Randomness, exchangeability, and conformal prediction

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #42

First posted January 20, 2025. Last revised November 20, 2025.

Project web site:
http://alrw.net

Abstract

This paper argues for a wider use of the functional theory of randomness, a modification of the algorithmic theory of randomness getting rid of unspecified additive constants. Both theories are useful for understanding relations between the assumptions of IID data and data exchangeability. While the assumption of IID data is standard in machine learning, conformal prediction relies on the weaker assumption of data exchangeability. Nouretdinov, V'yugin, and Gammerman showed, using the language of the algorithmic theory of randomness, that conformal prediction is a universal method under the assumption of IID data. In this paper, prepared for the Alex Gammerman Festschrift, I will selectively review connections between exchangeability and the property of being IID, early history of conformal prediction, my encounters and collaboration with Alex and other interesting people, and a translation of Nouretdinov et al.'s results into the language of the functional theory of randomness, which moves it closer to practice. Namely, the translation says that every confidence predictor that is valid for IID data can be converted into a conformal predictor without losing much in predictive efficiency.

Contents

1 Introduction

The functional theory of randomness was proposed in [50] under the name of non-algorithmic theory of randomness. The algorithmic theory of randomness originated with Kolmogorov in the 1960s [25] and has been extensively developed in numerous papers and books (see, e.g., [38]). It has been a powerful source of intuition, but its weakness is the dependence on the choice of a specific universal partial computable function. This dependence leads to the presence of unspecified additive (sometimes multiplicative) constants in its mathematical results. Kolmogorov [24, Sect. 3] speculated that for natural universal partial computable functions the additive constants will be in hundreds rather than in tens of thousands of bits, but this accuracy is very far from being sufficient in machine-learning and statistical applications (an additive constant of 100 in the definition of Kolmogorov complexity translates into the astronomical multiplicative constant of 2^{100} in the corresponding p-value).

The way of dealing with unspecified constants proposed in [50] is to express statements of the algorithmic theory of randomness as relations between various function classes. It will be introduced in Sect. 6. In this paper we call this approach the functional theory of randomness. While it loses somewhat in intuitive simplicity, it is closer to practical machine learning and statistics.

The main message of this paper is that the functional theory of randomness can be useful in the foundations of machine learning in general and conformal prediction in particular. The most standard assumption in machine learning is that the data are generated in the IID manner (are independent and identically distributed). An *a priori* weaker assumption is that of exchangeability, although for infinite data sequences being generated in the IID manner and exchangeability turn out to be essentially equivalent by the celebrated de Finetti representation theorem. The classical work on relations between the two assumptions, being IID and being exchangeable, will be the topic of Sect. 2.

The word "random" is often used in two very different senses: in the sense of statistical randomness referring to IID data (as in the title of [56]) and in the sense of algorithmic randomness (as in the title of [38]). In this paper I will try not to use "random" and its derivative "randomness" often, apart from expressions such as "algorithmic theory of randomness" and "functional theory of randomness". For the former sense, I will usually replace derivatives of "random" by compounds containing "IID", such as the *IID assumption* for the assumption that the data are generated in the IID manner (so that simply replacing "IID" by "independent identically distributed" becomes impossible, as in [57]). For the latter sense, I will often use the word "typical" and its derivative "typicalness", which was endorsed by Kolmogorov [38, Appendix 2, footnote 1] in its Russian form типичность and used in [27] (written by Uspensky [27, Introduction]).

Sections 3 and 4 contain personal elements. In Sect. 3 I recount meeting Andrei Kolmogorov and working under his supervision on the relation between IID and exchangeability. In Kolmogorov's frequentist philosophy of probability, the IID property was at the very basis of the notion of probability. In Sect. 4 I recount meeting Alex and then Vladimir Vapnik a decade later. Vapnik was a

second person who impressed me with his wholehearted acceptance of the IID assumption, which quickly led to the development of conformal prediction.

In my work under Kolmogorov, I realized that for finite data sequences the difference between IID and exchangeability is important. However, conformal prediction uses only exchangeability. This raises the question whether it is possible to improve on conformal prediction by using the stronger IID assumption. The topic of Sect. 5 is the fundamental result by Nouretdinov, V'yugin, and Gammerman saying that only limited improvement is possible. The result is stated in terms of the algorithmic theory of randomness, making it very intuitive, though the intuition can sometimes be obscured by dense notation.

Despite Nouretdinov et al.'s result being fundamental, it inherits the conceptual weakness of the algorithmic theory of randomness discussed earlier. As it involves unspecified constants, it cannot, strictly speaking, have any practical implications. In Sect. 6 I will state this and several related results in terms of the functional theory of randomness.

In this paper italic e may stand for an e-value. Euler's number (the base of the natural logarithms) is roman $\mathrm{e} \approx 2.72$. The notation for binary logarithm is lb [9, Sect. 10.1.2]. No detailed knowledge of the algorithmic theory of randomness is assumed on the part of the reader, but knowing the basics would be useful.

2 IID and exchangeability

Modelling data as IID observations is an ancient notion. Already Jacob Bernoulli [7] was using the IID assumption to state his weak law of large numbers. As Glenn Shafer reminds us in [18, Shafer's comment], the IID case has been central to probability and statistics ever since, "but its inadequacy was always obvious, and Leibniz made the point in his letters to Bernoulli: the world is in constant flux; causes do not remain constant, and so probabilities do not remain constant". There is no doubt the IID assumption is highly restrictive.

The real question is whether the IID assumption is a good *starting point*. It can be fundamental without being all-encompassing; many other, perhaps much more realistic, scenarios may reduce in some way to the IID case. For example, when dealing with prediction or decision making, is it a good strategy first to explore in detail what can be achieved under the IID assumption and then to try and relax it? Or is the assumption so restrictive that it is best to start elsewhere? My views about this have been drifting over time, and even now I remain uncertain.

One relatively modest extension of the IID assumption is the assumption of exchangeability; for potentially infinite data sequences one may even argue that it is not an extension at all. Philip Dawid says in [60, Sect. 7], "For so long, and it's still true of 97% of everything done in statistics and machine learning and everything, the fundamental assumption is basically we have just a bag of exchangeable goodies. And I thought that was just so limiting; how boring. There's a big wide world beyond that." This has been my feeling as well

since I started thinking about such things (and among my colleagues, Shafer's and Dawid's views on the philosophy of probability are perhaps closest to mine). However, I have also been impressed by the power of algorithms developed under IID and exchangeability and by many ingenuous ways of greatly relaxing these assumptions.

It is not clear who introduced exchangeability (see [11] for the complicated history), but the most well-known theorem about exchangeability is de Finetti's (generalized in later papers by other people), which connects it with IID. Let $\mathbf{Z}$ be an *observation space* (formally, a measurable space), and suppose that we observe its elements $z_i \in \mathbf{Z}$, $i = 1, 2, \ldots$, sequentially. The IID assumption is that the observations z_i are generated from an *IID probability measure* Q^∞, Q being a probability measure on $\mathbf{Z}$. The assumption of exchangeability is that they are generated from an *exchangeable* probability measure on $\mathbf{Z}^\infty$, i.e., a probability measure that is invariant w.r. to permutations of finitely many observations.

Remark 1. In [56] we referred to probability measures of the form Q^n, with $n = \infty$ allowed, as "power probability measures". In this paper I am using the expression "IID probability measures" instead to simplify terminology. Exchangeability of a probability measure on $\mathbf{Z}^n$ for $n < \infty$ still means invariance w.r. to permutations of observations.

According to de Finetti's theorem (see, e.g., [33, Theorem 1.49]), for infinite sequences, the IID and exchangeability assumptions are equivalent. Namely, each exchangeable probability measure R on $\mathbf{Z}^\infty$ is a convex mixture of IID probability measures: there exists a probability measure μ on the family $\mathfrak{P}(\mathbf{Z})$ of all probability measures on $\mathbf{Z}$ such that

$$R = \int_{\mathfrak{P}(\mathbf{Z})} Q^\infty \, \mu(\mathrm{d}Q).$$

The theorem makes the weak assumption that $\mathbf{Z}$ is a standard Borel space (and then $\mathfrak{P}(\mathbf{Z})$ is equipped with the smallest σ-algebra making all evaluation functionals measurable).

