Generalized learnability of stochastic principles

Morenikeji Neri[1], Nicholas Pischke[2], and Thomas Powell[1]

[1] Department of Computer Science, University of Bath, United Kingdom
[2] Department of Mathematics, Technische Universität Darmstadt, Germany
{mn728,trjp20}@bath.ac.uk, pischke@mathematik.tu-darmstadt.de

Abstract. Motivated by recent applications of proof theory in probability, we introduce a novel computational interpretation of probabilistic ∃∀-formulas, called dependent learnability. This encompasses several important notions of quantitative stochastic convergence, where it represents a generalized version of the property – widely studied in probability and ergodic theory – that a sequence of random variables has bounded fluctuations. We study both deterministic and stochastic variants of this notion and relate these to other computational interpretations of ∃∀-formulas from the literature. In particular, we prove dependent learnability to be primitive recursively equivalent to the influential notion of metastability, which in conjunction with results from applied proof theory highlights that dependently learnable rates can be extracted from large classes of nonconstructive proofs of ∃∀-formulas. Furthermore, we present a primitive recursive algorithm for joining two (and thus finitely many) dependently learnable rates, which in particular proves to be considerably more mathematically intuitive than the corresponding functional for joining rates of metastability. Finally, we discuss our results in the light of game semantics.

Keywords: Quantitative convergence · Proof theory · Computability in probability theory

1 Introduction

Results on controlling the oscillation behaviour of a sequence of random variables are central to probability theory. One of the most well-known examples of such results is Doob's upcrossing inequality for supermartingales, asserting that the number of times an L_1-bounded supermartingale upcrosses some fixed interval $[\alpha, \beta]$ for $\alpha < \beta$ is bounded in mean, which in turn is used to establish the fundamental result that L_1-bounded supermartingales converge almost surely.

Doob's upcrossing inequalities, and similar results in general, further gain significance because they provide explicit quantitative convergence information

The first and third authors were partially funded by the EPSRC (grant numbers EP/L016540/1 and EP/W035847/1 respectively). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC-BY) licence to any Author Accepted Manuscript version arising.

in a situation where *direct, computable* rates of convergence generally do not exist (for the simple reason that supermartingales immediately inherit the Specker phenomena that apply to monotone bounded sequences of reals [23]). Indeed, such quantitative results on the oscillatory behaviour of stochastic processes abound in the probability literature. In particular, there are further inequalities that express oscillatory behavior in martingale theory [5, 7], as well as in the convergence of ergodic averages [4, 10, 11], along with results from information theory [8], just to name a few.

The computability theory of convergence has been long studied by logicians. Confronted with a convergence theorem for which no computable rates exist, there are several options available. One is to follow the approach of computable analysis and rebuild the underlying notions and the theory in a computable way (as done for measure theory e.g. in [9]). Another is to work in a general mathematical setting but to adjust what is meant by computable convergence, with the benefit of remaining compatible with the usual mathematical practice.

It is the second approach that is adopted in the proof mining program, a general methodology developed by Kohlenbach and his collaborators (see [13] for a comprehensive monograph on the subject and [14] for a recent survey) which applies ideas and techniques from proof theory to extract quantitative information from nonconstructive proofs in mathematics, substantiated and (re-)enforced by underlying proof-theoretic results known as general logical metatheorems.

In the context of this second approach, it has been shown that in the *non-stochastic* setting, there is in fact a hierarchy of quantitative effective data that one can associate to a convergent sequence, classified by the amount of classical logic permissible in proofs from which these data can be potentially extracted [15]. Concretely, this hierarchy is populated with computable rates of convergence at the top (which in general can only be extracted from semi-constructive proofs) and so-called rates of metastability at the bottom (which can be extracted in very general situations, in particular in the presence of full classical logic). Intermediate to these lies the property of a sequence having a computable bound on the number of *fluctuations* it experiences (which is extractable in contexts with a restricted use of the law of excluded middle).

It has been recently shown by the first and third author [19] that a natural lifting of this property to the stochastic setting – called uniform learnability – is enjoyed by large family of stochastic processes whose oscillatory behaviour is bounded in mean, including L_1-bounded supermartingales. To be more precise, for such processes $\{X_n\}$ it is possible to construct a computable function $\phi : (\mathbb{Q}^+)^2 \to \mathbb{N}$ such that for any $\varepsilon, \lambda \in \mathbb{Q}^+$ and sequences $a_0 < b_0 \leq a_1 < b_1 \leq \dots$ of natural numbers, there exists some $n \leq \phi(\varepsilon, \lambda)$ such that

$$\mathbb{P}\left(\exists i, j \in [a_n; b_n] \ \ (|X_i - X_j| \geq \varepsilon)\right) < \lambda,$$

or in other words, the probability of an ε-fluctuation occurring in the interval $[a_n; b_n]$ is very low. However, there are many convergent stochastic processes and algorithms that do not posses natural upcrossing inequalities in mean, and the question then arises: What kind of computational information can we hope to provide in those cases, i.e. what mathematically natural notion populates the bottom of an analogous hierarchy of quantitative effective data that one can associate to stochastic processes?

In this paper, we answer that question by introducing a new concept of computation for probabilistic formulas, which we call *dependent learnability*. This is defined abstractly in terms of measurable sets, but applies in particular to the convergence of stochastic processes, where we obtain a strict generalization of uniform learnability. We prove that dependent learnability is in fact equivalent to the natural lifting(s) of the metastability property in the probabilistic setting, thereby demonstrating that rates of dependent learnability can be extracted from rather general proofs (that make use of full classical logic and comprehension principles), and thus in principle can be found for any stochastic process whose convergence is provable in a theory amenable to proof mining.

We then provide a "joining theorem", which sets out a primitive recursive algorithm for combining a finite number of rates of dependent learnability. Informally speaking, this allows us to combine individual quantitative bounds for properties $A_1, \ldots, A_n$ into bounds for their conjunction $A_1 \wedge \ldots \wedge A_n$, which in the concrete setting of convergence enables us, for example, to combine quantitative fluctuation information about two individual stochastic processes into fluctuation information for their sum or product. In the simple case of uniform learnability this recently proved crucial for obtaining rates for almost supermartingales [20], and such combinations naturally feature in many other proof mining case studies. Our construction is extremely general and is intended to form a useful tool for future case studies of proof mining in probability, in addition to being of interest in its own right.

Finally, we discuss the logical and semantic aspects of our work. We show that dependent learnability has an elegant reading in terms of game semantics, as a winning strategy in a two-player probabilistic game. This interpretation extends to our joining constructions, highlighting that dependent learnability is a mathematically natural and intuitive concept, and providing insight into programs extracted from proofs in probability.

All of our main definitions and results can be formulated within the logical framework for probability developed by the first two authors [18]. In particular, we only make use of finite unions and intersections of events, giving a logical explanation of the uniformities underlying our constructions.

