k - X_l| > \varepsilon\right)\right) \leqslant \lambda.$$

A learnable rate of uniform convergence provides a bound on how far one must look, when given a selection of intervals, in order to locate one that contains no oscillations, with high probability. Moreover, this bound is independent of the particular selection of intervals. If the bound is allowed to depend on the selection of intervals, then one arrives at a notion that is computationally equivalent to the so-called *rates of (uniform) metastability* in the sense of Tao [27, 28] (see Remark 2.10). We also note that, as in the case of Theorem 1.3, we obtain a learnable rate of uniform convergence, in Theorem 1.5, for lacunary sequences. We make this notion precise in Definition 2.4.

1.2. **An outline of the proof of Theorem 1.5.** To prove Theorem 1.5, we set

$$Y_n := X_n I\{X_n \leqslant n\} - \mathbb{E}(X_n I\{X_n \leqslant n\}), \quad z_n := \sum_{k=1}^{n} \mathbb{E}(X_k I\{X_k \leqslant k\}).$$

Then we have the following decomposition of A_n into three terms:

$$A_n = \frac{1}{n}\sum_{k=1}^{n} Y_k + \frac{z_n}{n} + \frac{1}{n}\sum_{k=1}^{n} X_k I\{|X_k| > k\}. \tag{1}$$

Our result will follow by analysing the oscillations of the terms in this decomposition and applying results on the oscillations of sums of sequences of random variables, which we present in Section 2.

For the first term in the decomposition, we prove a quantitative version of Kolmogorov's Strong Law of Large Numbers for pairwise independent random variables, as given in [5]. More precisely, in Section 3 we show that if $\{Q_n\}$ is a sequence of pairwise independent random variables satisfying $\mathbb{E}(Q_n)/n \leqslant W$ for all $n \in \mathbb{Z}^+$, $\mathbb{E}(Q_n) = 0$, and

$$\sum_{k=1}^{\infty} \frac{\operatorname{Var}(Q_k)}{k^2} \leqslant V$$

for some $V > 1$ and $W > 0$, then for every $\varepsilon, \lambda \in (0,1]$, $R \geqslant 1$, and sequence of natural numbers $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ with $b_n/a_n > R$, we have

$$\exists m \leqslant \frac{cV\rho^3}{\varepsilon^5\lambda} \quad \mathbb{P}\left(\exists i, j \in [a_m; b_m]\left(\left|\frac{1}{i}\sum_{k=1}^{i} Q_k - \frac{1}{j}\sum_{k=1}^{j} Q_k\right| > \varepsilon\right)\right) \leqslant \lambda, \tag{2}$$

for a constant c (for which we provide a crude bound) and ρ that depends on R and W. This result applies directly to $\{Y_n\}$, noting that from standard calculations (see Lemma 2.4.3 of [7], for example) we have

$$\sum_{k=1}^{\infty} \frac{\operatorname{Var}(Y_k)}{k^2} \leqslant 4\mu.$$

The proof of (2) requires extending notions of fluctuations to arbitrary logical formulas, in a manner similar to [19, 22], and it also requires an analysis of how these notions behave with respect to disjunctions (see Lemma 2.11).

The oscillations of the last two terms in the decomposition (1) are not well behaved in general. However, if we assume that X_n is nonnegative, then $\{\mathbb{E}(X_n I\{X_n \leqslant n\})\}$ is a monotone sequence of real numbers. Hence, for all $n \in \mathbb{N}$,

$$\mathbb{E}(X_{n+1} I\{X_{n+1} \leqslant n+1\}) \geqslant \frac{z_n}{n},$$

and therefore,

$$\frac{z_{n+1}}{n} = \frac{z_n}{n} + \frac{1}{n+1}\left(\mathbb{E}(X_{n+1} I\{X_{n+1} \leqslant n+1\}) - \frac{z_n}{n}\right) \leqslant \frac{z_n}{n}.$$

Thus, $\{z_n/n\}$ is a monotone sequence of nonnegative real numbers bounded by μ, and we shall see in Section 2 that such sequences have well-behaved oscillatory properties.

The last term in (1) also behaves nicely if we assume X_n is nonnegative. A key observation is that we can establish an appropriate version of Theorem 1.5 separately for the positive and negative parts of X_n, and then use the decomposition $X_n = X_n^+ - X_n^-$ together with results on the oscillations of sums of random variables.

The complete proof of Theorem 1.5 will be presented in Section 4.

1.3. **The proof mining program.** This work can be viewed as a contribution to the *proof mining* program of Kohlenbach [13]. Proof mining is a subfield of mathematical logic that applies tools from proof theory to extract computational content from seemingly non-constructive mathematical proofs.

Approaching probability theory from the perspective of proof mining has led to an abstract framework for probability that is amenable to the extraction of quantitative information. For instance, the idea of studying the oscillatory behaviour of abstract logical formulas, introduced in Section 2.2, was already pursued in [19] and [22], with the latter using such abstract considerations to obtain a quantitative version of the Martingale Convergence Theorem. Furthermore, in [18], a formal logical system providing an abstract treatment of probability theory is developed, together with a corresponding *metatheorem* that guarantees the extraction of highly uniform quantitative information from proofs.

This abstract approach to quantitative probability theory has given rise to a growing body of case studies, including further results on the Strong Law of Large Numbers by the author [15, 17], where we studied the quantitative nature of the convergence of the sample means in the Strong Law of Large Numbers when the random variables carried dependencies, under the assumption that high moments exist. In addition, work has been carried out on the asymptotic behaviour of more general stochastic processes [16, 19, 22], on stochastic optimization [20, 21, 24], and other classical results in probability theory connected with the Borel–Cantelli lemmas and zero–one laws [1, 25].

2. Preliminary definitions and lemmas

2.1. **Deterministic convergence.** We call a function $b : (0,1] \to \mathbb{R}^+$ a *bound on the number of fluctuations* of $\{x_n\}$ if, for every $\varepsilon \in (0,1]$, we have

$$J_\varepsilon\{x_n\} \leqslant b(\varepsilon).$$

An equivalent formulation, which is more convenient for our purposes, is that of a *learnable rate of uniform convergence* [19, 21, 22].

Definition 2.1. Let $\{x_n\}$ be a sequence of real numbers. A function $\phi : (0,1] \to \mathbb{R}^+$ is called an *R-learnable rate of convergence* for $\{x_n\}$ if

$$\exists n \leqslant \phi(\varepsilon)\, \forall i, j \in [a_n; b_n]\, (|x_i - x_j| \leqslant \varepsilon)$$

for any $\varepsilon \in (0,1]$ and any sequence of natural numbers $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ satisfying $b_n/a_n > R$.

A 1-learnable rate of convergence is simply called a *learnable rate of convergence.*

It is clear that ϕ is a learnable rate of convergence if and only if it is a bound on the number of fluctuations. We shall also need the known fluctuation bounds for bounded monotone sequences.

Lemma 2.2. *Suppose $\{x_n\}$ is a monotone sequence of nonnegative real numbers such that there exists $C > 0$ with $x_n \leqslant C$ for all n. Then the function $\phi : (0,1] \to \mathbb{R}^+$ defined by*

$$\phi(\varepsilon) := \frac{C}{\varepsilon}$$

is a learnable rate of convergence (or equivalently, a bound on the number of fluctuations) for $\{x_n\}$.