I find it intuitively compelling that the assumption that the data are generated from a statistical model M (i.e., a family probability measures) is equivalent to the assumption that the data are generated from the convex hull $\bar{M}$ of that statistical model. However, there are people who do not share this intuition (e.g., a friendly reviewer for [58]), so let me try to make it more explicit. The most basic way of testing a statistical model M ("Cournot's principle") is to select in advance a *critical region* A of a small probability under any probability measure in M and reject M if the actual data happens to be in A. Since

$$\sup_{R \in M} R(A) = \sup_{R \in \bar{M}} R(A),$$

rejecting M and rejecting $\bar{M}$ are equivalent. This conclusion is not affected if Cournot's principle is replaced by more sophisticated ways of hypothesis testing, such as using p-variables or e-variables, to be discussed starting from the next

section. In particular, the IID and exchangeability assumptions are equivalent. This is far from being true for finite sequences, as will be discussed in detail in Sections 3–4.

The IID picture is fundamental in the frequentist theory of probability and statistics, at least as it was presented and developed by Richard von Mises [43, 44] and Andrei Kolmogorov [21, Sect. I.2], who was following von Mises. It is well known that Kolmogorov was the first to put the mathematical theory of probability on a firm axiomatic basis in his 1933 book [21]. However, while the axioms of probability introduced in this book eventually (albeit slowly) gained universal acceptance (see, e.g., [36]), the way in which Kolmogorov proposed to connect his axioms with reality [21, Sect. I.2] was informal and has never become widely accepted. According to Kolmogorov's frequentist Principle A, introduced in [21, Sect. I.2], we can say that an event A has a probability $\mathbb{P}(A)$ under a system of conditions $\mathfrak{S}$ if

> One can be practically certain that if the system of conditions $\mathfrak{S}$ is repeated a large number of times, n, and the event A occurs m times, then the ratio m/n will differ only slightly from $\mathbb{P}(A)$.

(When quoting [21, Sect. I.2] I am using the translation given in [36, Sect. 5.2.1].) This principle, which Kolmogorov traces back to von Mises [21, Sect. I.2, footnote 1], gives us a way of measuring $\mathbb{P}(A)$. Presumably the repetitions in Principle A are independent, and so IID observations are at the heart of frequentist probability.

Remark 2. Kolmogorov's approach was not purely frequentist. Alongside his frequentist Principle A he also had a non-frequentist Principle B, namely Cournot's principle:

> If $\mathbb{P}(A)$ is very small, then one can be practically certain that the event A will not occur on a single realization of the conditions $\mathfrak{S}$.

Kolmogorov [21, Sect. I.2] postulated both principles, but it can be argued that Principle B renders Principle A redundant [36, Sect. 5.2]. All the books that I have co-authored so far can be traced back either to Kolmogorov's Principle A [56] or to his Principle B [35, 37].

For many years after the publication of his book [21], Kolmogorov talked about connections of his axioms with reality only informally, believing that von Mises's approach, which was based on a flawed definition of an individual infinite sequence of IID observations, could not be cleanly applied to the real world. (See, e.g., [22].) A breakthrough came when Kolmogorov visited India in 1962 [23]. He realized that von Mises's picture can be made applicable to finite sequences (albeit in a way that looks awkward to me, no doubt in hindsight, in view of his later elegant algorithmic approach, which was much more typical of Kolmogorov).

Soon afterwards Kolmogorov came up with his algorithmic theory of randomness subsuming his theory developed in India. In particular, he formalized what it means for a finite binary sequence to be a typical IID sequence. Since

Kolmogorov was only dealing with binary sequences, he referred to typical IID sequences as "Bernoulli sequences". The key to his definition was a notion of algorithmic complexity ("Kolmogorov complexity"), and typicalness was defined as maximal complexity in a finite set. This theory was described in his papers [24–26].

Martin-Löf [28] translated Kolmogorov's definition of typicalness into a more standard statistical language defining universal p-values. Later Levin and Gács modified Martin-Löf's definition in an important way, which I will discuss in the next section.

3 Meeting Kolmogorov; IID vs exchangeability for finite sequences

In 1980 Kolmogorov became Head of the Department of Mathematical Logic at Moscow University, and in the same year I became his student. This happened after I attended his talk aimed at undergraduate students and afterwards spoke to Alexei Semenov, who in his role of the departmental scientific secretary (учёный секретарь кафедры) took care of the administrative side. First I did the specialized part of a Soviet combined BSc/MSc degree programme under Kolmogorov's supervision (the specialized part covering the last three years of the 5-year degree programme), and then I did a PhD under the joint supervision of Kolmogorov and Semenov.

One of the problems that Kolmogorov offered to me was to quantify the qualitative (and intuitively obvious) statement that Bernoulli sequences satisfy his frequentist definition as given in [23]. This was a difficult problem that did not look particularly appealing to me (some results in this direction were obtained later by Kolmogorov himself and his other student Eugene Asarin; see [5, Theorem 3] for Kolmogorov's result and [4, Theorem 1] for Asarin's). Instead, I chose to investigate the relation between IID and exchangeability for finite sequences.

As I mentioned in the previous section, Kolmogorov was only interested in finite sequences in his work on the foundations of probability and believed that infinite sequences, being empirically non-existent, are irrelevant when discussing connections between the mathematical theory of probability and reality. At one point during a walk to a train station Kolmogorov told me that we can only see finite sequences around us, but in the quote from Kolmogorov given in [2, Chap. 7, bottom of p. 57] the word "only" is misplaced (Kolmogorov could not see any infinite sequences, and neither could I).

In my first journal paper [45] I explored Kolmogorov's definition of Bernoulli sequences and argued that it was a formalization of a different kind of typicalness, not typicalness under IID. The term that I used, on Alexander Zvonkin's advice, for Kolmogorov's Bernoulli sequences was von Mises's "collectives", but in hindsight I meant exchangeability (and I talk about exchangeability in the technical report [45] containing the proofs). After that I introduced a definition

of typicalness under IID for binary sequences and characterized the difference between the two definitions. To state these results, let me give the relevant (standard) definitions.

I will use the notion of an aggregate of constructive objects, as in [41, Sect. 1.0.6]. This is an infinitely countable set whose elements can be effectively numbered, such as the set $\mathbb{Z}$ of all integer numbers or the set $\{0,1\}^*$ of all finite binary sequences. Let the observation space $\mathbf{Z}$ range over the finite non-empty subsets of a fixed aggregate of constructive objects. In Kolmogorov's work in this area and in my work reported in this section, $\mathbf{Z} = \{0,1\}$, but let me give more general definitions for later use (e.g., in the context of conformal prediction). Let $\mathbb{N}$ be the set of natural numbers; by default we do not include 0 in $\mathbb{N}$, so that $\mathbb{N} = \{1,2,\dots\}$.

A real-valued function f defined on an aggregate of constructive objects is *lower semicomputable* if there is an algorithm that, when fed with v in the domain of f and $r \in \mathbb{Q}$ ($\mathbb{Q}$ being the set of rational numbers), eventually stops if and only if $f(v) > r$. Similarly, it is *upper semicomputable* if this condition holds with $f(v) > r$ replaced by $f(v) < r$. (And computability is the conjunction of lower and upper semicomputability.)

Let me start from a definition equivalent to Kolmogorov's, which is closest to conformal prediction. A *p-test for exchangeability* is a $[0,1]$-valued upper semicomputable function P that takes as input $\mathbf{Z}$, $N \in \mathbb{N}$, and a sequence $z_1,\dots,z_N$ in $\mathbf{Z}^N$ and that satisfies, for all $\mathbf{Z}$, all N, and all exchangeable probability measures R on $\mathbf{Z}^N$,

$$\forall \epsilon \in (0,1) : R(\{\zeta : P(\zeta) \le \epsilon\}) \le \epsilon \tag{1}$$

(omitting, here and later, mentioning $\mathbf{Z}$ and N as arguments; remember that $\mathbf{Z}$ always ranges over the finite subsets of a fixed aggregate of constructive objects). The requirement (1) is usually expressed by saying that P, for fixed $\mathbf{Z}$ and N, is a *p-variable* (and its values are *p-values*). There exists a smallest, to within a constant factor, p-test for exchangeability, which is then called *universal*. Let us fix a universal p-test for exchangeability and let D^{pX} stand for its minus binary logarithm. We call $D^{\mathrm{pX}}(z_1,\dots,z_N)$ the *exchangeability p-deficiency* of the sequence $(z_1,\dots,z_N)$. We can also allow P to depend on a *condition* (an integer number) (and then P is required to be lower semicomputable as function of all its arguments, including the condition); the full notation for the exchangeability p-deficiency is then $D^{\mathrm{pX}}(z_1,\dots,z_N \mid k)$, where k is the condition.

The function D^{pX} can be defined to some degree arbitrarily; different choices of D^{pX}, however, will coincide to within an additive constant. This renders results of the algorithmic theory of randomness inapplicable in practice. When we say that two functions (such as D^{pX}) that are only defined to within an additive constant coincide, we mean, of course, that their difference is bounded. In general, when discussing relations (such as inequalities) between such functions, we will always ignore additive constants. Without loss of generality, we will assume that the function D^{pX}, and similar functions introduced later in this paper, are integer-valued.