2 Fluctuations and generalized learnability

In this section, we now survey the well-known notion of fluctuations for sequences of real numbers as well as the related notion of a learnable rate of convergence, which we take as motivation to introduce our new generalized notion of learnability, both in a deterministic and a probabilistic variant, which are then compared to other such computational interpretations of (probabilistic) principles.

2.1 The deterministic setting

A sequence of real numbers $\{x_n\}$ is said to be *(Cauchy) convergent* if

$$\forall \varepsilon > 0\ \exists n \in \mathbb{N}\ \forall i, j \geq n\ (|x_i - x_j| < \varepsilon).$$

Results establishing the convergence of sequences of real numbers are fundamental to analysis, and obtaining computational information on the asymptotic behavior of such sequences and their convergence from such theorems is of great interest and importance. The most direct computational interpretation one can give here is a *rate of convergence*, which is a function $r : \mathbb{Q}^+ \to \mathbb{N}$ such that:

$$\forall \varepsilon \in \mathbb{Q}^+\ \forall i, j \geq r(\varepsilon)\ (|x_i - x_j| < \varepsilon).$$

However, it is well known that there are many convergence results for which computable rates do not exist. Thus, one is tasked to seek other computational interpretations for the convergence of sequences.

A natural interpretation which has found favour amongst proof-theorists [2, 15, 19] as well as probability theorists [5, 11, 12] is that of a bound on the number of fluctuations, namely a function $\phi : \mathbb{Q}^+ \to \mathbb{N}$ such that for any $\varepsilon \in \mathbb{Q}^+$ and any $i_0 < j_0 \leq i_1 < j_1 \leq \ldots \leq i_{k-1} < j_{k-1}$ with $|x_{i_n} - x_{j_n}| \geq \varepsilon$ for all $n = 0, \ldots, k-1$, it holds that $k \leq \phi(\varepsilon)$.

Fluctuation bounds offer an intuitive computational interpretation for many theorems for which one cannot obtain a direct computable rate of convergence. The prime example here is the case of a monotone sequence $\{x_n\} \subseteq [0, 1]$, where one can easily show that a fluctuation bound is given by $\phi(\varepsilon) := \lceil 1/\varepsilon \rceil$, but following the fundamental results of Specker [23], one can construct a computable monotone sequence of rational numbers in $[0, 1]$ with no computable rate of convergence. In the stochastic setting, it is supermartingales and related classes of process that represent the canonical examples of sequences with controlled fluctuation behaviour, as we will discuss further in Section 2.2 below.

We now present fluctuations formally, lifting them from the specific setting of convergent sequences and defining them on arbitrary $\exists\forall$-formulas. We then show that the closely related concept of *learnability* represents an alternative way of capturing bounded fluctuations.

Definition 1 (Fluctuations for formulas). *For an arbitrary formula $A(i, j)$*[3] *on pairs of natural numbers, and $N \in \mathbb{N}$, we define $J_{N,A}$ to be the maximal $k \in \mathbb{N}$ for which there exist*

$$a_0 < b_0 \leq a_1 < b_1 \leq \ldots \leq a_{k-1} < b_{k-1} < N$$

such that for all $n = 0, \ldots, k-1$, there exist[4] *$i, j \in [a_n; b_n]$ such that $A(i, j)$ holds. We write*

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

[3] These formulas could represent formal statements in an axiomatic theory, but for the purpose of this paper we just consider them as arbitrary properties of pairs of natural numbers, i.e. as subsets of $\mathbb{N} \times \mathbb{N}$.

[4] Here, and in the following, we write $[k; l] := [k, l] \cap \mathbb{N}$.

where this quantity could be infinite. We say that the fluctuations are bounded if $J_{\infty,A} \leq e$ *for some* $e \in \mathbb{N}$*, which we call a fluctuation bound.*

Here we imagine $A(i, j)$ as representing a "fluctuation" from i to j. While we suppress it in this paper, the formula A could naturally depend on outside parameters. In that vein, we recover the previous notion of a sequence of real numbers $\{x_n\}$ having bounded fluctuations: Setting $A_\varepsilon(i, j) :\equiv |x_i - x_j| \geq \varepsilon$ for $\varepsilon \in \mathbb{Q}^+$, a function ϕ is a fluctuation bound for $\{x_n\}$ if, and only if, we have

$$J_{\infty,A_\varepsilon} \leq \phi(\varepsilon)$$

for all $\varepsilon \in \mathbb{Q}^+$. An alternative and particularly intuitive way of presenting fluctuation bounds (and so in particular on the number of ε-fluctuations of the sequence $\{x_n\}$), is the following, which was first isolated in this context in [20]:

Definition 2 (Learnability of formulas). *A formula* $\exists n \forall i, j \geq n\, A(i, j)$ *is learnable if there exists a number* $e \in \mathbb{N}$ *such that for all sequences of natural numbers* a, b *with* $a_0 < b_0 \leq a_1 < b_1 \leq \dots$*, we have:*

$$\exists n \leq e\ \forall i, j \in [a_n; b_n]\ A(i, j).$$

The number e *is called a* learnability bound *for* A*.*

The intuition behind the term "learnability" is that $a_0 < a_1 < \dots$ represent guesses of points for which $\forall i, j \geq a_n\, A(i, j)$, with $b_0 < b_1 < \dots$ counterexamples to those guesses. Whenever a guess fails, we learn from that guess and move on to the next. A learnability bound represents the maximum number of mind-changes required before a guess works. This semantic meaning is discussed in greater depth in Section 4, but for now we simply note that learnability represents a simple and intuitive alternative to the property of the (negated) formula having finite fluctuations, as it can be easily shown that

$$e \text{ is a learnability bound for } A \iff J_{\infty,\neg A} \leq e.$$

While fluctuations bounds allow one to give a computational interpretation to results where one cannot obtain direct rates, there are still cases of computable converging sequences for which one cannot obtain an effective fluctuation bound (see e.g. [15]). In such cases, inspired by the no counterexample interpretation of Kreisel [16, 17], one arrives at the computationally weaker concept of a *rate of metastability*, also discussed in more detail later on, the existence of which can be broadly substantiated through the perspective of proof mining, already for large classes of non-effective proofs in analysis (see also the discussion in Section 2.4 later on).

The following definition now offers a new computational interpretation for the convergence of sequences which do not possess effective learnable rates of convergence, which is mathematically different to the notion of metastability (albeit, as we will later show, being computationally and proof-theoretically equivalent) and in a way more closely represents this paradigm of learnability.