Proof. Let $\varepsilon > 0$ be given, and suppose there exists $k \in \mathbb{N}$ satisfying

$$i_1 < j_1 \leqslant i_2 < j_2 \leqslant \ldots \leqslant i_k < j_k \quad \text{with} \quad |x_{i_l} - x_{j_l}| > \varepsilon.$$

Since $x_n \leqslant C$, we must have $C \geqslant x_{j_k} > k\varepsilon$, which implies $k \leqslant C/\varepsilon$, as claimed. □

It will be convenient to apply the previous lemma to bounded sums of nonnegative real numbers.

Lemma 2.3. *Let $\{x_n\}$ be a sequence of nonnegative real numbers and let $C > 0$ satisfy*

$$\sum_{k=0}^{\infty} x_k \leqslant C.$$

Then, for any $\varepsilon \in (0,1]$ and any sequence of natural numbers $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$, there exists $n \leqslant 1 + 2C/\varepsilon$ such that

$$\sum_{k=a_n}^{b_n} x_k \leqslant \varepsilon.$$

Proof. Let $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ and $\varepsilon \in (0,1]$ be given. Since the partial sums

$$\sum_{k=0}^{n} x_k$$

form a monotone bounded sequence of nonnegative numbers, by considering the subsequence

$$a_1 - 1 < b_1 \leqslant a_3 - 1 < b_3 \leqslant a_5 - 1 < b_5 \leqslant \ldots$$

and applying Lemma 2.2, there exists $m \leqslant C/\varepsilon$ such that for all $i, j \in [a_{2m+1} - 1; b_{2m+1}]$,

$$\left| \sum_{k=0}^{i} x_k - \sum_{k=0}^{j} x_k \right| \leqslant \varepsilon.$$

In particular,

$$\sum_{k=a_{2m+1}}^{b_{2m+1}} x_k \leqslant \varepsilon,$$

so taking $n := 2m + 1 \leqslant 2C/\varepsilon + 1$ yields the desired result. □

2.2. **Stochastic convergence.** We now introduce the notion of an *R-learnable rate of uniform convergence*, which will serve as the central tool in our analysis of the oscillatory behaviour associated with the Strong Law of Large Numbers for pairwise iid random variables.

Definition 2.4. Let $\{X_n\}$ be a sequence of random variables, $R \geqslant 1$, and let X be a random variable. A function $\phi : (0,1] \to \mathbb{R}^+$ is called an *R-learnable rate of uniform convergence* for $\{X_n\}$ if, for any $\varepsilon, \lambda \in (0,1]$ and any sequence of natural numbers $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ with $b_n/a_n > R$ for all $n \in \mathbb{N}$,

$$\exists n \leqslant \phi(\lambda, \varepsilon)\, \mathbb{P}\left(\exists i, j \in [a_n; b_n]\, (|X_i - X_j| > \varepsilon)\right) \leqslant \lambda.$$

We say ϕ is an *R-learnable rate of uniform convergence to* X for $\{X_n\}$ if

$$\exists n \leqslant \phi(\lambda, \varepsilon)\, \mathbb{P}\left(\exists i \in [a_n; b_n]\, (|X_i - X| > \varepsilon)\right) \leqslant \lambda.$$

A 1-learnable rate of uniform convergence (to X) is simply called a learnable rate of uniform convergence (to X), corresponding to the notion introduced in [21, 22].

Remark 2.5. Learnable rates of uniform convergence generalise learnable rates of convergence for sequences of real numbers. More precisely, suppose $\{X_n\}$ is a sequence of real numbers with an R-learnable rate of convergence ψ. Then $\{X_n\}$, treated as a sequence of constant random variables, has an R-learnable rate of uniform convergence given by

$$\phi(\lambda, \varepsilon) := \psi(\varepsilon).$$

Lemma 2.6. *Let $\{X_n\}$ be a sequence of random variables, $R \geqslant 1$, and X a random variable. If ϕ is an R-learnable rate of uniform convergence to X for $\{X_n\}$, then*

$$\psi(\lambda, \varepsilon) := \phi(\lambda, \varepsilon/2)$$

is an R-learnable rate of uniform convergence for $\{X_n\}$.

Proof. Let $\varepsilon, \lambda \in (0,1]$ and $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ with $b_n/a_n > R$ be given. By assumption, there exists $n \leqslant \phi(\lambda, \varepsilon/2)$ such that

$$\mathbb{P}\left(\exists i \in [a_n; b_n]\, (|X_i - X| > \varepsilon/2)\right) \leqslant \lambda.$$

By the triangle inequality, if

$$\exists i, j \in [a_n; b_n]\, (|X_i - X_j| > \varepsilon),$$

then

$$\exists i \in [a_n; b_n]\, (|X_i - X| > \varepsilon/2),$$

and the result follows. ∎

We extend the notion of learnable rates of uniform convergence to arbitrary logical formulas. We also prove elementary results, which will be important for our later arguments, concerning the behaviour of such rates with respect to disjunctions. Similar results in this direction were important in [19, 21, 22].

Definition 2.7 (Measurable formula, c.f. [22])**.** A logical formula $\varphi(\omega, x_1, \ldots, x_n)$, with parameters $x_1, \ldots, x_n$ and $\omega \in \Omega$, is *measurable* if for all parameters $x_1, \ldots, x_n$,

$$\{\omega \in \Omega : \varphi(\omega, x_1, \ldots, x_n)\} \in \mathcal{F}.$$

For such a formula, we write $\varphi(x_1, \ldots, x_n) := \{\omega \in \Omega : \varphi(\omega, x_1, \ldots, x_n)\}$.

If $\varphi(\omega, n)$ is measurable with $n \in \mathbb{N}$, we define

$$\exists n\, \varphi(n) := \bigcup_{n \in \mathbb{N}} \varphi(n), \quad \forall n\, \varphi(n) := \bigcap_{n \in \mathbb{N}} \varphi(n).$$

Definition 2.8. Let $A(\omega, x)$ be a measurable formula with parameters $x = (x_1, \ldots, x_n)$ taking natural numbers and $R \geqslant 1$. A function $\phi : (0,1] \to \mathbb{R}^+$ is called an *R-learnable uniform rate* for A if, for any $\lambda \in (0,1]$ and sequence of natural numbers $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ with $b_n/a_n > R$,

$$\exists n \leqslant \phi(\lambda)\, \mathbb{P}\big(\exists x_1, \ldots, x_n \in [a_n; b_n]\, A(x)\big) \leqslant \lambda.$$

Remark 2.9. Let $\{X_n\}$ be a sequence of random variables. A function ϕ is an R-learnable rate of uniform convergence for $\{X_n\}$ if and only if, for each $\varepsilon \in (0,1]$, the function $\phi_\varepsilon(\lambda) := \phi(\lambda, \varepsilon)$ is an R-learnable uniform rate for the formula

$$A_\varepsilon(\omega, i, j) :\equiv |X_i(\omega) - X_j(\omega)| > \varepsilon.$$

Remark 2.10. *R-learnable uniform rates* are special instance of *dependent learnable uniform rates* [19] which are functions $\phi : (0,1] \times \mathbb{N}^\mathbb{N} \times \mathbb{N}^\mathbb{N} \to \mathbb{R}^+$ that satisfy, for any $\lambda \in (0,1]$ and sequence of natural numbers $a := \{a_n\}$ and $b := \{b_n\}$ with $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ and $b_n/a_n > R$,

$$\exists n \leqslant \phi(\lambda, a, b)\, \mathbb{P}\big(\exists x_1, \ldots, x_n \in [a_n; b_n]\, A(x)\big) \leqslant \lambda.$$

In the case of *R-learnable uniform rates*, the dependence on the sequences a, b is only through a subsequence which grows hyper-geometrically with some fixed rate.