In the case $\mathbf{Z} = \{0,1\}$, $D^{\mathrm{pX}}(z_1,\dots,z_N)$ coincides with Kolmogorov's def-

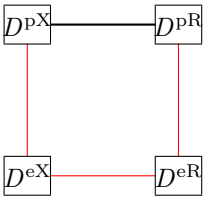

Figure 1: Connections between 4 deficiencies of typicalness: The connection between D^{pX} and D^{pR} (shown as thick black line) is established via the chain D^{pX}–D^{eX}–D^{eR}–D^{pR} (shown as thin red lines).

inition of Bernoulliness, as follows from [53, Proposition 11]. Kolmogorov's expression for "$D^{\mathrm{pX}}(z_1,\dots,z_N) \le m$" was "$z_1,\dots,z_N$ is m-Bernoulli" ("m-бернуллиевская").

In a similar way, we can define the *IID p-deficiency* $D^{\mathrm{pR}}(z_1,\dots,z_N)$ of a sequence $(z_1,\dots,z_N)$. The only difference is that we replace p-tests for exchangeability by *p-tests for IID* defined by letting R in the definition (1) range over the IID probability measures Q^N, Q being a probability measure on $\mathbf{Z}$ generating one observation. A universal p-test for IID also exists; we let D^{pR} stands for its minus binary logarithm and call $D^{\mathrm{pR}}(z_1,\dots,z_N)$ the *IID p-deficiency* of $z_1,\dots,z_N$. This is equivalent to my definition proposed in [45].

To establish connections between D^{pX} and D^{pR}, I followed a strategy that is standard in the algorithmic theory of randomness. While hypothesis testing in classical statistics is based on p-values, e-values are mathematically much more convenient and often serve as a useful tool. Universal e-values were introduced in the algorithmic theory of randomness by Levin and then simplified by Gács [15] (see also [61]), without using this expression, and nowadays non-universal e-values are gaining popularity in statistics (see, e.g., [20,31,62]). Therefore, the Martin-Löf-style functions D^{pX} and D^{pR} were connected in [45] by connecting D^{pX} and D^{eX}, then D^{eX} and D^{eR}, and finally D^{eR} and D^{pR}, where D^{eX} and D^{eR} are Levin-style analogues of D^{pX} and D^{pR}, to be defined momentarily. (The connections are shown in Fig. 1.)

An *e-test for exchangeability* is a nonnegative lower semicomputable function E on the same domain as a p-test for exchangeability, but it is required to satisfy

$$\sum_{\zeta\in\mathbf{Z}^N} E(\zeta)R(\{\zeta\}) \le 1 \tag{2}$$

in place of (1) for all $\mathbf{Z}$, N, and exchangeable R. Both upper semicomputability for P (in (1)) and lower semicomputability for E (in (2)) are natural requirements: we reject the null hypothesis of exchangeability (or IID) when a p-value is small (say, below some threshold such as 1% or 5%) or an e-value is large, and the decision to reject should be taken in finite time. The condition (2) means that, for fixed $\mathbf{Z}$ and N, E is an *e-variable*, and then its values are *e-values*. We fix a universal (this time meaning largest to within a constant factor) e-test for

exchangeability and call its binary logarithm D^{eX} *exchangeability e-deficiency*. Replacing exchangeable R by $R := Q^N$, we get the definition of D^{eR}, *IID e-deficiency*.

For any data sequence $z_1, \dots, z_N$, let us define the IID e-deficiency of the corresponding *configuration* $\wr z_1, \dots, z_N \int$, i.e., of the bag (multiset) of its elements, as

$$D^{\mathrm{eR}}(\wr z_1, \dots, z_N \int) := \min_{\pi} D^{\mathrm{eR}}(z_{\pi(1)}, \dots, z_{\pi(N)}), \tag{3}$$

π ranging over all permutations of $\{1, \dots, N\}$. In other words, a bag is IID (compatible with the IID assumption) if it can arise from an IID data sequence. If ζ is a data sequence, we let $\wr \zeta \int$ stand for the bag of its elements.

The following relation between IID and exchangeability is stated in [45, Theorem 1] for $\mathbf{Z} = \{0, 1\}$ and in [55, Theorem 3] in general. It uses "$=^+$" to mean coincidence of two functions to within an additive constant; $\mathbf{Z}^+$ is the family of all non-empty finite sequences of observations. Remember that $\mathbf{Z}$ varies over the finite non-empty subsets of a fixed aggregate of constructive objects.

Theorem 1. *Let ζ range over $\mathbf{Z}^+$ for a variable $\mathbf{Z}$. Then*

$$D^{\mathrm{eR}}(\zeta) =^+ D^{\mathrm{eR}}(\wr \zeta \int) + D^{\mathrm{eX}}(\zeta \mid D^{\mathrm{eR}}(\wr \zeta \int)). \tag{4}$$

Theorem 1 clarifies the relation between exchangeability and IID in the case of finite sequences: a sequence is IID if and only if it is exchangeable and its configuration is IID. The difference between the deficiencies of IID and exchangeability of a data sequence is, roughly, the IID deficiency of its configuration; the condition "$\mid D^{\mathrm{eR}}(\wr \zeta \int$" in (4) slightly obscures this, but (4) implies, e.g.,

$$D^{\mathrm{eR}}(\wr \zeta \int) + D^{\mathrm{eX}}(\zeta) \le^+ D^{\mathrm{eR}}(\zeta) \le^+ 1.01\, D^{\mathrm{eR}}(\wr \zeta \int) + D^{\mathrm{eX}}(\zeta)$$

(with "$\le^+$" denoting the inequality to within an additive constant).

Another result that I obtained in [45] (Theorem 2) was about how big the difference given by (4) between Kolmogorov's and my definitions can be. In the binary case considered in that paper, the configuration $\wr \zeta \int$ carries the same information as the number of 1s in ζ given its length N. Let k be the number of 1s. Then $D^{\mathrm{eR}}(k)$ can be characterized as the typicalness deficiency of k in its neighbourhood of size $\sqrt{k(N-k)/N}$ (approximately $\sqrt{N}$ if k is neither very small nor very large). Informally, this is a requirement of local typicalness; e.g., $k = \lfloor N/2 \rfloor$ is untypical for a large N since it is described in such a simple way (given N, which the definition assumes). This characterization implies that the difference between D^{eX} and D^{eR} can be as large as $\frac{1}{2} \operatorname{lb} N$ on data sequences of length N, but not larger.

In the binary case the difference between IID and exchangeability appears small, of the order of magnitude $O(\log N)$, which is much less than the attainable upper bound of $N + o(N)$ on D^{eX} and D^{eR}. In the algorithmic theory of randomness, coincidence to within a logarithmic term is often considered as being sufficiently close to disregard the difference.

These statements are also true about the definitions D^{pX} and D^{pR} in terms of p-values, as these inequalities show:

$$\begin{aligned} D^{\mathrm{eX}} &\le^+ D^{\mathrm{pX}} \le^+ D^{\mathrm{eX}} + 2 \operatorname{lb} D^{\mathrm{eX}}, \\ D^{\mathrm{eR}} &\le^+ D^{\mathrm{pR}} \le^+ D^{\mathrm{eR}} + 2 \operatorname{lb} D^{\mathrm{eR}} \end{aligned} \tag{5}$$

(they can be proved in the same way as Proposition 1 in [30]). We can see that D^{eX} and D^{pX}, as well as D^{eR} and D^{pR}, coincide to within logarithmic terms. Therefore, D^{pX} and D^{pR} also coincide to within a logarithmic term. This may be the reason why Kolmogorov used D^{pX} rather than D^{pR} as formalization, in the binary case, of "a result of independent tests with a probability p of getting a one during each test" [25, Sect. 2].

Remark 3. In [25, Sect. 2] Kolmogorov says (in translation) about his proposed definition, "We view, approximately, in this manner 'Bernoulli sequences' where separate signs are 'independent' and appear with a certain probability p." It is natural to assume, which I did, that the word "approximately" here means that he is ignoring the $O(\log N)$ difference between the two deficiencies (IID vs exchangeability). However, Kolmogorov told me that this was not what he meant (and he did not elaborate further).

From the vantage point of conformal prediction, however, the difference of $\frac{1}{2} \operatorname{lb} N$ between D^{eX} and D^{eR} is not small at all. Before discussing this, let us check that this difference persists when we move to the p-versions, D^{pX} and D^{pR}. Indeed, let us consider a data sequence $\zeta \in \{0,1\}^N$ which is a typical element of the set of all binary sequences of length N containing exactly $\lfloor N/2 \rfloor$ 1s, for a large N. Then $D^{\mathrm{pX}}(\zeta)$ will be close to 0, while, by the local limit theorem [39, Sect. 1.6], $D^{\mathrm{pR}}(\zeta)$ will be close to $\frac{1}{2} \operatorname{lb} N$. Therefore, the difference between D^{pX} and D^{pR} can also be as large as $\frac{1}{2} \operatorname{lb} N$.