Definition 3 (Dependent learnability of formulas). *A formula* $\exists n \forall i, j \geq n\, A(i,j)$ *is dependently learnable if for all sequences of natural numbers* a, b *with* $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$, *we have:*

$$\exists n \in \mathbb{N}\ \forall i, j \in [a_n; b_n]\ A(i,j).$$

A function $e : \mathbb{N}^{\mathbb{N}} \times \mathbb{N}^{\mathbb{N}} \to \mathbb{N}$ *which bounds* $n \leq e(a,b)$ *in terms of* a, b *is called a* dependently learnable rate *for* A.

In the special case that e is a constant function, it is a learnability bound for A or equivalently a bound for $J_{\infty,\neg A}$ as before, and therefore we can view dependent learnability as a generalization of the property of having bounded fluctuations. Rephrased in terms of fluctuations, a formula is dependently learnable if, and only if, for any fixed sequences $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$, the length of fluctuation sequences with the property that each fluctuation is contained within a unique interval $[a_n; b_n]$ is bounded.

2.2 The stochastic setting

We now arrive at the first main contribution of the paper: A concept of generalized learnability in the stochastic setting, which in particular represents a broader notion of "good fluctuation behaviour" for stochastic processes.

Everything that follows now takes place over an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$. While not explicitly highlighted any further in this paper, all of the main definitions and theorems that follow will only involve finite unions and intersections of events. This means that technically we can restrict $\mathbb{P}$ to being a probability *content* (also called a charge) [3], that is a finitely additive function $\mathbb{P} : \mathcal{F} \to [0,1]$ over some algebra $\mathcal{F}$. This then explains our results in light of the recent logical metatheorem for probability [18].

In any case, we say that a logical formula $\varphi(\omega, x_1, \ldots, x_n)$, with parameters $x_1, \ldots, x_n$ and ω a variable taking values in Ω, is *measurable* if for all parameters $x_1, \ldots, x_n$, we have $\varphi(x_1, \ldots, x_n) := \{\omega \in \Omega : \varphi(\omega, x_1, \ldots, x_n)\} \in \mathcal{F}$. If $\varphi(n)$ is a measurable formula, with $n \in \mathbb{N}$, and $p, q \in \mathbb{N}$ we define $\neg\varphi := \varphi^c$ as well as

$$\exists n \in [p;q]\ \varphi(n) := \bigcup_{n \in [p;q]} \varphi(n) \quad \text{and} \quad \forall n \in [p;q]\ \varphi(n) := \bigcap_{n \in [p;q]} \varphi(n).$$

We now come straight to our definition of generalized stochastic learnability. As with many concepts from probability, a single definition in the nonstochastic setting (in this case Definition 3) splits into multiple stochastic notions, in this case *uniform* and *pointwise*, which we justify and explain in more detail below.

Definition 4 (Dependent learnability of measurable formulas). *Let* $A(i,j)$ *be a measurable formula. A function* $e : \mathbb{Q}^+ \times \mathbb{N}^{\mathbb{N}} \times \mathbb{N}^{\mathbb{N}} \to \mathbb{N}$ *is*

(a) a uniform dependent learnable rate for $\exists n \forall i, j \geq n\, A(i,j)$ *if*

$$\exists n \leq e(\lambda, a, b)\, (\mathbb{P}(\exists i, j \in [a_n; b_n]\ \neg A(i,j)) < \lambda)$$

for any $\lambda \in \mathbb{Q}^+$ *and* $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ *(and if* e *is independent of the sequences* a, b *then it is a uniform learnable rate for* A*),* *(b) a pointwise dependent learnable rate for* $\exists n \forall i, j \geq n\, A(i,j)$ *if*

$$\mathbb{P}(\forall n \leq e(\lambda, a, b)\ \exists i, j \in [a_n; b_n]\ \neg A(i,j)) < \lambda$$

for any $\lambda \in \mathbb{Q}^+$ *and* $a_0 < b_0 \leq a_1 < b_1 \leq \dots$ *(and if* e *is independent of the sequences* a, b *then it is a pointwise learnable rate for* A*).*

The simpler notions of pointwise and uniform learnability for arbitrary measurable formulas, which represent stochastic versions of Definition 2, were introduced in [19], and these were in turn inspired by the quantitative formulations of pointwise and uniform convergence of measurable functions in [1].

Analogously to the nonstochastic case, also pointwise and uniform learnability are intimately connected to the previously introduced notion of fluctuations. More precisely, for a measurable formula A, define $J_{\infty,A}$ just as in Definition 1, where it now becomes random variable. In [19, Theorem 3.6 (i)], it is shown that if $\mathbb{E}[J_{\infty,\neg A}] < K$, then $e(\lambda) := \lceil K/\lambda \rceil$ is a uniform learnable rate for $\exists n \forall i, j \geq n\, A(i,j)$, and in the case of stochastic processes, uniform learnable rates exist generally when those processes enjoy L_1-bounds on fluctuations or upcrossings (thus L_1-bounded sup- or supermartingales can be given uniform learnable rates cf. [19, Theorem 7.4]). Similarly, it is shown in [19, Theorem 3.6 (ii)] that if e is a rate of convergence for

$$\lim_{n \to \infty} \mathbb{P}\left(J_{\infty,\neg A} \geq n\right) = 0,$$

then e is also a pointwise learnable rate for $\exists n \forall i, j \geq n\, A(i,j)$.

However, not all convergent stochastic processes enjoy the kind of strong quantitative fluctuation behaviour from which we can expect to obtain simple pointwise or uniform rates, in the same way that not all deterministic sequences posses computable learnability rates. We now show that dependent learnability is a much broader notion.

2.3 The computational strength of dependent learnability

We prove that having a uniform/pointwise dependent learnable rate for A is *computationally* equivalent to having a uniform/pointwise *rate of metastability* for the statement that $\exists n \forall i, j \geq n\, A(i,j)$ holds almost surely. We thereby demonstrate that dependent learnable rates can be in principle extracted from general convergence proofs, and thus represent a kind of quantitative fluctuation behaviour that is enjoyed by *any* stochastic processes that converge almost surely, provably so within a system amenable to proof mining.