It is shown in [19] that dependent learnable uniform rates are computationally equivalent to so-called *rates of uniform metastability* (formulated for arbitrary measurable formulas), with explicit computations demonstrating how to convert between the two notions. We do not spell out the definition of metastability here, since we will not need it, but we note that this notion provides a natural quantitative interpretation of the convergence of deterministic and stochastic sequences, and it has found applications in many areas of mathematics, from ergodic theory to stochastic optimization. For further discussions of metastability we refer the reader to [2, 3, 22, 26, 27, 28].

Lemma 2.11. *Let $A_1(\omega, y^{(1)}), \ldots, A_L(\omega, y^{(L)})$ be measurable formulas with respective tuples of parameters $y^{(1)}, \ldots, y^{(L)}$ and R-learnable uniform rates $\phi_1, \ldots, \phi_L$, for some $R \geqslant 1$. Then $\bigvee_{i=1}^{L} A_i$ has an R-learnable uniform rate given by*

$$\phi(\lambda) := \sum_{i=1}^{L} \phi_i(\lambda/L).$$

Proof. The proof is by induction on L. The case $L = 1$ is trivial. Suppose the lemma holds for L, and consider $L + 1$ formulas $A_1, \ldots, A_{L+1}$. Define

$$\psi(\lambda) := \sum_{i=1}^{L} \phi_i(\lambda/L), \quad \phi(\lambda) := \psi\left(\frac{\lambda L}{L+1}\right) + \phi_{L+1}\left(\frac{\lambda}{L+1}\right).$$

Assume for contradiction that for all $n \leqslant \phi(\lambda)$,

$$\mathbb{P}\left(\exists y^{(1)}, \ldots, y^{(L+1)} \in [a_n; b_n] \bigcup_{i=1}^{L+1} A_i(y^{(i)})\right) \geqslant \lambda.$$

We have, by sub-additivity of $\mathbb{P}$,

$$\mathbb{P}\left(\exists y^{(1)}, \ldots, y^{(L+1)} \in [a_n; b_n] \bigcup_{i=1}^{L+1} A_i(y^{(i)})\right) \leqslant \mathbb{P}\Big(\exists y^{(1)}, \ldots, y^{(L)} \in [a_n; b_n] \bigcup_{i=1}^{L} A_i(y^{(i)})\Big) + \mathbb{P}\left(\exists y^{(L+1)} \in [a_n; b_n]\, A_{L+1}(y^{(L+1)})\right).$$

Hence, either

$$\mathbb{P}\left(\exists y^{(1)},\dots,y^{(L)}\in[a_n;b_n]\bigcup_{i=1}^{L}A_i(y^{(i)})\right)\geqslant\frac{\lambda L}{L+1}$$

or

$$\mathbb{P}\left(\exists y^{(L+1)}\in[a_n;b_n]\,A_{L+1}(y^{(L+1)})\right)\geqslant\frac{\lambda}{L+1},$$

for all $n\leqslant\phi(\lambda)$. But then it follows that either there exists a subsequence $a_{n_0}<b_{n_0}\leqslant a_{n_1}<b_{n_1}\leqslant\dots$ such that

$$\forall k\leqslant\psi\left(\frac{\lambda L}{L+1}\right)\mathbb{P}\left(\exists y^{(1)},\dots,y^{(L)}\in[a_{n_k};b_{n_k}]\left(\bigcup_{i=1}^{L}A_i(y^{(i)})\right)\right)\geqslant\frac{\lambda L}{L+1}$$

or a subsequence $a_{m_0}<b_{m_0}\leqslant a_{m_1}<b_{m_1}\leqslant\dots$ such that

$$\forall k\leqslant\phi_{L+1}\left(\frac{\lambda}{L+1}\right)\mathbb{P}\left(\exists y^{(L+1)}\in[a_{m_k};b_{m_k}]\,A_{L+1}\left(y^{(L+1)}\right)\right)\geqslant\frac{\lambda}{L+1}.$$

The first case contradicts the induction hypothesis and the second the defining property of ϕ_{L+1}. Therefore, ϕ is a R-learnable uniform rate. ☐

The previous lemma immediately allows us to calculate R-learnable rates of uniform convergence for the sums of random variables.

Lemma 2.12. *Suppose $\{X_n^{(1)}\},\dots,\{X_n^{(L)}\}$ are sequences of random variables with respective R-learnable rates of uniform convergence (to $X^{(1)},\dots,X^{(L)}$) $\phi_1,\dots,\phi_L$, for $R\geqslant1$. Then $\{\sum_{m=1}^{L}X_n^{(m)}\}$ has an R-learnable rate of uniform convergence (to $\sum_{m=1}^{L}X^{(m)}$) given by*

$$\phi(\lambda,\varepsilon):=\sum_{i=1}^{L}\phi_i(\lambda/L,\varepsilon/L).$$

Proof. We only prove the Cauchy case; the other case can be proven similarly.

Fix $\varepsilon\in(0,1]$ and define

$$A_\varepsilon(\omega,i,j):=\left|\sum_{m=1}^{L}X_i^{(m)}-\sum_{m=1}^{L}X_j^{(m)}\right|>\varepsilon,\quad A_\varepsilon^m(\omega,i,j):=|X_i^{(m)}-X_j^{(m)}|>\varepsilon/L.$$

Since each A_ε^m has an R-learnable uniform rate $\phi_i(\lambda,\varepsilon/L)$, Lemma 2.11 implies

$$\phi(\lambda,\varepsilon):=\sum_{i=1}^{L}\phi_i(\lambda/L,\varepsilon/L)$$

is an R-learnable uniform rate for $\bigvee_{m=1}^{L}A_\varepsilon^m$. By the triangle inequality,

$$A_\varepsilon\implies\bigvee_{m=1}^{L}A_\varepsilon^m$$

and so ϕ is an R-learnable uniform rate for A_ε, which proves the result. ☐

3. A Quantitative Kolmogorov's Strong Law for Pairwise Independent Random Variables

In this section, we establish a quantitative version of Kolmogorov's Strong Law of Large Numbers for pairwise independent random variables, following [5]. The quantitative rates derived here will be used to prove our main result, Theorem 1.5, but are also of independent interest.