An ideal picture of conformal prediction will be introduced in the next section, but what is important for us now is that the largest p-deficiency at which an ideal conformal predictor can reject a false label for a test object is $\operatorname{lb} N$, where N is the length of the "augmented training sequence" (for details, see (10) below). Another manifestation of this phenomenon, which we will call the "fundamental limitation of conformal prediction", is the fact that the smallest possible conformal p-value is $1/N$. Now $\frac{1}{2} \operatorname{lb} N$ does not look small anymore. Even in the binary case, the difference between D^{pX} and D^{pR} can eat up half of the largest p-deficiency achievable by an ideal conformal predictor. In the non-binary case, the difference between IID and exchangeability becomes even more substantial; see inequality (8) below and its discussion.

The paper [45] did not contain any proofs. Full proofs were first published only in 2016, but the main components appeared in [48], as indicated in [45, Appendix C].

Remark 4. There are versions of de Finetti's theorem for finite sequences that assert near equivalence between a finite sequence being IID and being a prefix of a much longer finite sequence that is exchangeable (see, e.g., [13, 14] for much stronger results). This idea of using exchangeable extensions makes it possible

to adapt de Finetti's theorem to finite sequences, but in this paper we are only interested in basic exchangeability, with a fixed length of the data sequence.

4 Meeting Alex Gammerman and Vladimir Vapnik; emergence of conformal prediction

I first met Alex in Barcelona at EuroCOLT 1995, the Second European Conference on Computational Learning Theory (Barcelona, Spain, 13–15 March 1995). Shortly before that, Norman Gowar, the Principal of Royal Holloway, University of London, had suggested that Alex become the next Head of Department of Computer Science. Despite some initial misgivings, Alex agreed. He proposed establishing a machine-learning group in the department, and the Principal enthusiastically supported him. One of Alex's goals in attending EuroCOLT 1995 was to meet, as new Head of Department, active researchers in machine learning.

It is interesting that the First European Conference on Computational Learning Theory had been held at Royal Holloway, University of London, on 20–22 December 1993, yet neither Alex nor I attended it (even though Alex had started teaching at Royal Holloway in September of that year).

Apart from discussing research at EuroCOLT 1995 (in particular, I learned about Alex's interest in Kolmogorov complexity), I remember an enjoyable walk past the Columbus monument at the bottom of La Rambla, Barcelona's iconic pedestrian street. My paper presented at the conference (and published in the proceedings as [47]) elaborated on the key element of the connection between IID and exchangeability found in [45] (I talked about it at length in the previous section). Later Paul Vitányi invited me to submit an extended version of [47] to a Special Issue of the *Journal of Computer and System Sciences* devoted to EuroCOLT 1995; the extended journal version appeared as [48] and later led to the publication of the proofs in [45].

In the summer of 1995 I moved to Stanford to spend a year at the Centre for Advanced Studies in the Behavioral Sciences (now part of Stanford University but then an independent institution). It had been difficult to survive doing science in Russia (I even attended a Business School in Moscow for a year, with an internship in the USA in the summer of 1992), and when Alex invited me to apply for a lectureship position at Royal Holloway to join the emerging machine-learning group (later called CLRC, Computer Learning Research Centre), I saw it as an exciting opportunity. In December 1995 I had an interview there, and at the same time my friend Philip Dawid arranged a backup interview at UCL in case I was unsuccessful. As it turned out, I was successful at Royal Holloway—and I have no idea how I fared at UCL. Even though I was appointed as lecturer, Alex told me that I would be promoted to Professorship within three years—and indeed, I was.

Alex's first two hires were Vladimir Vapnik (part-time) and, shortly afterwards, me. My family and I moved permanently to the UK in June 1996. That summer, Alex, Vapnik, and I had very fruitful discussions which later led, among

other things, to the development of conformal prediction. Vapnik was working (mainly or even exclusively) on support vector machines and writing his 1998 book [42], we discussed them repeatedly, and I was eager to contribute.

Before meeting Vapnik, I had not taken the IID assumption particularly seriously. My philosophy was affected by my work on what Shafer and I later called "game-theoretic probability" ([12, 46], with later books [35, 37] joint with Shafer), and as I mentioned earlier, this assumption appeared narrow to me. But Vapnik was taking it very seriously and in many cases did not even mention explicitly that he was making it (which at first even made it difficult for me to follow his arguments). It was a live demonstration of its importance, and indeed I soon realized that it was the most fundamental assumption in machine learning. And the problem of prediction under the IID assumption looked much more down-to-earth and less philosophical than providing frequentist foundations of probability (my preferred approach to the foundations of probability being based on Kolmogorov's Principle B rather than Principle A).

It was very natural to apply what I knew about typicalness deficiency to Vapnik's IID picture, as described briefly in [56, Sect. 2.9.2]. The ideal picture of prediction under IID or exchangeability is straightforward (and described in [55]). Let us suppose that each observation z consists of two components, an object x and its label y, and our task is to predict the label of a test object. Suppose the observation space $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is finite, where $\mathbf{X}$ is the object space and $\mathbf{Y}$ is the label space, both non-empty. Let $|\mathbf{Y}| > 1$. The possibility of the decomposition $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ does not restrict generality since we allow $|\mathbf{X}| = 1$. The upper or lower semicomputable functions producing IID and exchangeability deficiencies are given both $\mathbf{X}$ and $\mathbf{Y}$ as inputs (which are subsets of fixed aggregates of constructive objects). Given a training sequence $z_1, \dots, z_n$ and a test object x_{n+1}, our task is to predict the label y_{n+1} of x_{n+1}. We say that y_{n+1} is the *true label* of the test object x_{n+1} while labels $y \neq y_{n+1}$ are *false*. The number $N := n + 1$ can be interpreted, as is often done in conformal prediction, as the length of the "augmented training sequence" (the training sequence extended by the test object x_{n+1} with a possible label y).

In "universal prediction" we can use typicalness deficiency for evaluating the plausibility of various potential labels for the test object x_{n+1}. For that we can use any of D^{pX}, D^{pR}, D^{eX}, or D^{eR}, but for concreteness, let us concentrate on Kolmogorov's D^{pX} (which is particularly close to conformal prediction).

Remark 5. In the rest of this paper I will avoid using the expression "universal prediction" because of another unfortunate terminological clash (in addition to that between statistical randomness and algorithmic randomness discussed in Sect. 1). On one hand, the adjective "universal" may mean "related to universal partial computable functions", as in "universal p-test". On the other hand, Nouretdinov et al. [29] applied it to conformal prediction meaning that, under IID, it does not lose much in efficiency as compared with any other prediction method that is valid in the same sense. The two meanings are very different, so I will usually say "ideal prediction" rather than "universal prediction" when talking about prediction in the ideal picture based on universal partial computable

functions.

Our prediction and how confident we can be in it can be figured out by looking at the exchangeability deficiencies

$$f(y) := D^{\mathrm{pX}}(z_1, \dots, z_n, x_{n+1}, y) \tag{6}$$

for various potential labels $y \in \mathbf{Y}$ for the test object. (I am omitting parentheses in expressions such as $D^{\mathrm{pX}}(z_1, \dots, z_n, (x_{n+1}, y))$ if this is unlikely to lead to a misunderstanding.) For example, we can use

$$\hat{y}_{n+1} \in \arg\min_{y \in \mathbf{Y}} D^{\mathrm{pX}}(z_1, \dots, z_n, x_{n+1}, y)$$

(let us assume, for simplicity, that the arg min is attained at one point only) as the *point prediction* for the true test label y_{n+1}. However, the full *prediction function* f defined by (6) contains a lot of other useful information. For example, we can be confident that our point prediction is correct, $\hat{y}_{n+1} = y_{n+1}$, if the second smallest value $f(y)$ is large (presumably the smallest value is $f(y_{n+1})$ under exchangeability).

The point prediction $\hat{y}_{n+1}$ complemented by the second smallest value of f is a useful summary of the full prediction function f. Another way to summarize the prediction function (6) is to fix a significance level $\epsilon \in \mathbb{Q}$ (such as 5% or 1%) and output a prediction set using $-\operatorname{lb} \epsilon$ as threshold,

$$\Gamma^\epsilon := \{y \in \mathbf{Y} : D^{\mathrm{pX}}(z_1, \dots, z_n, x_{n+1}, y) < -\operatorname{lb} \epsilon\} \tag{7}$$

(as in [56, Sect. 2.2.4] but with p-values measured on the logarithmic scale). Notice that in this ideal picture the prediction sets are constructively closed (i.e., their indicator functions are upper semicomputable), which is natural: when computing the ideal prediction sets we keep making them narrower and narrower (i.e., better and better) as time passes.