Metastability for stochastic processes was introduced (in the context of logic) by Avigad et al. in [1], and adapted to arbitrary $\exists\forall$-formulas in [19]. We briefly recall the key definitions as they relate to the notions presented here. We first observe that $\exists n \forall i, j \geq n\, A(i,j)$ holding with probability one can be equivalently formulated in terms of finite unions and intersections as

$$\forall \lambda \in \mathbb{Q}^+\ \exists n \in \mathbb{N}\ \forall m \in \mathbb{N}\left(\mathbb{P}\left(\exists i, j \in [n; n+m]\ \neg A(i,j)\right) < \lambda\right). \qquad (*)$$

As recently confirmed by the results of [18], in many situations we can reasonably expect that even when a proof of the above statement is nonconstructive (and so the above n cannot be computably witnessed), we can still obtain a witness for the monotone Dialectica interpretation of

$$\forall \lambda \in \mathbb{Q}^+ \; \neg\neg\exists n \in \mathbb{N} \; \forall m \in \mathbb{N} \, (\mathbb{P}\,(\exists i, j \in [n; n+m] \; \neg A(i,j)) < \lambda) \qquad (**)$$

which is a functional $\varphi : \mathbb{Q}^+ \times (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ satisfying

$$\forall \lambda \in \mathbb{Q}^+, g : \mathbb{N} \to \mathbb{N} \, \exists n \leq \varphi(\lambda, g) \, (\mathbb{P}\,(\exists i, j \in [n; n+g(n)] \neg A(i,j)) < \lambda) \, .$$

We call such a functional a *rate of uniform metastability* for $\exists n \forall i, j \geq n \, A(i,j)$. In many cases, notably when analysing convergence proofs that work in a pointwise way on elements ω of the underlying probability space, it is considerably easier to instead produce a functional φ satisfying the following apparently weaker condition

$$\forall \lambda \in \mathbb{Q}^+ \; \forall g : \mathbb{N} \to \mathbb{N} \, (\mathbb{P}\,(\forall n \leq \varphi(\lambda, g) \; \exists i, j \in [n; n+g(n)] \; \neg A(i,j)) < \lambda)$$

and we label such a functional a *rate of pointwise metastablity* for $\exists n \forall i, j \geq n \, A(i,j)$. It is immediate that a rate of uniform metastability is also a pointwise rate, and in fact, the two are computationally equivalent, where (as made explicit in [19, Theorem 3.2]) the route from the latter to the former uses a computational version of Egorov's theorem given in [1] and inspired by a construction of Tao [24, Theorem A.2], though with the apparent cost of a significant blowup in complexity (where the construction in [1] uses bar recursion).

In the main result of this subsection, we demonstrate that one can easily translate between uniform (resp. pointwise) dependent learnability and uniform (resp. pointwise) metastability, using the same construction in both cases. In particular, this means that the complexity difference between pointwise and uniform dependent learnability matches that between the two corresponding forms of metastability. This generalises a similar correspondence between nondependent learnability and metastability proven recently as [19, Lemma 3.5].

Theorem 1. *Suppose $A(i,j)$ is a measurable formula and e is a pointwise (resp. uniform) dependent learnable rate for $\exists n \forall i, j \geq n \, A(i,j)$. Then*

$$\varphi(\lambda, g) := a^g_{e(\lambda, a^g, b^g)} = \tilde{g}^{(e(\lambda, a^g, b^g))}(0)$$

is a rate of pointwise (resp. uniform) metastability for $\exists n \forall i, j \geq n \, A(i,j)$, with a^g, b^g defined by $a^g_i := \tilde{g}^{(i)}(0)$ and $b^g_i := \tilde{g}^{(i+1)}(0)$, where $\tilde{g}(n) := n + g(n) + 1$. Conversely, if φ is a rate of pointwise (resp. uniform) metastability for $\exists n \forall i, j \geq n \, A(i,j)$. Then

$$e(\lambda, a, b) := k_{a,b}(\varphi(\lambda, g_{a,b}))$$

is a rate of pointwise (resp. uniform) dependent learnability for $\exists n \forall i, j \geq n \, A(i,j)$, with $k_{a,b}, g_{a,b}$ defined by $g_{a,b}(n) := b_{k_{a,b}(n)} - n$ and $k_{a,b}(n) := \min\{i : n \leq a_i\}$.

Proof. We only consider the case of pointwise rates, the case of uniform rates follows analogously. To see the first part, assume that e is a pointwise dependent learnable rate for $\exists n \forall i, j \geq n\, A(i,j)$ and let $\lambda > 0$ and $g : \mathbb{N} \to \mathbb{N}$ be fixed. It is easy to see that a^g and b^g satisfy $a^g_0 < b^g_0 \leq a^g_1 < b^g_1 \leq \ldots$. We therefore have

$$\mathbb{P}(\forall i \leq e(\lambda, a^g, b^g)\ \exists k, l \in [a^g_i; b^g_i]\ \neg A(k,l)) < \lambda.$$

Let $\omega \in \Omega$ be such that

$$\exists i \leq e(\lambda, a^g, b^g)\ \forall k, l \in [a^g_i; b^g_i]\ A(\omega, k, l).$$

We then have $[a^g_i; b^g_i] = [a^g_i; \tilde{g}(a^g_i)] = [a^g_i; a^g_i + g(a^g_i) + 1] \supset [a^g_i; a^g_i + g(a^g_i)]$ and as a^g is monotone increasing, we have $a^g_i \leq a^g_{e(\lambda, a^g, b^g)} = \varphi(\lambda, g)$ so that we have shown

$$\exists n \leq \varphi(\lambda, g)\ \forall k, l \in [n; n + g(n)]\ A(\omega, k, l).$$

As ω was arbitrary, we in particular have

$$\begin{aligned}&\mathbb{P}(\forall n \leq \varphi(\lambda, g)\ \exists k, l \in [n; n + g(n)]\ \neg A(k,l)) \\ &\qquad \leq \mathbb{P}(\forall i \leq e(\lambda, a^g, b^g)\ \exists k, l \in [a^g_i; b^g_i]\ \neg A(k,l)) < \lambda.\end{aligned}$$

To see the second part, assume that φ is a rate of pointwise metastability for $\exists n \forall i, j \geq n\, A(i,j)$ and let $\lambda > 0$ as well as a, b with $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ be fixed. Note first that $k_{a,b}$ and $g_{a,b}$ are well-defined since a is strictly increasing and $n \leq a_{k_{a,b}(n)} < b_{k_{a,b}(n)}$, and thus in particular we have

$$\mathbb{P}(\forall n \leq \varphi(\lambda, g_{a,b})\ \exists k, l \in [n; n + g_{a,b}(n)]\ \neg A(k,l)) < \lambda.$$

Let $\omega \in \Omega$ be such that

$$\exists n \leq \varphi(\lambda, g_{a,b})\ \forall k, l \in [n; n + g_{a,b}(n)]\ A(\omega, k, l).$$

By definition we have $n \leq a_{k_{a,b}(n)}$ and $b_{k_{a,b}(n)} = n + g_{a,b}(n)$ and therefore $A(\omega, k, l)$ holds for all $k, l \in [a_{k_{a,b}(n)}; b_{k_{a,b}(n)}]$. As $k_{a,b}$ is monotone, we have $k_{a,b}(n) \leq k_{a,b}(\varphi(\lambda, g_{a,b})) = e(\lambda, a, b)$ so that we have shown

$$\exists i \leq e(\lambda, a, b)\ \forall k, l \in [a_i; b_i]\ A(\omega, k, l).$$

As ω was arbitrary, we again have

$$\begin{aligned}&\mathbb{P}(\forall i \leq e(\lambda, a, b)\ \exists k, l \in [a_i; b_i]\ \neg A(k,l)) \\ &\qquad \leq \mathbb{P}(\forall n \leq \varphi(\lambda, g_{a,b})\ \exists k, l \in [n; n + g_{a,b}(n)]\ \neg A(k,l)) < \lambda.\end{aligned}$$

This completes the proof.