Throughout this section, let $\{X_n\}$ be a sequence of pairwise independent random variables, and define

$$S_n := \sum_{k=1}^{n} X_k, \quad A_n := \frac{S_n}{n}.$$

Set

$$\mu_n := \mathbb{E}(|X_n|), \quad z_n := \sum_{k=1}^{n} \mu_k.$$

Assume that there exists a constant $W \geqslant 0$ such that, for all $n \in \mathbb{N}$,

$$\frac{z_n}{n} \leqslant W.$$

We adopt the notation from [17], which is based on [5]:

- For each $\delta > 0$, let
$$L_\delta := \left\lfloor \frac{W}{\delta} \right\rfloor.$$
- For $\delta > 0$, $\alpha > 1$, $m \in \mathbb{N}$, and $0 \leqslant s \leqslant L_\delta$, define
$$C_{\alpha,s,\delta,m} := \left\{ n \in \mathbb{N} \mid \alpha^m \leqslant n < \alpha^{m+1}, \ \frac{z_n}{n} \in [s\delta, (s+1)\delta) \right\}.$$
- If $C_{\alpha,s,\delta,m}$ is nonempty, let
$$k_s^-(m) := \min C_{\alpha,s,\delta,m}, \quad k_s^+(m) := \max C_{\alpha,s,\delta,m}.$$
- If $C_{\alpha,s,\delta,m}$ is empty, set
$$k_s^-(m) = k_s^+(m) := \lfloor \alpha^m \rfloor.$$

As in [5, 17], the dependence of $k_s^\pm(m)$ on δ and α is suppressed in the notation.

We first consider the case when $\{X_n\}$ are nonnegative.

Theorem 3.1. *Suppose $\{X_n\}$ is a sequence of nonnegative, pairwise independent random variables satisfying*

$$\sum_{k=1}^{\infty} \frac{\mathrm{Var}(X_k)}{k^2} \leqslant V$$

for some $V > 1$. Then, for any $R > 1$, the sequence $\{A_n - z_n/n\}$ has an R-learnable rate of uniform convergence to 0 given by

$$e(\lambda, \varepsilon) := \frac{\kappa_1 V \rho^3}{\varepsilon^5 \lambda},$$

where $\kappa_1 \leqslant 6000$ is a universal constant, and

$$\rho := \max\left\{1, \frac{1}{3(R-1)}, W\right\}.$$

Proof. From the proof of Theorem 1 in [5], we have for all $\delta > 0$, $\alpha > 1$, and $0 \leqslant s \leqslant L_\delta$:

$$\sum_{n=0}^{\infty} \mathbb{E}\left(\frac{\left(S_{k_s^\pm(n)} - z_{k_s^\pm(n)}\right)^2}{(k_s^\pm(n))^2}\right) \leqslant \frac{\alpha^2}{\alpha^2-1}\sum_{k=1}^{\infty}\frac{\mathrm{Var}(X_k)}{k^2} \leqslant \frac{\alpha^2}{\alpha^2-1}V.$$

Applying Chebyshev's inequality gives, for all $\varepsilon > 0$:

(3) $$\sum_{n=0}^{\infty}\mathbb{P}\left(\left|\frac{S_{k_s^\pm(n)}}{k_s^\pm(n)} - \frac{z_{k_s^\pm(n)}}{k_s^\pm(n)}\right| > \varepsilon\right) \leqslant \frac{\alpha^2 V}{\varepsilon^2(\alpha^2-1)}.$$

By Lemma 2.3, for any sequence $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \dots$ and $\lambda \in (0,1]$, there exists

$$m^\pm \leqslant \frac{2\alpha^2 V}{\varepsilon^2\lambda(\alpha^2-1)} + 1$$

such that

$$\sum_{n=a_{m^\pm}}^{b_{m^\pm}}\mathbb{P}\left(\left|\frac{S_{k_s^\pm(n)}}{k_s^\pm(n)} - \frac{z_{k_s^\pm(n)}}{k_s^\pm(n)}\right| > \varepsilon\right) \leqslant \lambda.$$

Consequently,

$$\mathbb{P}\left(\exists n \in [a_{m^\pm}, b_{m^\pm}]\ \left(\left|\frac{S_{k_s^\pm(n)}}{k_s^\pm(n)} - \frac{z_{k_s^\pm(n)}}{k_s^\pm(n)}\right| > \varepsilon\right)\right) \leqslant \lambda.$$

Define measurable formulas

$$A^+_{\varepsilon,s}(\omega,n) := \left|\frac{S_{k_s^+(n)}(\omega)}{k_s^+(n)} - \frac{z_{k_s^+(n)}}{k_s^+(n)}\right| > \varepsilon, \quad A^-_{\varepsilon,s}(\omega,n) := \left|\frac{S_{k_s^-(n)}(\omega)}{k_s^-(n)} - \frac{z_{k_s^-(n)}}{k_s^-(n)}\right| > \varepsilon.$$

Applying Lemma 2.11 gives

$$m \leqslant \frac{4\alpha^2 V}{\varepsilon^2\lambda(\alpha^2-1)} + 2 =: \phi_\alpha(\lambda,\varepsilon)$$

such that

$$\mathbb{P}\Big(\exists n \in [a_m, b_m]\ \left(A^+_{\varepsilon,s}(n) \cup A^-_{\varepsilon,s}(n)\right)\Big) \leqslant \lambda.$$

Now, let $\varepsilon, \lambda \in (0,1]$ and $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \dots$ with $b_n/a_n > R$ be given. Set $\delta := \varepsilon/3$ and $\alpha := 1 + \varepsilon/(3\rho)$, so that

(4) $$(\alpha-1)\rho = \varepsilon/3, \quad -\varepsilon/3 \leqslant -(1-1/\alpha)\rho.$$

Define $L(n) := \lfloor \log_\alpha n \rfloor$. Since $b_n/a_n > R \geqslant 1 + 1/(3\rho) \geqslant \alpha$, we have $L(b_n) > L(a_n)$ and hence $L(a_0) < L(b_0) \leqslant L(a_1) < L(b_1) \leqslant \dots$.

Now, from our previous discussion, for each $s \in [0, L_\delta]$ the formula

$$A^+_{\varepsilon,s}(\omega,n) \cup A^-_{\varepsilon,s}(\omega,n)$$

has an R-learnable uniform rate given by ϕ_α and so, by applying Lemma 2.11, we can obtain a R-learnable uniform rate for the formula

$$\bigvee_{s=0}^{L_\delta} A^+_{\varepsilon/3\alpha,s}(\omega,n) \cup A^-_{\varepsilon/3\alpha,s}(\omega,n)$$

which would give an

$$m \leqslant (L_\delta+1)\phi_\alpha\left(\frac{\lambda}{L_\delta+1}, \frac{\varepsilon}{3\alpha}\right)$$

such that

$$\mathbb{P}\left(\exists p \in [L(a_m); L(b_m)]\, \exists s \in [0, L_\delta]\, (A^+_{\varepsilon/3\alpha,s}(p) \cup A^-_{\varepsilon/3\alpha,s}(p))\right) \leqslant \lambda.$$

For this choice of m, we show

$$\exists n \in [a_m, b_m] \left|\frac{S_n}{n} - \frac{z_n}{n}\right| > \varepsilon$$

is contained in

$$\exists p \in [L(a_m), L(b_m)]\, \exists s \in [0, L_\delta]\, (A^+_{\varepsilon/(3\alpha),s}(p) \cup A^-_{\varepsilon/(3\alpha),s}(p)).$$

Hence, the result follows after simplification, using the facts that

$$\alpha \leqslant 4/3, \quad \alpha^2 - 1 = (\alpha-1)(\alpha+1) \geqslant \varepsilon/3\rho, \quad L_\delta + 1 \leqslant 4W/\varepsilon.$$