The next question is how to make this ideal picture computable, so that we could use, e.g., support vector machines to find some practical approximations to the ideal prediction sets (7). This was a well-rehearsed step, which I had done earlier in, e.g., [46] when developing game-theoretic probability. The idea is to use the algorithmic theory of randomness for getting a clear intuitive picture of some area of probability or statistics, and then to strip the picture of its algorithmic content. This makes results more precise and, in particular, eliminates unspecified additive constants. The process is described in detail in [59, Sect. 6], where Shafer and I present the algorithmic theory of randomness as a tool of discovery.

Conformal prediction in its primitive binary form, which is a special case of both conformal prediction and conformal e-prediction, was introduced in [16]. It was applicable only to binary classification, since it was based on support vector machines. The nonconformity measure used in that paper assigns nonconformity scores of 1 to support vectors and nonconformity scores of 0 to all other observations. Let us call conformal prediction based on binary conformity

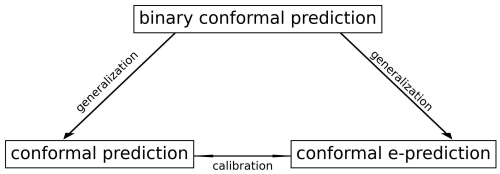

Figure 2: Binary conformal prediction as special case of both conformal prediction and conformal e-prediction. Calibration will be discussed in Sect. 6.4.

measures *binary conformal prediction*. See Fig. 2 for a pictorial representation, and see [63, Sect. 2] for a more general comparison of hypothesis testing based on p-values and e-values with binary testing based on Cournot's principle. The exposition in [16], however, emphasized conformal e-prediction much more prominently than conformal prediction.

Remark 6. In [16] we apply binary conformal prediction to a binary classification problem. This is a random coincidence, and binary conformal prediction is applicable to a wide range of prediction problems (see [52, Sect. 3] for an example).

Perhaps the first public announcement of conformal prediction in a wide sense (namely, of binary conformal prediction) happened in Alex's inaugural lecture in December 1996, which was later published locally as [17]. He and I regarded it as a public report about the work carried out at CLRC and worked on it together. My main contribution was to the section "Transduction" describing the binary conformal predictor based on support vector machines and connecting it to Vapnik's idea of transductive inference. Conformal prediction proper was introduced in [55] (and soon afterwards in [32], which mainly concentrated on support vector machines).

Untypically for literature on conformal prediction, the paper [55] that introduced it paid some attention to the ideal picture based on (6). One remark that it makes [55, Remark 5] is that, in the binary classification problem ($\mathbf{Y} = \{-1, 1\}$), if the true data sequence $z_1, \dots, z_N$, where $N = n + 1$ and $z_N = (x_N, y_N)$ is the true test observation, is typical under IID, then the maximal value of the prediction function (6) will be $\operatorname{lb} N$. This is again a manifestation of the fundamental limitation of conformal prediction already mentioned in the previous section.

Two other results stated in [55] extended relations between IID and exchangeability discussed in the previous section to the case of a general observation space $\mathbf{Z}$; it turned out that Theorems 1 and 2 of [45] behave very differently when the assumption $\mathbf{Z} = \{0, 1\}$ is dropped. Theorem 1 carries over to the case of general $\mathbf{Z}$ without any problems. On the other hand, a chasm between IID and exchangeability opens up when $\mathbf{Z}$ is large (or even infinite, which is allowed in [55]); namely,

$$\sup_{\zeta \in \mathbf{Z}^N} D^{\mathrm{pR}}(\wr\zeta\int) \geq^+ \sup_{\zeta \in \mathbf{Z}^N} D^{\mathrm{eR}}(\wr\zeta\int) \geq^+ N \operatorname{lb} \mathrm{e} - \frac{1}{2} \operatorname{lb} N \tag{8}$$

[55, Theorem 4], where $D^{\mathrm{pR}}(\wr \cdot \int)$ is defined analogously to (3). The difference between IID and exchangeability deficiencies now dwarfs the best p-deficiency of $\operatorname{lb} N$ that can be used for prediction (the fundamental limitation of conformal prediction; cf. (10) below). Formally, we did not allow $|\mathbf{Z}| = \infty$ (just for simplicity of definitions), but it is sufficient to assume that $|\mathbf{Z}| \geq N$. There are no proofs in [55], but the argument in the proof of [50, Theorem 2, (16)] also proves (8).

In principle, the vast difference (8) per se does not necessarily imply that IID and exchangeability are so very different: *a priori*, both can be very large, much greater than the difference. However, we can complement (8) by

$$\sup_{\zeta \in \mathbf{Z}^N} D^{\mathrm{pX}}(\wr \zeta \int) =^+ \sup_{\zeta \in \mathbf{Z}^N} D^{\mathrm{eX}}(\wr \zeta \int) =^+ 0,$$

where $D^{\mathrm{pX}}(\wr \cdots \int)$ and $D^{\mathrm{eX}}(\wr \cdots \int)$ are also defined analogously to (3). Therefore, exchangeability deficiency can be small while IID deficiency is large. A specific example of a data sequence demonstrating this is an algorithmically random permutation of $1, \dots, N$; while it is perfectly exchangeable, it does not look IID at all: given N, its IID p- and e-deficiency is

$$\operatorname{lb} \frac{N^N}{N!} \sim N \operatorname{lb} \mathrm{e} - \frac{1}{2} \operatorname{lb} N.$$

5 Nouretdinov et al.'s discovery: universality of conformal prediction

In the previous section, we saw that the difference between IID and exchangeability deficiencies can be very large. Does it mean that, under the IID assumption, we can achieve much more than what can be achieved by conformal prediction, which only relies on exchangeability? An important discovery by Nouretdinov, V'yugin, and Alex [29] was that conformal prediction is universal: we do not lose much even under IID when using conformal prediction. (I was among the authors of early versions of this paper, but at some point Volodya V'yugin's exposition became too technical for me, and I switched to other projects. The final version of the paper is still very generous about my contribution.)

To discuss the universality of conformal prediction under IID, it is useful to distinguish between two sides of our prediction problem. For concreteness, let us talk about IID p-deficiency D^{pR}.

- If the true data sequence $z_1, \dots, z_n, x_{n+1}, y_{n+1}$ looks IID, i.e.,

  $$D^{\mathrm{pR}}(z_1, \dots, z_n, x_{n+1}, y_{n+1})$$

  is small, we are in the situation of *prediction proper*; we can output y_{n+1} as a confident prediction for the label of the test object x_{n+1} if

  $$D^{\mathrm{pR}}(z_1, \dots, z_n, x_{n+1}, y)$$

is large for all false labels y.

- If the true data sequence $z_1, \dots, z_n, x_{n+1}, y_{n+1}$ does not look IID, i.e., $D^{\mathrm{pR}}(z_1, \dots, z_n, x_{n+1}, y_{n+1})$ is large, we are in the situation of *anomaly detection*; in this case all of $D^{\mathrm{pR}}(z_1, \dots, z_n, x_{n+1}, y)$, $y \in \mathbf{Y}$, can be expected to be large.

Nouretdinov et al. were interested in prediction proper, which is the most natural setting of the prediction problem. While the vast difference between IID and exchangeability might well show in anomaly detection, it does not have to show in prediction proper. In this terminology, the remark in [55, Remark 5] mentioned in the previous section says that, in the situation of prediction proper, the maximal value of the prediction function is $\mathrm{lb}\, N$ (in the case of binary classification; remember that $N := n + 1$). Now at least we have a rough coincidence of the upper bounds: what can be achieved under IID (namely, deficiency of $\mathrm{lb}\, N$) can also be achieved already by conformal prediction ($\mathrm{lb}\, N$ is allowed by its fundamental limitation). This coincidence hints at the universality of conformal prediction, but Nouretdinov et al. [29] paint a much fuller picture.