Considering the one-point space immediately allows us to derive an analogous result for the deterministic case, showing the equivalence for dependent learnability as discussed in Definition 3 and rates of metastability for formulas $\exists n \forall i, j \geq n\, A(i,j)$ in the form of functionals $\varphi : (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ satisfying

$$\exists n \leq \varphi(g)\ \forall i, j \in [n; n + g(n)]\ A(i,j)$$

for all $g : \mathbb{N} \to \mathbb{N}$. More specifically, in one direction, a rate of metastability for $\exists n \forall i, j \geq n\, A(i,j)$ is given by $\varphi(g) := a^g_{e(a^g, b^g)}$, and conversely a rate of dependent learnability by $k_{a,b}(\varphi(g_{a,b}))$. Remaining this time in the stochastic setting but restricting our attention to *nondependent* learnability, a similar and simpler variant of the proof of Theorem 1 was given in [19] to show that $e : \mathbb{Q}^+ \to \mathbb{N}$ is a pointwise (resp. uniform) learnable rate for $\exists n \forall i, j \geq n\, A(i,j)$ if and only if $\varphi(\lambda, g) := \tilde{g}^{(e(\lambda))}(0)$ is a rate of pointwise (resp. uniform) metastability. By considering the one-point space as before, one regains a deterministic correspondence result from this which was previously obtained in [15].

2.4 Dependent learnability and program extraction

We conclude by making more precise the significance of dependent learnability from the perspective of program extraction, particularly in light of the recent logical approach to probability in [18]. Almost sure convergence of a stochastic process $\{X_n\}$ is equivalent to the following property (analogous to $(*)$ above):

$$\forall \varepsilon, \lambda > 0\ \exists n \in \mathbb{N}\ \forall m \in \mathbb{N}\, (\mathbb{P}\, (\exists i, j \in [n; n+m]\ \ (|X_i - X_j| \geq \varepsilon)) < \lambda)\,.$$

As shown in [18, Section 9], one can formally represent this property in a logical system for reasoning about probability contents, such that the usual program extraction mechanism of proof mining (arising typically by a combination of a negative translation with a monotone variant of Gödel's functional interpretation, see [13] for more details) guarantees the existence of a computable rate of uniform metastability for the convergence property, whenever the above is provable in that system. Theorem 1 then guarantees the existence of a computable uniform dependent learnable rate. A similar argument holds in the pointwise setting.

As such, unlike for ordinary learnable rates, which require strong assumptions on the fluctuation behaviour of stochastic processes, the existence of computable rates of dependent learnability relies only on a stochastic process being provably convergent in a system amenable to proof mining. The existence of a core formal system for probability theory [18] along with empirical evidence from recent case studies [19, 20] suggest that dependent learnable rates should therefore be broadly extractible from convergence proofs in core probability theory.

3 Joining learnable rates

We now move on to constructions on rates of dependent learnability, demonstrating how two dependent rates for formulas $\exists n \forall i, j \geq n\, A_1(i,j)$ and $\exists n \forall i, j \geq n\, A_2(i,j)$ respectively can be combined to obtain a dependent rate for $\exists n \forall i, j \geq n\ (A_1(i,j) \wedge A_2(i,j))$ (and hence iteratively combined to produce a dependent rate for finite conjunctions).

Constructions of this kind are crucial in proof mining. In the deterministic setting, the joining of rates of metastability are discussed in abstract terms in

[21], where it is connected to the finite double negation shift, with a concrete example of the deterministic joining presented in detail in e.g. [22], among many other examples. In the stochastic setting, the joining of uniform learnable rates was central in the recent analysis of the Robbins-Siegmund theorem [20], where multiple instances were required to arrive at a uniform rate of learnability for almost-supermartingales.

We now present a series of joining results covering all notions of learnability introduced in this paper. For a sequence of natural numbers a and $n \in \mathbb{N}$, let us write $a^{(n)} := (a_i^{(n)})$ for the sequence defined by $a_i^{(n)} := a_{n+i}$ for all $i \in \mathbb{N}$.

Theorem 2. *Suppose $A_1(i,j), A_2(i,j)$ are measurable formulas and let e_1 and e_2, respectively, be pointwise dependent learnable rates for $\exists n \forall i,j \geq n\, A_1(i,j)$ and $\exists n \forall i,j \geq n\, A_2(i,j)$. Then $\exists n \forall i,j \geq n\ (A_1(i,j) \wedge A_2(i,j))$ has a pointwise dependent learnable rate given by*

$$e(\lambda, a, b) := p_{2e_1(\lambda/2, a', b')+1}$$

where a', b' are defined by $a'_i := a_{p_{2i}}$, $b'_i := b_{p_{2i+1}}$ for all $i \in \mathbb{N}$, and with (p_i) defined inductively by

$$p_0 := 0 \text{ and } p_{i+1} := p_i + e_2(\lambda/2^{i+1}, a^{(p_i)}, b^{(p_i)}) + 1.$$

Proof. Suppose a, b are sequences of natural numbers satisfying $a_0 < b_0 \leq a_1 < b_1 \leq \dots$ and let $\lambda \in \mathbb{Q}^+$ be given. Set

$$Q := \{\omega \in \Omega : \exists k \leq e_1(\lambda/2, a', b')\ \forall i,j \in [a'_k; b'_k]\ A_1(\omega, i, j)\}.$$

From the fact that e_1 is a pointwise dependent learnable rate for $\exists n \forall i,j \geq n\, A_1(i,j)$, we have $\mathbb{P}(\neg Q) < \lambda/2$. For this, note in particular that a', b' satisfy

$$a'_i = a_{p_{2i}} < b_{p_{2i}} < b_{p_{2i+1}} = b'_i \leq a_{p_{2i+2}} = a'_{i+1},$$

for any $i \in \mathbb{N}$ as we have $p_i < p_{i+1}$ for any $i \in \mathbb{N}$. Further, for each $k \in \mathbb{N}$, set

$$R_k := \{\omega \in \Omega : \exists l \leq e_2(\lambda/2^{k+1}, a^{(p_k)}, b^{(p_k)})\ \forall i,j \in [a_l^{(p_k)}; b_l^{(p_k)}]\ A_2(\omega, i, j)\}.$$

From the fact that e_2 is a pointwise dependent learnable rate for $\exists n \forall i,j \geq n\, A_2(i,j)$, we get $\mathbb{P}(\neg R_k) < \lambda/2^{k+1}$ where, for this, we note that $a_0^{(p_k)} < b_0^{(p_k)} \leq a_1^{(p_k)} < b_1^{(p_k)} \leq \dots$ holds for any $k \in \mathbb{N}$.