Suppose we take $n \in [a_m, b_m]$ such that

$$\left|\frac{S_n}{n} - \frac{z_n}{n}\right| > \varepsilon. \tag{5}$$

Then $p := L(n) \in [L(a_m), L(b_m)]$ and we can find $0 \leqslant s \leqslant L_\delta$ such that

$$\frac{1}{n} z_n \in [s\delta, (s+1)\delta).$$

Now, taking $p \in \mathbb{N}$ such that $\alpha^p \leqslant n < \alpha^{p+1}$, ensures that $n \in C_{\alpha,s,\delta,p}$. Now, since $k^\pm_s(p) \in C_{\alpha,s,\delta,p}$, we have

$$\frac{1}{k^\pm_s(p)} z_{k^\pm_r(p)} \in [s\delta, (s+1)\delta)$$

which implies

$$\left|\frac{1}{n} z_n - \frac{1}{k^\pm_s(p)} z_{k^\pm_s(p)}\right| \leqslant \delta.$$

Now, following the exact same reasoning as [5, 17], we have

$$\begin{aligned}
&-\delta - \left(1 - \frac{1}{\alpha}\right)\rho + \frac{1}{\alpha}\frac{1}{k^-_s(p)}\left(S_{k^-_s(p)} - z_{k^-_s(p)}\right) \\
&\quad \leqslant -\delta - \left(1 - \frac{1}{\alpha}\right)\frac{1}{k^-_s(p)} z_{k^-_s(p)} + \frac{1}{\alpha}\frac{1}{k^-_s(p)}\left(S_{k^-_s(p)} - z_{k^-_s(p)}\right) \\
&\quad \leqslant \frac{1}{n} S_{k^-_s(p)} - \frac{1}{n} z_n \\
&\quad \leqslant \frac{1}{n}(S_n - z_n) \\
&\quad \leqslant \frac{1}{n} S_{k^+_s(p)} - \frac{1}{k^+_s(p)} z_{k^+_s(p)} + \delta \\
&\quad \leqslant \frac{\alpha}{k^+_s(p)}\left(S_{k^+_s(p)} - z_{k^+_s(p)}\right) + (\alpha-1)\rho + \delta.
\end{aligned}$$

So, (4) implies that (recalling that $\delta = \varepsilon/3$),

$$\begin{aligned}
-\frac{2\varepsilon}{3} + \frac{1}{\alpha}\frac{1}{k^-_s(p)}\left(S_{k^-_s(p)} - z_{k^-_s(p)}\right) &\leqslant \frac{1}{m}(S_n - z_n) \\
&\leqslant \frac{\alpha}{k^+_s(p)}\left(S_{k^+_s(p)} - z_{k^+_s(p)}\right) + \frac{2\varepsilon}{3}.
\end{aligned} \tag{6}$$

So, it must be the case that

$$(A^+_{\varepsilon/3\alpha,s}(p) \cup A^-_{\varepsilon/3\alpha,s}(p))$$

as if not, we have

$$\left|\frac{1}{k_s^\pm(p)}\left(S_{k_s^\pm(p)} - z_{k_s^\pm(p)}\right)\right| \leqslant \frac{\varepsilon}{3\alpha}$$

and so (6) implies

$$\left|\frac{S_n}{n} - \frac{z_n}{n}\right| \leqslant \varepsilon$$

which contradicts (5). □

We can now obtain our results without the assumption that $\{X_n\}$ are nonnegative.

Theorem 3.2. *Suppose $\{X_n\}$ satisfy $\mathbb{E}(X_n) = 0$ for all $n \in \mathbb{N}$ and*

$$\sum_{k=1}^{\infty} \frac{\mathrm{Var}(X_k)}{k^2} \leqslant V$$

for some $V > 1$. For all $R > 1$, $\{A_n\}$ has a R-learnable rate of uniform convergence to 0 given by

$$e(\lambda, \varepsilon) := \frac{\kappa_2 V \rho^3}{\varepsilon^5 \lambda}$$

for a universal constant $\kappa_2 \leqslant 2^7 \times \kappa_1$, where,

$$\rho := \max\left\{1, \frac{1}{3(R-1)}, \frac{W}{2}\right\}.$$

Proof. For all $n \in \mathbb{N}$, we have $\mathrm{Var}(X_n) = \mathbb{E}(X_n^2) \geqslant \mathrm{Var}(X_n^+) + \mathrm{Var}(X_n^-) \geqslant \mathrm{Var}(X_n^\pm)$. Thus, we have

$$\sum_{k=1}^{\infty} \frac{\mathrm{Var}(X_k^\pm)}{k^2} \leqslant V.$$

Furthermore, $\mathbb{E}(X_n^+) = \mathbb{E}(X_n^-)$ since $\mathbb{E}(X_n) = 0$ and so $\mathbb{E}(|X_n|) = \mathbb{E}(X_n^+) + \mathbb{E}(X_n^-) = 2\mathbb{E}(X_n^\pm)$. Therefore, we have for all $n \in \mathbb{N}$

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(X_n^\pm) = \frac{z_n}{2n} \leqslant \frac{W}{2}.$$

From the above and the fact that $\{X_n^\pm\}$ are sequences of pairwise independent nonnegative random variables, we can apply Theorem 3.1 to obtain an R-learnable rate of uniform convergence to 0 for

$$\frac{1}{n}\sum_{k=1}^{n}(X_k^\pm - \mathbb{E}(X_k^\pm)).$$

is given by

$$e(\lambda, \varepsilon) := \frac{\kappa_1 V \rho^3}{\varepsilon^5 \lambda}.$$

The result then follows from Lemma 2.12, noting that

$$A_n = \frac{1}{n}\sum_{k=1}^{n}(X_k^+ - \mathbb{E}(X_k^+)) - \frac{1}{n}\sum_{k=1}^{n}(X_k^- - \mathbb{E}(X_k^-)).$$

and so $\{A_n\}$ will have a learnable rate of uniform convergence given by $2e(\lambda/2, \varepsilon/2)$. □

4. Proof of Theorem 1.5

We follow the proof strategy outlined in Section 1.2. First, let us assume we have a sequence of nonnegative pairwise independent random variables $\{X_n\}$ and a constant $\mu > 0$ with $\mathbb{E}(X_n) = \mu$ for all $n \in \mathbb{N}$. Setting

$$Y_n := X_n I\{X_n \leqslant n\} - \mathbb{E}(X_n I\{X_n \leqslant n\})$$

and

$$z_n := \sum_{k=1}^{n} \mathbb{E}(X_k I\{X_k \leqslant k\}),$$

we recall the following decomposition for A_n:

$$A_n = \frac{1}{n}\sum_{k=1}^{n} Y_k + \frac{z_n}{n} + \frac{1}{n}\sum_{k=1}^{n} X_k I\{|X_k| > k\}.$$

Hence, if we calculate respective R-learnable rates of uniform convergence ϕ_1, ϕ_2, ϕ_3 for

$$\frac{1}{n}\sum_{k=1}^{n} Y_k, \quad \frac{z_n}{n}, \quad \frac{1}{n}\sum_{k=1}^{n} X_k I\{|X_k| > k\},$$

then Lemma 2.12 implies that $\{A_n\}$ will have an R-learnable rate of uniform convergence given by

$$\phi(\lambda, \varepsilon) := \phi_1(\lambda/3, \varepsilon/3) + \phi_2(\lambda/3, \varepsilon/3) + \phi_3(\lambda/3, \varepsilon/3).$$