An e-test E for exchangeability or IID is said to be *train-invariant* if, for all n and for all data sequences $(z_1, \dots, z_n, z_{n+1}) \in \mathbf{Z}^{n+1}$,

$$E(z_1, \dots, z_n, z_{n+1}) = E(z_{\sigma(1)}, \dots, z_{\sigma(n)}, z_{n+1})$$

for all permutations σ of $\{1, \dots, n\}$. In this definition, $z_1, \dots, z_n$ is interpreted as training sequence and z_{n+1} as test observation. If such an E is used as predictor (e.g., replacing D^{pX} in (6) by E), we can refer to $\lbag z_1, \dots, z_n \rbag$ as *training bag*, or, colloquially, as *training set*, which is a standard expression in machine learning; E does not depend on the ordering of the bag. In the same way we define train-invariant p-tests for exchangeability. (Nouretdinov et al. used the expression "invariant" for our "train-invariant", but in this paper we will use "invariant" in a different, much narrower, sense.)

The first result reported in [29], their Proposition 1, is Ilia Nouretdinov's observation that the class of conformal predictors (understood to be functions producing conformal p-values for all possible labels $y \in \mathbf{Y}$ for the test object) essentially coincides with the class of train-invariant p-tests for exchangeability. Namely, each function in the former class is dominated (in the sense of being less than or equal to) by some function in the latter class, and vice versa.

It is also easy to check that the class of train-invariant e-tests for exchangeability essentially coincides, in the same sense, with the class of conformal e-predictors as defined in [51]. (The only difference is that "dominates" means "is greater than or equal to" in the case of e-tests.)

There exist a universal train-invariant p-test for exchangeability, a universal train-invariant e-test for exchangeability, a universal train-invariant p-test for IID, and a universal train-invariant e-test for IID. We fix such tests and denote their binary logarithms (with the sign reversed in the case of the p-tests) by D^{ptX}, D^{etX}, D^{ptR}, and D^{etR}, respectively.

Remark 7. Nouretdinov et al. [29, Sect. 4.2] used the expression "i-test" rather than "e-test", and I had used "i-values" for "e-values" earlier in [49, Sect. 5]. When working on [62], I misremembered "i-" as "e-". In hindsight, "e-" (standing for "expectation") appears to be a better counterpart of "p-" (standing for "probability" in the context of p-values) than "i-" (standing for "integral").

The following theorem is the main result of [29] (Theorem 2, slightly simplified). In it, n ranges over $\mathbb{N}$, $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ as before, $(z_1, \dots, z_n)$ (training sequence) ranges over $\mathbf{Z}^n$, (x_{n+1}, y_{n+1}) (test observation) over $\mathbf{Z}$, and y (possible labels of x_{n+1}) over $\mathbf{Y}$.

Theorem 2. *Letting $\zeta := (z_1, \dots, z_n)$ stand for the training sequence, we have*

$$D^{\mathrm{pR}}(\zeta, x_{n+1}, y) - 2 \operatorname{lb} D^{\mathrm{pR}}(\zeta, x_{n+1}, y) - 4 D^{\mathrm{pR}}(\zeta, x_{n+1}, y_{n+1}) - 4 \operatorname{lb} |\mathbf{Y}| \\ \le^+ D^{\mathrm{ptX}}(\zeta, x_{n+1}, y) \le^+ D^{\mathrm{pR}}(\zeta, x_{n+1}, y). \quad (9)$$

Theorem 2 is Nouretdinov et al.'s statement of universality for conformal prediction in classification problems. By classification I mean, informally, prediction with a small number $|\mathbf{Y}|$ of classes. In this case and in the situation of prediction proper (i.e., $D^{\mathrm{pR}}(\zeta, x_{n+1}, y_{n+1})$ also being small), (9) implies that

$$D^{\mathrm{ptX}}(\zeta, x_{n+1}, y) \approx D^{\mathrm{pR}}(\zeta, x_{n+1}, y),$$

i.e., ideal conformal prediction is almost as efficient as ideal prediction under IID.

Since $D^{\mathrm{pR}} \approx D^{\mathrm{eR}}$ and $D^{\mathrm{ptX}} \approx D^{\mathrm{etX}}$ (with the approximate equalities holding to within logarithmic terms), (9) also holds, perhaps with the coefficients 2 and 4 replaced by larger ones, for D^{eR} and D^{etX} in place of D^{pR} and D^{ptX}. In other words, conformal e-prediction is also universal in classification problems.

Theorem 1 connecting IID and exchangeability was an important component of the proof of Theorem 2 in [29] (that component was stated there as Proposition 7). The role of this connection will be clearly seen in the functional version of Nouretdinov et al.'s result stated in the following section.

Now we can discuss properly the fundamental limitation of conformal prediction and its significance. Because of the upper limit of $1/(n+1)$ on conformal p-values, we have

$$D^{\mathrm{etX}}(\zeta, x_{n+1}, y) \le^+ D^{\mathrm{ptX}}(\zeta, x_{n+1}, y) \le^+ \operatorname{lb}(n+1) = \operatorname{lb} N, \quad (10)$$

where $\zeta := (z_1, \dots, z_n)$. In the situation of classification and prediction proper, $D^{\mathrm{pR}}(\zeta, x_{n+1}, y_{n+1}) =^+ 0$, (9) implies

$$D^{\mathrm{pR}}(\zeta, x_{n+1}, y) \le^+ \operatorname{lb} N + O(\operatorname{lb} \operatorname{lb} N).$$

This continues to hold with D^{eR} in place of D^{pR}. Therefore, the fundamental limitation of conformal prediction is also a limitation of prediction proper under IID in general classification problems (not necessarily binary classification, as in [55, Remark 5]).

Remark 8. The inequalities "$\le^+$" in (10) are, of course, tight. The most confident prediction can be made where, e.g., the training sequence is $(0,\dots,0)$ (n zeros for a large n); then the confidence with which we can predict that the test label is 0 as well is reflected in the large value of

$$D^{\mathrm{etX}}(0,\dots,0,1) =^+ D^{\mathrm{ptX}}(0,\dots,0,1) =^+ \mathrm{lb}(n+1).$$

6 Perspective of the functional theory of randomness

As already mentioned, most of the groundbreaking results in [29] (all but Proposition 1) involve unspecified constants. The goal of this section is to explain how the functional theory of randomness makes those results more practical: instead of dealing with functions defined to within an additive constant, now we are dealing with inclusions and other relations between various function classes. Very few proofs will be given, and most of them can be found in [54] (which also covers the case of regression, while this paper is constrained to classification, similarly to [29]). Otherwise, this section is more detailed and self-contained than the previous ones.

In one respect, the setting of this section is simpler than our setting so far; since unspecified constants are gone, the observation space $\mathbf{Z}$ and the length of the training sequence n do not need to vary explicitly. Our setting can also be made more general for free; since the theory of algorithms is also gone, now we just assume that the object space $\mathbf{X}$ is a non-empty measurable space. However, we are still interested in the classification problem, where the label space $\mathbf{Y}$ is finite with $|\mathbf{Y}| \ge 2$ and equipped with the discrete σ-algebra. The observation space $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ is then also a measurable space. In informal explanations, I will assume that $\mathbf{Y}$ is a small set, such as in the case of binary classification $|\mathbf{Y}| = 2$ (it might be a good idea for the reader to concentrate on this case, at least at first). Now both $\mathbf{Z}$ and the length n of the training sequence $z_1,\dots,z_n$ can be fixed throughout the section. Given a new test object x_{n+1}, our task is to predict x_{n+1}'s label y_{n+1}. We will be interested in "confidence predictors", i.e., algorithms for this prediction problem producing valid measures of confidence, such as p-values or e-values. While the notion of confidence predictor is informal, later I will give formal definitions of several classes of confidence predictors.

6.1 Eight function classes

To translate Nouretdinov et al.'s results into the functional theory of randomness, it is useful to introduce eight function classes representing eight kinds of confidence predictors based on three dichotomies:

- the assumption about the data-generating mechanism can be IID (R) or exchangeability (X);

- with each potential label of a test object we can associate its p-value or e-value;
- optionally, we can require the train-invariance (abbreviated to "t") of the confidence predictor.

The combination X/p/t corresponds to the conformal predictors, while the combinations R/p and R/e correspond to the most general confidence predictors under the IID assumption. Following [29], one of our goals will be to establish the closeness of the conformal predictors (i.e., X/p/t predictors) to the R/p predictors; this goal is attractive since confidence predictors based on p-values enjoy a more intuitive property of validity than those based on e-values. Our argument will also establish the closeness of the conformal e-predictors (i.e., X/e/t predictors) to the R/e predictors. Such closeness can be interpreted as the universality of conformal prediction and conformal e-prediction. We will also consider simplified versions of these goals.

In the rest of this section we will explore

- the difference between IID and exchangeability predictors in Sect. 6.2 (and these results will be summarized in Corollary 1 and simplified in Corollary 2),
- the effect of imposing the requirement of train-invariance in Sect. 6.3 (summarized in Theorem 6),
- and the difference between confidence predictors based on p-values and those based on e-values in Sect. 6.4.