Now, let

$$\omega \in P := Q \cap \bigcap_{k \leq e_1(\lambda/2, a', b')} R_{2k}.$$

As $\omega \in Q$, we can take $k \leq e_1(\lambda/2, a', b')$ such that

$$\forall i,j \in [a'_k; b'_k]\ A_1(\omega, i, j) \equiv \forall i,j \in [a_{p_{2k}}; b_{p_{2k+1}}]\ A_1(\omega, i, j).$$

As also $\omega \in R_{2k}$, we can take $l \leq e_2(\lambda/2^{2k+1}, a^{(p_{2k})}, b^{(p_{2k})})$ such that

$$\forall i,j \in [a_l^{(p_{2k})}; b_l^{(p_{2k})}]\ A_2(\omega, i, j) \equiv \forall i,j \in [a_{p_{2k}+l}; b_{p_{2k}+l}]\ A_2(\omega, i, j).$$

Now we have $a_{p_{2k}} \leq a_{p_{2k}+l}$ and $b_{p_{2k}+l} \leq b_{p_{2k}+e_2(\lambda/2^{2k+1},a^{(p_{2k})},b^{(p_{2k})})+1} = b_{p_{2k+1}}$. Thus, we also have $A_1(\omega,i,j)$ for all $i,j \in [a_{p_{2k}+l}, b_{p_{2k}+l}]$ and so combined with the above we have

$$\forall i,j \in [a_{p_{2k}+l}; b_{p_{2k}+l}]\ (A_1(\omega,i,j) \wedge A_2(\omega,i,j)).$$

As furthermore

$$p_{2k}+l \leq p_{2k}+e_2(\lambda/2^{2k+1}, a^{(p_{2k})}, b^{(p_{2k})})+1 = p_{2k+1} \leq p_{2e_1(\lambda/2,a',b')+1} = e(\lambda,a,b),$$

we have thereby shown that

$$P \subseteq \{\omega \in \Omega : \exists n \leq e(\lambda,a,b)\ \forall i,j \in [a_n; b_n]\ (A_1(\omega,i,j) \wedge A_2(\omega,i,j))\}.$$

Thus we conclude that

$$\begin{aligned} &\mathbb{P}(\forall n \leq e(\lambda,a,b)\ \exists i,j \in [a_n; b_n]\ \ (\neg A_1(i,j) \vee \neg A_2(i,j))) \\ &\quad \leq \mathbb{P}(\neg Q) + \mathbb{P}\left(\bigcup_{i \leq e_1(\lambda/2,a',b')} \neg R_{2i} \right) < \frac{\lambda}{2} + \sum_{i=0}^{\infty} \mathbb{P}(\neg R_{2i}) < \lambda. \end{aligned}$$

By considering the one-point probability content space, an immediate consequence of the above is the following result in the deterministic case:

Theorem 3. *Suppose $A_1(i,j), A_2(i,j)$ are formulas and let e_1 and e_2, respectively, be dependent learnable rates for $\exists n \forall i,j \geq n\, A_1(i,j)$ and $\exists n \forall i,j \geq n\, A_2(i,j)$. Then $\exists n \forall i,j \geq n\ (A_1(i,j) \wedge A_2(i,j))$ has a dependent learnable rate given by*

$$e(a,b) := p_{2e_1(a',b')+1}$$

where a', b' are defined by $a'_i := a_{p_{2i}}$, $b'_i := b_{p_{2i+1}}$ for all $i \in \mathbb{N}$, and with (p_i) defined inductively by

$$p_0 := 0 \text{ and } p_{i+1} := p_i + e_2(a^{(p_i)}, b^{(p_i)}) + 1.$$

The case for uniform dependent learnable rates can be obtained in a similar manner to the pointwise case, in which case however some simplifications apply. We collect this in the following theorem, the proof-sketch of which we defer to the appendix.

Theorem 4. *Suppose $A_1(i,j), A_2(i,j)$ are measurable formulas and let e_1 and e_2, respectively, be uniform dependent learnable rates for $\exists n \forall i,j \geq n\, A_1(i,j)$ and $\exists n \forall i,j \geq n\, A_2(i,j)$. Then $\exists n \forall i,j \geq n\ (A_1(i,j) \wedge A_2(i,j))$ has a uniform dependent learnable rate given by*

$$e(\lambda,a,b) := p_{2e_1(\lambda/2,a',b')+1}$$

where a', b' are defined by $a'_i := a_{p_{2i}}$, $b'_i := b_{p_{2i+1}}$ for all $i \in \mathbb{N}$, and with (p_i) defined inductively by

$$p_0 := 0 \text{ and } p_{i+1} := p_i + e_2(\lambda/2, a^{(p_i)}, b^{(p_i)}) + 1.$$

In the case of pointwise learnable rates in the ordinary sense, one can also drastically simplify the above bounds. Again, we provide a proof-sketch in the appendix.

Theorem 5. *Suppose $A_1(i,j), A_2(i,j)$ are measurable formulas and let e_1 and e_2, respectively, be pointwise learnable rates for $\exists n \forall i,j \geq n\, A_1(i,j)$ and $\exists n \forall i,j \geq n\, A_2(i,j)$. Then $\exists n \forall i,j \geq n\, (A_1(i,j) \wedge A_2(i,j))$ has a pointwise learnable rate given by*

$$e(\lambda) := \left(e_2\left(\frac{\lambda}{2(e_1(\lambda/2)+1)}\right)+1\right)(2e_1(\lambda/2)+1).$$

The last of the further simplifications that we want to highlight is the case of uniform learnable rates in the ordinary sense where the above bound just collapses to a sum. As this follows rather immediately using arguments similar to [20, Lemma 2.10] and [19, Lemma 7.3], we omit the proof.

Theorem 6. *Suppose $A_1(i,j), A_2(i,j)$ are measurable formulas and let e_1 and e_2, respectively, be uniform learnable rates for $\exists n \forall i,j \geq n\, A_1(i,j)$ and $\exists n \forall i,j \geq n\, A_2(i,j)$. Then $\exists n \forall i,j \geq n\, (A_1(i,j) \wedge A_2(i,j))$ has a uniform learnable rate given by*

$$e(\lambda, a, b) := e_1(\lambda/2) + e_2(\lambda/2).$$

4 Our results in the light of game semantics

Another benefit of our various notions of learnability is that they have an intuitive reading in terms of game semantics. This was already alluded to in Section 2.1 above, and this in short section, we extend this intuition to our new notion of stochastic learnability and our joining constructions.