Now, $\{Y_n\}$ are pairwise independent with $\mathbb{E}(Y_n) = 0$ and for all $n \in \mathbb{N}$,

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(|Y_k|) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\big(|X_k I\{X_k \leqslant k\} - \mathbb{E}(X_k I\{X_k \leqslant k\})|\big) \leqslant 2\mu,$$

where we use the nonnegativity of $\{X_n\}$, so that in particular $\mathbb{E}(X_n I\{X_n \leqslant n\}) \leqslant \mu$. Furthermore, we have (c.f. Lemma 2.4.3 of [7])

$$\sum_{k=1}^{\infty} \frac{\mathrm{Var}(Y_k)}{k^2} \leqslant 4\mu.$$

Thus, Theorem 3.2 gives us an R-learnable rate of uniform convergence to 0. Applying Lemma 2.6, we obtain

$$\phi_1(\lambda, \varepsilon) := \frac{\kappa_3 \rho^4}{\varepsilon^5 \lambda},$$

where

$$\rho := \max\left\{1, \frac{1}{3(R-1)}, \mu\right\},$$

and $\kappa_3 \leqslant 2^7 \times \kappa_2$ is a universal constant.

We now follow Section 1.2 to calculate ϕ_2. Namely, $\{z_n/n\}$ is a monotone sequence of nonnegative real numbers bounded by μ, so that

$$\phi_2(\lambda, \varepsilon) = \frac{\mu}{\varepsilon}$$

by Lemma 2.2 (see also Remark 2.5).

The calculation of ϕ_3 will be the content of the following lemma.

Lemma 4.1. *The sequence*

$$\frac{1}{n}\sum_{k=1}^{n} X_k I\{X_k > k\}$$

has an R-learnable rate of uniform convergence given by

$$\phi_3(\lambda,\varepsilon) := \left(\frac{4\rho}{\lambda}+2\right)\left\lceil \log_R\left(\frac{4\rho}{\lambda\varepsilon}\right)\right\rceil$$

for all $R > 1$, where

$$\rho := \max\{1,\mu\}.$$

Proof. Fix $R > 1$. We show that

$$\psi(\lambda,\varepsilon) := \left(\frac{4\rho}{\lambda}+2\right)\left\lceil \log_R\left(\frac{2\rho}{\lambda\varepsilon}\right)\right\rceil$$

is an R-learnable rate of uniform convergence to 0, and then the result follows from Lemma 2.6.

First, observe that

$$\sum_{k=1}^{\infty}\mathbb{P}(X_k > k) = \sum_{k=1}^{\infty}\mathbb{P}(X > k) \leqslant \mu \leqslant \rho. \tag{7}$$

Let $\varepsilon,\lambda \in (0,1]$ and $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \dots$ be given such that $b_n/a_n > R$, and set

$$\beta := \left\lceil \log_R\left(\frac{2\rho}{\lambda\varepsilon}\right)\right\rceil.$$

Our choice of ρ ensures that $\beta \geqslant 1$, so

$$a_1 < b_\beta \leqslant a_{\beta+1} < b_{2\beta} \leqslant a_{2\beta+1} < b_{3\beta} \leqslant \dots$$

Thus, Lemma 2.3 applied to (7) implies there exists

$$m \leqslant \frac{4\rho}{\lambda}+1$$

such that

$$\mathbb{P}\Big(\exists p \in [a_{m\beta+1}; b_{(m+1)\beta}]\,(X_p > p)\Big) \leqslant \sum_{p=a_{m\beta+1}}^{b_{(m+1)\beta}}\mathbb{P}(X_p > p) \leqslant \frac{\lambda}{2}.$$

We claim that

$$n := (m+1)\beta \leqslant \left(\frac{4\rho}{\lambda}+2\right)\beta \leqslant \psi(\lambda,\varepsilon)$$

satisfies

$$\mathbb{P}\left(\exists l \in [a_n; b_n]\left(\frac{1}{l}\sum_{k=1}^{l} X_k I\{X_k > k\} > \varepsilon\right)\right) \leqslant \lambda,$$

and so the result follows.

Observe that $[a_n; b_n] \subseteq [a_{m\beta+1}; b_{(m+1)\beta}]$, and if

$$\exists l \in [a_n; b_n]\left(\frac{1}{l}\sum_{k=1}^{l} X_k I\{X_k > k\} > \varepsilon\right) \quad \text{and} \quad \forall p \in [a_{m\beta+1}; b_{(m+1)\beta}]\,(X_p \leqslant p),$$

then

$$\frac{1}{a_n}\sum_{k=1}^{a_{m\beta}} X_k I\{X_k > k\} > \varepsilon.$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(\exists l \in [a_n; b_n] \left(\frac{1}{l}\sum_{k=1}^{l} X_k I\{X_k > k\} > \varepsilon\right)\right)\\
&\leqslant \frac{\lambda}{2} + \mathbb{P}\left(\exists l \in [a_n; b_n] \left(\frac{1}{l}\sum_{k=1}^{l} X_k I\{X_k > k\} > \varepsilon \cap \forall p \in [a_{m\beta+1}; b_{(m+1)\beta}]\,(X_p \leqslant p)\right)\right)\\
&\leqslant \frac{\lambda}{2} + \mathbb{P}\left(\frac{1}{a_n}\sum_{k=1}^{a_{m\beta}} X_k I\{X_k > k\} > \varepsilon\right)\\
&\leqslant \frac{\lambda}{2} + \frac{1}{\varepsilon}\mathbb{E}\left(\frac{1}{a_n}\sum_{k=1}^{a_{m\beta}} X_k I\{X_k > k\}\right)\\
&\leqslant \frac{\lambda}{2} + \frac{a_{m\beta}\mu}{a_n\varepsilon}.
\end{aligned}$$

We use the fact that X_n are nonnegative and apply Markov's inequality to obtain the penultimate inequality in the above. Finally, since $b_n/a_n > R$, one can show by induction that for each $k \in \mathbb{Z}^+$

$$\frac{a_{m\beta}}{a_{m\beta+k}} < R^{-k}.$$

Setting $k = \beta$ yields,

$$\frac{a_{m\beta}}{a_n} = \frac{a_{m\beta}}{a_{(m+1)\beta}} < R^{-\beta},$$

so

$$\frac{a_{m\beta}\mu}{a_n\varepsilon} < \frac{\mu}{\varepsilon R^{\beta}} \leqslant \frac{\lambda}{2}$$

and the proof is complete. □

Therefore, for nonnegative $\{X_n\}$, A_n will have a R-learnable rate of uniform convergence given by

$$\phi(\lambda, \varepsilon) := 3^6 \cdot \frac{\kappa_3\rho^4}{\varepsilon^5\lambda} + \frac{3\mu}{\varepsilon} + \left(\frac{12\rho}{\lambda} + 2\right)\left\lceil \log_R\left(\frac{36\rho}{\lambda\varepsilon}\right)\right\rceil \leqslant \frac{\kappa_4\rho^4}{\varepsilon^5\lambda}$$

where

$$\rho := \max\left\{1, \frac{1}{3(R-1)}, \mu\right\}.$$

and $\kappa_4 \leqslant 3^6\kappa_3 + 3 + 14 \times 36$.