The overall picture will be summarized in Corollary 3 and simplified in Corollary 4, both in Sect. 6.4.

My informal explanations will sometimes be couched in the language of the "naive theory of randomness" postulating the existence of the largest or smallest, as appropriate, element in each function class; this element will be called "universal". Even though formally self-contradictory, this postulate makes some intuitive sense along the lines of the algorithmic theory of randomness. (To make statements of the naive theory of randomness more palatable, it may sometimes be helpful to qualify them using words such as "almost", but not in this paper.) In this way, instead of using the algorithmic theory of randomness for both formal analysis and intuitive considerations, we may use the functional theory of randomness for the former and the naive theory of randomness for the latter.

Figure 3 shows the eight function classes as a cube in each panel. Let us concentrate on its left panel for now ignoring the right one. We start from the function class in the top left corner (of the exterior square), $\mathcal{P}^{\mathrm{R}}$. It consists of all *IID p-variables* on $\mathbf{Z}^{n+1}$, i.e., functions $P : \mathbf{Z}^{n+1} \to [0,1]$ such that, for all IID probability measures $R = Q^{n+1}$ on $\mathbf{Z}^{n+1}$, we have (1). (By default all functions referred to as "variables" are assumed to be measurable.) The importance of the class $\mathcal{P}^{\mathrm{R}}$ stems from IID being the standard assumption of machine learning.

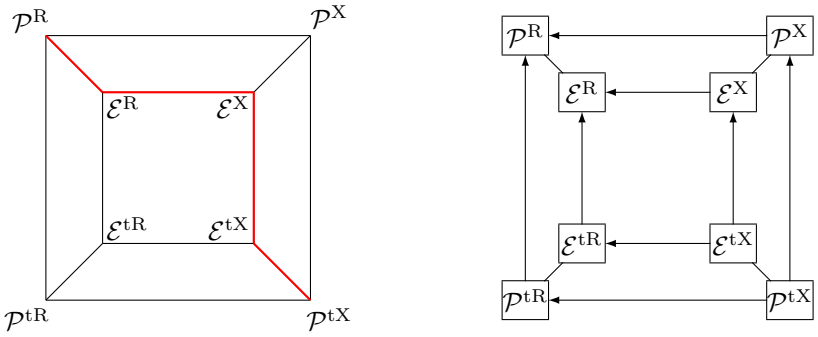

Figure 3: A cube representing eight function classes. The polygonal chain $\mathcal{P}^{\mathrm{R}}$–$\mathcal{E}^{\mathrm{R}}$–$\mathcal{E}^{\mathrm{X}}$–$\mathcal{E}^{\mathrm{tX}}$–$\mathcal{P}^{\mathrm{tX}}$ is shown in red in the left panel.

The p-variable P can be used as a "confidence transducer", in the terminology of [56, Sect. 2.7.1]. Given a training sequence $z_1, \dots, z_n$ and a test object x_{n+1}, we can compute the p-value $P(z_1, \dots, z_n, x_{n+1}, y)$ for each possible label y for x_{n+1} (as before, p-values and e-values are just values taken by p-variables and e-variables, respectively). We can regard $P(z_1, \dots, z_n, x_{n+1}, \cdot)$ to be a fuzzy set predictor for y_{n+1}. To obtain a crisp set predictor, we can choose a *significance level* $\epsilon \in (0, 1)$ and define the prediction set

$$\Gamma^\epsilon(z_1, \dots, z_n, x_{n+1}) := \{y \in \mathbf{Y} : P(z_1, \dots, z_n, x_{n+1}, y) > \epsilon\} \tag{11}$$

by thresholding (cf. (7)). By the definition of p-variables, the probability of error (meaning $y_{n+1} \notin \Gamma^\epsilon(z_1, \dots, z_n, x_{n+1})$) for this crisp set predictor will not exceed ϵ.

In [56, Sect. 2.1.6] confidence predictors were defined as nested families Γ^ϵ, $\epsilon \in (0, 1)$, and were called *conservatively valid* if Γ^ϵ makes an error with probability at most ϵ. This includes the families defined by (11) as a subclass, and in general the inclusion is proper. However, the difference is not essential, as spelled out in Propositions 2.14 and 2.15 of [56]. We will refer to the IID p-variables $P : \mathbf{Z}^{n+1} \to [0, 1]$ as *IID p-predictors* (or, more fully, IID confidence p-predictors).

The top right corner ("Kolmogorov's corner") in Fig. 3, $\mathcal{P}^{\mathrm{X}}$, is the class that consists of all *exchangeability p-variables* on $\mathbf{Z}^{n+1}$, i.e., functions $P : \mathbf{Z}^{n+1} \to [0, 1]$ satisfying (1) for all $\epsilon \in (0, 1)$ and all exchangeable probability measures R on $\mathbf{Z}^{n+1}$. Such p-variables serve as *exchangeability p-predictors*. Naively, a data sequence $\zeta \in \mathbf{Z}^{n+1}$ is exchangeable (resp. IID) if $U(\zeta)$ is not small, U being the universal exchangeability (resp. IID) p-variable.

The bottom left corner of Fig. 3, $\mathcal{P}^{\mathrm{tR}}$, is the class of the train-invariant IID p-variables (elements of $\mathcal{P}^{\mathrm{R}}$), and its bottom right corner $\mathcal{P}^{\mathrm{tX}}$ is the class of the train-invariant exchangeability p-variables.

The top left corner $\mathcal{E}^{\mathrm{R}}$ of the interior square in Fig. 3 consists of all *IID e-variables* on $\mathbf{Z}^{n+1}$, i.e., functions $E : \mathbf{Z}^{n+1} \to [0, \infty]$ such that, for all IID

probability measures R on $\mathbf{Z}^{n+1}$,

$$\int E \, \mathrm{d}R \le 1 \tag{12}$$

(which generalizes (2)). By Markov's inequality, $1/\mathcal{E}^{\mathrm{R}} \subseteq \mathcal{P}^{\mathrm{R}}$ (where $1/\mathcal{E}^{\mathrm{R}}$ consists of all $1/E$, $E \in \mathcal{E}^{\mathrm{R}}$). We will also refer to e-variables $E \in \mathcal{E}^{\mathrm{R}}$ as *IID e-predictors*. For a given training sequence $z_1, \dots, z_n$ and test object x_{n+1}, the e-value $E(z_1, \dots, z_n, x_{n+1}, y)$ for each $y \in \mathbf{Y}$ tells us how unlikely y is as label y for x_{n+1}. Therefore, $E(z_1, \dots, z_n, x_{n+1}, \cdot)$ is again a soft set predictor for y_{n+1}.

The other function classes in Fig. 3 are defined in a similar way; $\mathcal{E}^{\mathrm{X}}$ consists of all *exchangeability e-variables* on $\mathbf{Z}^{n+1}$, i.e., functions $E : \mathbf{Z}^{n+1} \to [0, \infty]$ satisfying (12) for all exchangeable R. Finally, $\mathcal{E}^{\mathrm{tR}}$ and $\mathcal{E}^{\mathrm{tX}}$ consist of all train-invariant functions in $\mathcal{E}^{\mathrm{R}}$ and $\mathcal{E}^{\mathrm{X}}$, respectively. As before, these e-variables may be referred to as e-predictors, depending on context. The right panel of Fig. 3 shows all inclusions between our eight classes, with an arrow $A \to B$ from A to B meaning $A \subseteq B$.

It is interesting that all four confidence predictors on the right of the cubes in Fig. 3 have names (either existing or trivial modifications of existing) containing the word "conformal":

- $\mathcal{P}^{\mathrm{X}}$ (the exchangeability p-predictors) are the weak conformal predictors [56, Sect. 2.2.8 and Proposition 2.9];
- $\mathcal{E}^{\mathrm{X}}$ (the exchangeability e-predictors) are the weak conformal e-predictors;
- $\mathcal{P}^{\mathrm{tX}}$ (the train-invariant exchangeability p-predictors) are the conformal predictors [56, Proposition 2.9];
- $\mathcal{E}^{\mathrm{tX}}$ (the train-invariant exchangeability e-predictors) are the conformal e-predictors (see Sect. 6.3 below).

As already mentioned, the equivalence of $\mathcal{P}^{\mathrm{tX}}$ and conformal prediction was first established in [29, Proposition 1].

Returning to the very informal language of the naive theory of randomness, our discussion of calibration in Sect. 6.4 will show that IID data sequences $\zeta \in \mathbf{Z}^{n+1}$ can be defined as those for which $U(\zeta)$ is not large, U being the universal (i.e., largest in this context) IID e-variable. Similarly, exchangeable data sequences ζ can be defined as those for which $U(\zeta)$ is not large, U being the universal (largest) exchangeability e-variable.