It has long been known that one can view the computational content of classical reasoning in terms of winning strategies in backtracking games, an idea that is already implicit Hilbert's epsilon calculus, and has been explored in numerous different settings since. A particularly readable account is given by Coquand in [6], where an explicit connection is made between games and double negative translations (as represented in the present paper by $(**)$ above), also in the context of a simple "joining" operation (cf. [6, Section 3]). While we do not go into details of the general situation here, we discuss how both our notions of learning and the stochastic joining constructions can be connected back to game semantics in the sense of [6].

For all of our main definitions of learnability, we now interpret $a_0 < b_0 \leq a_1 < b_1 \leq \dots$ as a dialogue between two players (E) and (A), where $a_0 < a_1 < \dots$ represent attempts by (E) to find a number a such that $\forall i,j \geq a\, A(i,j)$, while $b_0 < b_1 < \dots$ represent counter-moves by (A) to refute these attempts, i.e. for each a_n find some b_n such that $\exists i,j \in [a_n; b_n]\, \neg A(i,j)$. In this sense, a_n represents a winning move whenever $\forall i,j \in [a_n; b_n]\, A(i,j)$. With these intuitions in mind, all of our learnability bounds represents the maximum number of mind changes that (E) might need to make before finding a winning move, with slight differences in meaning as set out below:

- **Pointwise dependent:** For any λ and dialogue a, b, the probability that (E) wins with at most $e(\lambda, a, b)$ mind changes is $> 1 - \lambda$.
- **Uniform dependent:** For any λ and dialogue a, b, with at most $e(\lambda, a, b)$ mind changes (E) finds a move that wins with probability $> 1 - \lambda$.

In the nonstochastic case both variants collapse to the same notion, namely that (E) wins with at most $e(a, b)$ mind changes, and whenever $e(\lambda, a, b)$ is independent of the dialogue, we obtain the previously studied notions of pointwise/uniform learnability.

More interestingly, the main constructions of Section 3 can be understood as constructions on games. For example, informally speaking, in Theorem 2, we decompose the dialogue into a sequence of consecutive finite subdialogues marked by the points (p_i), where within each subdialogue the probability that (E) finds a winning move for A_2 is $> 1 - \lambda/2^{i+1}$. The endpoints (p_i) are then used to define a sparser dialogue, for which we can provide a bound such that the probability that (E) finds a winning move for A_1 on this sparse dialogue is $> 1 - \lambda/2$. The proof of Theorem 2 makes formal the simple idea that if (E) fails to find a winning move for $A_1 \wedge A_2$ within the stated number of mind changes, then this represents a failure either in the sparse dialogue for A_1, or one of the subdialogues for A_2, the total probability of which is bounded by $\lambda/2 + \sum_{i=0}^{\infty} \lambda/2^{i+1} = \lambda$.

The other constructions in Section 3 can be interpreted in a similar way, where the situation for uniform rates (Theorem 6) is notable in that the construction is *symmetric* in A_1 and A_2, which as discussed in [6, Section 3.2] is not generally the case for strategies arising from double negative translations.

5 Concluding remarks

The main achievements of this paper have been a new notion of learnability in the stochastic setting, which is general enough that we can expect computable bounds to be extractable for any convergent stochastic process for which there exists a convergence proof that lies within the scope of techniques from proof mining, along with a series of joining results which we expect to be crucial for future applications of proof theory in probability.

Several interesting questions directly leading off from the work in this paper present themselves: Are the constructions given in Section 3 optimal? This is particularly relevant given that, in practice, the joining constructions would likely be iterated to form a finite joining operation. Can we flesh out our hierarchy of quantitative notions for stochastic convergence with further levels? For example, the hierarchy for deterministic convergence in [15] has a more complex notion of learnability that, computationally, lies strictly between learnability and dependent learnability in our sense. Finally, can we give a more rigorous definition of a learning semantics for classical probability theory, along the lines of [6] but applicable to a logical system for probability as in [18]?

References

1. Avigad, J., Dean, E., Rute, J.: A metastable dominated convergence theorem. Journal of Logic and Analysis **4**(3) (2012), 19pp.
2. Avigad, J., Gerhardy, P., Towsner, H.: Local stability of ergodic averages. Transactions of the American Mathematical Society **362**(1), 261–288 (2010)
3. Bhaskara Rao, K., Bhaskara Rao, M.: Theory of Charges, Pure and Applied Mathematics, vol. 109. Elsevier (1983)
4. Bishop, E.: Foundations of Constructive Analysis. Mcgraw-Hill, New York, NY, USA (1967)
5. Chashka, A.: Fluctuations in martingales. Uspekhi Matematicheskikh Nauk **49**(2), 179–180 (1994), english translation in Russian Math. Surveys 49:2 (1994)
6. Coquand, T.: A semantics of evidence for classical arithmetic. J. Symb. Log. **60**(1), 325–337 (Mar 1995)
7. Dubins, L.E.: Some upcrossing inequalities for uniformly bounded martingales. In: Symposia Mathematica. vol. 9, pp. 169–177 (1972)
8. Hochman, M.: Upcrossing inequalities for stationary sequences and applications. Annals of Probability **37**(6), 2135–2149 (2009)
9. Hoyrup, M., Rute, J.: Computable Measure Theory and Algorithmic Randomness, pp. 227–270. Springer (2021)
10. Ivanov, V.: Oscillations of averages in the ergodic theorem. Doklady Akademii Nauk **347**(6), 736–738 (1996)
11. Jones, R., Kaufman, R., Rosenblatt, J., Wierdl, M.: Oscillation in ergodic theory. Ergodic Theory and Dynamical Systems **18**(4), 889–935 (1998)
12. Kachurovskii, A.: The rate of convergence in ergodic theorems. Russian Mathematical Surveys **51**(4), 653–703 (1996)
13. Kohlenbach, U.: Applied Proof Theory: Proof Interpretations and their Use in Mathematics. Springer Monographs in Mathematics, Springer (2008)
14. Kohlenbach, U.: Proof-theoretic Methods in Nonlinear Analysis. In: Proceedings ICM 2018. vol. 2, pp. 61–82 (2019)
15. Kohlenbach, U., Safarik, P.: Fluctuations, effective learnability and metastability in analysis. Annals and Pure and Applied Logic **165**, 266–304 (2014)
16. Kreisel, G.: On the Interpretation of Non-Finitist Proofs, Part I. Journal of Symbolic Logic **16**, 241–267 (1951)
17. Kreisel, G.: On the Interpretation of Non-Finitist Proofs, Part II: Interpretation of Number Theory. Journal of Symbolic Logic **17**, 43–58 (1952)
18. Neri, M., Pischke, N.: Proof mining and probability theory (2024), preprint, available at `https://arxiv.org/abs/2403.00659`
19. Neri, M., Powell, T.: On quantitative convergence for stochastic processes: Crossings, fluctuations and martingales (2024), preprint, available at `https://arxiv.org/abs/2406.19979`
20. Neri, M., Powell, T.: A quantitative Robbins-Siegmund theorem (2024), preprint, available at `https://arxiv.org/abs/2410.15986`
21. Oliva, P.: Understanding and using Spector's bar recursive interpretation of classical analysis. In: Logical Approaches to Computational Barriers, pp. 423–434. Lecture Notes in Computer Science, Springer (2006)
22. Powell, T.: A note on the finitization of Abelian and Tauberian theorems. Mathematical Logic Quarterly **66**(3), 300–310 (2020)
23. Specker, E.: Nicht Konstruktiv Beweisbare Sätze der Analysis. Journal of Symbolic Logic **14**, 145–158 (1949)
24. Tao, T.: Norm convergence of multiple ergodic averages for commuting transformations. Ergodic Theory and Dynamical Systems **28**, 657–688 (2008)