Now to prove Theorem 1.5, we note that $\{X_n^{\pm}\}$ are pairwise iid random variables with $\mathbb{E}(X_n^+) = \mathbb{E}(X_n^-) = \mu/2$ and so each have an R-learnable rate of uniform convergence given by ϕ above replacing μ by $\mu/2$ in the definition of ρ. The result then follows from Lemma 2.3 and simplification, since we can write $X_n = X_n^+ - X_n^-$. Furthermore, we will have $\kappa \leqslant 2^7\kappa_4 \leqslant 10^{11}$.

5. Oscillatory operators

In the introduction, we introduced the notion of a learnable rate of pointwise convergence. It is not hard to see that both learnable rates of uniform convergence and moduli of finite fluctuations are also learnable rates of pointwise convergence. Furthermore, given a modulus of finite fluctuations, one can obtain a learnable rate of uniform convergence by an argument due to Powell:

Proposition 5.1. *Suppose $\{X_n\}$ is a sequence of random variables with a modulus of finite fluctuations ϕ. Then $\{X_n\}$ has a learnable rate of uniform convergence given by*

$$\psi(\lambda) := \frac{2\phi(\lambda/2)}{\lambda}.$$

Proof. Fix $\lambda, \varepsilon \in (0, 1]$ and $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$. Define the event

$$B :\equiv J_\varepsilon\{X_n\} \geqslant \phi(\lambda/2).$$

So $\mathbb{P}(B) \leqslant \lambda/2$. To ease notation, write for $a, b \in \mathbb{N}$, write $C(a, b) :\equiv \exists i, j \in [a; b]\, |X_i - X_j| > \varepsilon$.

Now suppose, for contradiction, that for all $n \leqslant \psi(\lambda)$ we have $\mathbb{P}(C(a_n, b_n)) \geqslant \lambda$. Then, for all $n \leqslant \psi(\lambda)$ we must have

$$\mathbb{P}(C(a_n, b_n) \wedge B^c) = \mathbb{P}(C(a_n, b_n)) - \mathbb{P}(C(a_n, b_n) \wedge B) \geqslant \lambda - \mathbb{P}(B) \geqslant \frac{\lambda}{2}.$$

Now, it is clear that

$$\sum_{n \leqslant \psi(\lambda)} I\{C(a_n, b_n) \wedge B^c\} < \phi(\lambda/2).$$

Thus,

$$\frac{(\psi(\lambda) + 1)\lambda}{2} \leqslant \sum_{n \leqslant \psi(\lambda)} \mathbb{P}(C(a_n, b_n) \wedge B^c) = \mathbb{E}\left(\sum_{n \leqslant \psi(\lambda)} I\{C(a_n, b_n) \wedge B^c\}\right) < \phi(\lambda/2)$$

and we obtain a contradiction by examining the extremes of the above inequality. □

Remark 5.2. The fact that this bound is indeed optimal follows from Example 4.18 in [22]. The converse relationship is currently open, that is, to give an explicit bound on the modulus of finite fluctuations in terms of a given learnable rate of uniform convergence.

An adaptation of the argument in Proposition 5.1 to finitely additive probability spaces played a crucial role in [16].

Fluctuation bounds, such as Theorem 1.1, which correspond to moduli of finite fluctuations, appear throughout the ergodic theory literature. In [18, 22], almost sure convergence is studied from an abstract logical perspective, giving rise to more natural quantitative interpretations of stochastic convergence, such as learnable rates.

However, in ergodic theory, quantitative notions that are stronger than almost sure convergence have been developed and used to provide strengthenings to classical convergence results.

The first is based on so-called oscillation semi-norms, where for a sequence of real numbers $\{x_n\}$, an increasing sequence of natural numbers $\{n_k\}$ and $r > 1$ we define

$$\mathcal{O}^r_{\{n_k\}}(\{x_n\}) := \left(\sum_{k=1}^{\infty} \sup_{n_k \leqslant u \leqslant v < n_{k+1}} |x_u - x_v|^r\right)^{1/r}.$$

One can show that $\{x_n\}$ convergence if $\mathcal{O}^r_{\{n_k\}}(\{x_n\}) < \infty$ for all increasing sequences of natural numbers $\{n_k\}$ and $r > 1$.

The second is based on variation semi-norms, where we define

$$\mathcal{V}^r(\{x_n\}) := \sup_{\{n_k\}} \left(\sum_{k=1}^{\infty} |x_{n_{k+1}} - x_{n_k}|^r\right)^{1/r}$$

with the supremum taken over all increasing sequences of natural numbers $\{n_k\}$.

It is clear that for all increasing sequences of natural numbers $\{n_k\}$ and $r > 1$ we have $\mathcal{O}^r_{\{n_k\}}(\{x_n\}) \leqslant \mathcal{V}^r(\{x_n\})$, with the possibility that this inequality is indeed strict (see Remark 2.9 of [9]), thus the finiteness of the $\mathcal{V}$ operator is a stronger property that also entails convergence.

Strengthenings of ergodic theorems are given by weak type (1,1) inequalities on the oscillation and the variation semi-norms (we refer to [14] for a survey detailing the study of these operators with regard to polynomial ergodic averages), and such results easily give rise to the moduli arising from proof mining. For example, in [9, Theorem B] it is shown that for $r > 2$, $\mathcal{V}^r$ is weak type (1,1). That is, there exists $C > 0$ such that for all $f \in L_1$ and $a > 0$

$$\mathbb{P}\left(\sup_{\{n_k\}}\left(\sum_{k=1}^{\infty}|A_{n_{k+1}}f - A_{n_k}f|^r\right)^{1/r} \geqslant a\right) \leqslant \frac{C\|f\|_1}{a}.$$

From this, one can quickly calculate that a modulus of finite fluctuations is given by

$$e(\lambda,\varepsilon) = \left(\frac{C\|f\|_1}{\lambda\varepsilon}\right)^r$$

for all $r > 2$. Setting $r = 2$ in the above expression would yield Theorem 1.1. However, the above only holds for $r > 2$ and for all such r, this results in a worse bound.

In addition, in [9] it is shown that for all increasing sequences of natural numbers $\{n_k\}$ and $r > 1$ we have $\mathcal{O}^2_{\{n_k\}}$ is also weak type (1,1), with constant independent of $\{n_k\}$. That is, there exists $C > 0$ such that for all sequences of indices $n_1 < n_2 < \ldots$, $f \in L_1$, and $a > 0$

$$\mathbb{P}\left(\left(\sum_{k=1}^{\infty}\sup_{n_k\leqslant u\leqslant v<n_{k+1}}|A_uf - A_vf|^2\right)^{1/2} \geqslant a\right) \leqslant \frac{C\|f\|_1}{a}.$$

Such a bound does not naturally give moduli of finite fluctuations. However, one can obtain a learnable rate of pointwise convergence, as follows:

Suppose $a_0 < b_0 \leqslant a_1 < b_1 \leqslant \ldots$ and ε are given. Then

$$\begin{aligned}&\mathbb{P}\left(\forall i \leqslant e\,\exists k, l \in [a_i, b_j]\,(|A_kf - A_lf| \geqslant \varepsilon)\right)\\ &\leqslant \mathbb{P}\left(\left(\sum_{k=1}^{\infty}\sup_{n_k\leqslant u\leqslant v<n_{k+1}}|A_uf - A_vf|^2\right)^{1/2} \geqslant \varepsilon\sqrt{\frac{e}{3}}\right) \leqslant \frac{C\|f\|_1}{a},\end{aligned}$$

where $n_{2k+1} = a_{3k}$ and $n_{2k+2} := b_{3k} + 1$. This yields that,

$$e(\lambda,\varepsilon) := \frac{3C^2\|f\|_1^2}{\varepsilon^2\lambda^2}$$

is a learnable rate of pointwise convergence. The way the authors establish the weak type (1,1) inequality for $\mathcal{O}^2_{\{n_k\}}$ is of particular interest to us. Their strategy proceeds in two steps. First, they prove a strong type (2,2) inequality, demonstrating that there exists $C > 0$ such that for all increasing sequences of indices $\{n_k\}$ and $f \in L_2$

$$||\mathcal{O}^2_{\{n_k\}}(\{A_nf\})||_2 \leqslant C||f||_2. \tag{8}$$

They then invoke an abstract transfer principle based on the Calderón-Zygmund decomposition (c.f. [9, Theorem 3.1]) to pass from the strong type (2,2) inequality to the desired weak type (1,1) bound.

It had already been shown by several authors [4, 9, 23, 29] that (8) held if the sequence of indices $\{n_k\}$ was lacunary, that is, when there exists $R > 1$ such that $n_{i+1}/n_i > R$. In this case, the constant C will depend on R. Our main result, Theorem 1.5, provides uniform bounds under a similar lacunary condition.

This naturally raises two further questions. First, can one obtain bounds in Theorem 1.5 without imposing lacunarity? Second, what structural properties do the oscillation and variation operators $\mathcal{O}$ and $\mathcal{V}$ exhibit when applied to the sample means of pairwise independent, identically distributed random variables? More precisely:

Problem 5.3. If $\{X_n\}$ are pairwise iid random variables and $\{A_n\}$ are their sample means, is it the case that for all increasing sequences of natural numbers $\{n_k\}$ and $r > 1$ we have:

$$\mathcal{O}^r_{\{n_k\}}(\{A_n\}) < \infty \text{ and } \mathcal{V}^r(\{A_n\}) < \infty$$

almost surely? Furthermore, can one establish weak type and strong type inequalities?

Acknowledgements. We thank Thomas Powell for allowing us to include the proof of Proposition 5.1.

References

[1] Arthan, R., and Oliva, P. On the Borel-Cantelli Lemmas, the Erdős-Rényi Theorem, and the Kochen-Stone Theorem. *Journal of Logic and Analysis 13* (2021).

[2] Avigad, J., Dean, E., and Rute, J. A metastable dominated convergence theorem. *Journal of Logic and Analysis* (2012).

[3] Avigad, J., Gerhardy, P., and Towsner, H. Local stability of ergodic averages. *Transactions of the American Mathematical Society 362*, 1 (2010), 261–288.

[4] Bourgain, J. On the maximal ergodic theorem for certain subsets of the integers. *Israel Journal of Mathematics 61*, 1 (1988), 39–72.

[5] Csörgő, S., Tandori, K., and Totik, V. On the strong law of large numbers for pairwise independent random variables. *Acta Mathematica Hungarica 42* (1983), 319–330.

[6] Doob, J. L. *Stochastic Processes*. New York: Wiley, 1953.

[7] Durrett, R. *Probability: theory and examples*, vol. 49. Cambridge university press, 2019.

[8] Hochman, M. Upcrossing inequalities for stationary sequences and applications. *Annals of Probability 37*, 6 (2009), 2135–2149.

[9] Jones, R., Kaufman, R., Rosenblatt, J., and Wierdl, M. Oscillation in ergodic theory. *Ergodic Theory and Dynamical Systems 18*, 4 (1998), 889–935.

[10] Jones, R. L., Rosenblatt, J. M., and Wierdl, M. Counting in ergodic theory. *Canadian Journal of Mathematics 51*, 5 (1999), 996–1019.

[11] Kachurovskii, A. The rate of convergence in ergodic theorems. *Russian Mathematical Surveys 51*, 4 (1996), 653–703.

[12] Kalikow, S., and Weiss, B. Fluctuations of ergodic averages. *Illinois Journal of Mathematics 43*, 3 (1999), 480–488.

[13] Kohlenbach, U. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Monographs in Mathematics. Springer, 2008.

[14] Mirek, M., Szarek, T. Z., and Wright, J. Oscillation inequalities in ergodic theory and analysis: one-parameter and multi-parameter perspectives. *Revista matemática iberoamericana 38*, 7 (2022), 2249–2284.

[15] Neri, M. A finitary Kronecker's lemma and large deviations in the Strong Law of Large Numbers on Banach spaces. *Annals of Pure and Applied Logic 176*, 6 (2025), 103569.

[16] Neri, M. The pointwise ergodic theorem on finitely additive spaces. Preprint, available at `https://arxiv.org/abs/2511.01676`, 2025.

[17] Neri, M. Quantitative strong laws of large numbers. *Electronic Journal of Probability 30* (2025), 1–22.

[18] Neri, M., and Pischke, N. Proof mining and probability theory. *Forum of Mathematics, Sigma* (2025+). To appear.

[19] Neri, M., Pischke, N., and Powell, T. Generalized learnability of stochastic principles. In *Proceedings of Computability in Europe (CiE 2025)* (2025), Lecture Notes in Computer Science, Springer.

[20] Neri, M., Pischke, N., and Powell, T. On the asymptotic behaviour of stochastic processes, with applications to supermartingale convergence, Dvoretzky's approximation theorem, and stochastic quasi-Fejér monotonicity. Preprint, available at `https://arxiv.org/abs/2504.12922`, 2025.

[21] Neri, M., and Powell, T. A quantitative Robbins-Siegmund theorem. *Annals of Applied Probability* (2025). To appear.

[22] Neri, M., and Powell, T. On quantitative convergence for stochastic processes: Crossings, fluctuations and martingales. *Transactions of the American Mathematical Society, Series B 12* (2025), 974–1019.

[23] Petersen, K. E., and Salama, I. A. *Ergodic theory and its connection with harmonic analysis: proceedings of the 1993 Alexandria conference*, vol. 205. Cambridge university press, 1995.

[24] Pischke, N., and Powell, T. Asymptotic regularity of a generalised stochastic Halpern scheme with applications. Preprint, available at `https://arxiv.org/abs/2411.04845`, 2024.

[25] Powell, T., and Wan, A. An approximate zero-one law via the dialectica interpretation. Preprint, available at `https://arxiv.org/abs/2508.20849`, 2025.

[26] Robbins, H., and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*. Elsevier, 1971, pp. 233–257.

[27] Tao, T. Soft analysis, hard analysis, and the finite convergence principle. Essay posted 23 May 2007, 2007. Appeared in: ‘T. Tao, Structure and Randomness: Pages from Year One of a Mathematical Blog. AMS, 298pp., 2008’.

[28] Tao, T. Norm convergence of multiple ergodic averages for commuting transformations. *Ergodic Theory and Dynamical Systems 28* (2008), 657–688.

[29] Wierdl, M. Pointwise ergodic theorem along the prime numbers. *Israel Journal of Mathematics 64*, 3 (1988), 315–336.