A Detailed proofs

Proof (of Theorem 4). Suppose a, b are sequences of natural numbers satisfying $a_0 < b_0 \leq a_1 < b_1 \leq \dots$ and let $\lambda \in \mathbb{Q}^+$ be given. From the fact that e_1 is a uniform dependent learnable rate for A_1, and noting the properties of a', b' as in the proof of Theorem 2, we get that there exists a $k \leq e_1(\lambda/2, a', b')$ such that $\mathbb{P}(\neg Q_k) < \lambda/2$ for Q_k defined by

$$Q_k := \{\omega \in \Omega : \forall i, j \in [a'_k; b'_k]\ A_1(\omega, i, j)\}.$$

Further, from the fact that e_2 is a uniform dependent learnable rate for A_2, and noting the properties of $a^{(p_i)}, b^{(p_i)}$ for $i \in \mathbb{N}$ as in the proof of Theorem 2, we get that there exists an $l \leq e_2(\lambda/2, a^{(p_{2k})}, b^{(p_{2k})})$ such that $\mathbb{P}(\neg R_{k,l}) < \lambda/2$ for $R_{k,l}$ defined by

$$R_{k,l} := \{\omega \in \Omega : \forall i, j \in [a_l^{(p_{2k})}; b_l^{(p_{2k})}]\ A_2(\omega, i, j)\}.$$

Letting $\omega \in P := Q_k \cap R_{k,l}$, we thus have

$$\forall i, j \in [a_{p_{2k}}; b_{p_{2k+1}}]\ A_1(\omega, i, j) \text{ and } \forall i, j \in [a_{p_{2k}+l}; b_{p_{2k}+l}]\ A_2(\omega, i, j)$$

from which we can conclude

$$\forall i, j \in [a_{p_{2k}+l}; b_{p_{2k}+l}]\ (A_1(\omega, i, j) \wedge A_2(\omega, i, j))$$

as in the proof of Theorem 2. With $n = p_{2k} + l \leq e(\lambda, a, b)$ as in Theorem 2, we have derived that

$$P \subseteq \{\omega \in \Omega : \forall i, j \in [a_n; b_n]\ (A_1(\omega, i, j) \wedge A_2(\omega, i, j))\}$$

from which we conclude that

$$\mathbb{P}(\exists i, j \in [a_n; b_n]\, (\neg A_1(i, j) \vee \neg A_2(i, j))) \leq \mathbb{P}(\neg Q_k) + \mathbb{P}(\neg R_{k,l}) < \lambda$$

as was to show.

Proof (of Theorem 5). Suppose a, b are sequences of natural numbers satisfying $a_0 < b_0 \leq a_1 < b_1 \leq \dots$ and let $\lambda \in \mathbb{Q}^+$ be given. Define $p_n := n(e_2(\lambda/2(e_1(\lambda/2)+1))+1)$ and define a', b' relative to this as in Theorem 2. Set

$$Q := \{\omega \in \Omega : \exists k \leq e_1(\lambda/2)\ \forall i, j \in [a'_k; b'_k]\ A_1(\omega, i, j)\}.$$

From the fact that e_1 is a pointwise learnable rate for A_1, and noting the properties of a', b' which can be discussed analogously as in the proof of Theorem 2, we have $\mathbb{P}(\neg Q) < \lambda/2$. Further, for each $k \in \mathbb{N}$, set

$$R_k := \{\omega \in \Omega : \exists l \leq e_2(\lambda/2(e_1(\lambda/2) + 1))\ \forall i, j \in [a_l^{(p_k)}; b_l^{(p_k)}]\ A_2(\omega, i, j)\}.$$

From the fact that e_2 is a pointwise learnable rate for A_2, and noting the properties of $a^{(p_i)}, b^{(p_i)}$ for $i \in \mathbb{N}$ as in the proof of Theorem 2, we get $\mathbb{P}(\neg R_k) < \lambda/2(e_1(\lambda/2)+1)$. For any

$$\omega \in P := Q \cap \bigcap_{i \leq e_1(\lambda/2)} R_{2i},$$

we can then conclude as in the proof of Theorem 2 that there are $k \leq e_1(\lambda/2)$ and $l \leq e_2(\lambda/2(e_1(\lambda/2)+1))$ such that

$$\forall i,j \in [a_{p_{2k}}; b_{p_{2k+1}}]\ A_1(\omega,i,j) \text{ and } \forall i,j \in [a_{p_{2k}+l}; b_{p_{2k}+l}]\ A_2(\omega,i,j)$$

from which it as before follows that

$$\forall i,j \in [a_{p_{2k}+l}; b_{p_{2k}+l}]\ (A_1(\omega,i,j) \wedge A_2(\omega,i,j)).$$

As furthermore

$$p_{2k} + l \leq p_{2k} + e_2(\lambda/2(e_1(\lambda/2)+1)) + 1 = p_{2k+1} \leq p_{2e_1(\lambda/2)+1} = e(\lambda),$$

we have thereby shown that

$$P \subseteq \{\omega \in \Omega : \exists n \leq e(\lambda)\ \forall i,j \in [a_n; b_n]\ (A_1(\omega,i,j) \wedge A_2(\omega,i,j))\}$$

and can hence conclude that

$$\begin{aligned}
&\mathbb{P}(\forall n \leq e(\lambda)\ \exists i,j \in [a_n; b_n]\ \ (\neg A_1(i,j) \vee \neg A_2(i,j))) \\
&\quad \leq \mathbb{P}(\neg Q) + \mathbb{P}\left(\bigcup_{i \leq e_1(\lambda/2)} \neg R_{2i}\right) \\
&\quad < \frac{\lambda}{2} + \sum_{i=0}^{e_1(\lambda/2)} \frac{\lambda}{2(e_1(\lambda/2)+1)} = \lambda.
\end{aligned}$$