Following [29], we will connect two opposite vertices of the cube in the left panel of Fig. 3, $\mathcal{P}^{\mathrm{R}}$ (IID p-prediction) and $\mathcal{P}^{\mathrm{tX}}$ (conformal prediction). These vertices are important since $\mathcal{P}^{\mathrm{R}}$ corresponds to most general confidence prediction under the standard assumption of machine learning and $\mathcal{P}^{\mathrm{tX}}$ is understood very well and has been widely implemented (see, e.g., [8] and [10]).

A convenient path connecting $\mathcal{P}^{\mathrm{R}}$ and $\mathcal{P}^{\mathrm{tX}}$ is shown as the bold red polygonal chain in the left panel of Fig. 3. It will be used in stating the closeness of $\mathcal{P}^{\mathrm{R}}$ and $\mathcal{P}^{\mathrm{tX}}$ considered as predictors (Corollary 3 below, analogous to Nouretdinov

et al.'s main result). Namely, we will establish the closeness for each step in the path separately:

- The step from $\mathcal{P}^{\mathrm{R}}$ to $\mathcal{E}^{\mathrm{R}}$ (from p-values to e-values for IID) is the calibration step, to be discussed in Sect. 6.4.
- The step from $\mathcal{E}^{\mathrm{R}}$ to $\mathcal{E}^{\mathrm{X}}$ (from IID to exchangeability) is the key one; we will call it *Kolmogorov's step*. It is the topic of Sect. 6.2.
- The step from $\mathcal{E}^{\mathrm{X}}$ to $\mathcal{E}^{\mathrm{tX}}$ (adding train-invariance) is easier (if we do not worry about its optimality). We will call it the *train-invariance step*. It is discussed in Sect. 6.3.
- The step from $\mathcal{E}^{\mathrm{tX}}$ (conformal e-prediction) to $\mathcal{P}^{\mathrm{tX}}$ (conformal prediction) is the e-to-p calibration step, and it is also one of the topics of Sect. 6.4.

However, we will also be interested in other edges of the cube in the left panel of Fig. 3, first of all the edge connecting $\mathcal{P}^{\mathrm{tR}}$ and $\mathcal{P}^{\mathrm{tX}}$. A useful connection between these two classes will be obtained as a by-product (Corollary 4).

6.2 Kolmogorov's step

In principle, besides the eight function classes shown in Fig. 3, we are also interested in the following two:

- the class $\mathcal{E}^{\mathrm{iR}}$ consisting of *invariant IID e-variables*: $E \in \mathcal{E}^{\mathrm{iR}}$ if $E \in \mathcal{E}^{\mathrm{R}}$ and E is invariant w.r. to all permutations of its arguments;
- the analogous class $\mathcal{P}^{\mathrm{iR}}$ consisting of *invariant IID p-variables*, which is the class of all $P \in \mathcal{P}^{\mathrm{R}}$ that are invariant w.r. to all permutations of their arguments.

In this paper we will only use $\mathcal{E}^{\mathrm{iR}}$. When $E \in \mathcal{E}^{\mathrm{iR}}$ is chosen in advance and $E(z_1, \dots, z_{n+1})$ is large for the realized data sequence $z_1, \dots, z_{n+1}$, we are entitled to reject the hypothesis that its configuration $\wr z_1, \dots, z_{n+1} \int$ was generated in the IID manner. Therefore, $\mathcal{E}^{\mathrm{iR}}$ is the analogue of $D^{\mathrm{eR}}(\wr \cdot \int)$ in the functional theory of randomness (and $\mathcal{P}^{\mathrm{iR}}$ is the analogue of $D^{\mathrm{pR}}(\wr \cdot \int)$).

The following theorem is the functional version of the relation (4) between IID and exchangeability.

Theorem 3. *The class $\mathcal{E}^{\mathrm{R}}$ is the pointwise product of $\mathcal{E}^{\mathrm{X}}$ and $\mathcal{E}^{\mathrm{iR}}$:*

$$\mathcal{E}^{\mathrm{R}} = \mathcal{E}^{\mathrm{X}} \mathcal{E}^{\mathrm{iR}} . \tag{13}$$

The pointwise product of function classes $\mathcal{E}_1$ and $\mathcal{E}_2$ is defined as the class of all products $E_1 E_2$ for $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$, where $E_1 E_2$ is the pointwise product of functions, $(E_1 E_2)(\zeta) := E_1(\zeta) E_2(\zeta)$. For a proof of Theorem 3, see [50, Corollary 3]. However, since this result is derived in [50] as a corollary of a much more general statement [50, Theorem 1] and in order to make the exposition more self-contained, let me give a simple independent derivation.

Proof of Theorem 3. We consider the probability space $\mathbf{Z}^{n+1}$ equipped with an IID probability measure R and consider Z_i, $i = 1, \dots, n+1$, to be z_i regarded as a random observation. Formally, Z_i is the random element on that probability space defined by $Z_i(z_1, \dots, z_{n+1}) := z_i$. The expectation symbol $\mathbb{E} = \mathbb{E}_R$ refers to this probability space.

To prove the inclusion "$\subseteq$" in (13), let $E \in \mathcal{E}^{\mathrm{R}}$. Set

$$F(z_1, \dots, z_{n+1}) := \frac{1}{(n+1)!} \sum_{\pi} E(z_{\pi(1)}, \dots, z_{\pi(n+1)}),$$
$$E'(z_1, \dots, z_{n+1}) := \frac{E(z_1, \dots, z_{n+1})}{F(z_1, \dots, z_{n+1})}$$

(with $0/0 := 1$), π ranging over the permutations of $\{1, \dots, n+1\}$. It is obvious that $E' \in \mathcal{E}^{\mathrm{X}}$, and it is also easy to check that $F \in \mathcal{E}^{\mathrm{iR}}$:

$$\mathbb{E}(F(Z_1, \dots, Z_{n+1})) = \frac{1}{(n+1)!} \sum_{\pi} \mathbb{E}(E(Z_{\pi(1)}, \dots, Z_{\pi(n+1)})) \le \frac{1}{(n+1)!} \sum_{\pi} 1 = 1$$

(the inequality uses the fact that $Z_{\pi(1)}, \dots, Z_{\pi(n+1)}$ are IID).

To prove the inclusion "$\supseteq$" in (13), let $E \in \mathcal{E}^{\mathrm{X}}$ and $F \in \mathcal{E}^{\mathrm{iR}}$. Let us check that their product is in $\mathcal{E}^{\mathrm{R}}$:

$$\begin{aligned}
&\mathbb{E}\left(E(Z_1, \dots, Z_{n+1}) F(Z_1, \dots, Z_{n+1})\right) \\
&\quad = \mathbb{E}\left(\mathbb{E}\left(E(Z_1, \dots, Z_{n+1}) F(Z_1, \dots, Z_{n+1}) \mid \mathcal{G}\right)\right) \\
&\quad = \mathbb{E}\left(F(Z_1, \dots, Z_{n+1}) \mathbb{E}\left(E(Z_1, \dots, Z_{n+1}) \mid \mathcal{G}\right)\right) \\
&\quad \le \mathbb{E}\left(F(Z_1, \dots, Z_{n+1})\right) \le 1,
\end{aligned}$$

where $\mathcal{G}$ is the bag σ-algebra as defined in [56, Sect. A.5.2]; the first inequality follows from [56, Lemma A.3]. □

In terms of the naive theory of randomness, (13) implies that the universal IID e-variable is the product of the universal exchangeability e-variable and the universal invariant IID e-variable. This shows that the difference between being IID and exchangeability lies in the configuration being IID. Therefore, the following theorem establishes a connection between IID e-predictors and exchangeability e-predictors.

Theorem 4. *For each invariant IID e-variable F there exists an IID e-variable G such that, for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \ne y_{n+1}$,*

$$G(z_1, \dots, z_n, z_{n+1}) \ge \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} F(z_1, \dots, z_n, x_{n+1}, y). \tag{14}$$

We apply this theorem (see Corollary 1 below) in the context where $z_1, \dots, z_n, z_{n+1}$ is the true data sequence with z_{n+1} being the test observation

and y is a false label; ideally such y should be excluded by our confidence predictor. If $z_1, \dots, z_{n+1}$ is IID and F is the universal invariant IID e-variable, $F(z_1, \dots, x_{n+1}, y)$ will be small, and so there will be little difference between the degrees to which a false label for the test object will be rejected by the universal e-predictors under IID and exchangeability.

Let me give an informal argument why $\wr z_1, \dots, z_n, x_{n+1}, y \int$ not being IID for $y \neq y_{n+1}$ implies the true data sequence

$$(z_1, \dots, z_n, x_{n+1}, y_{n+1})$$

not being IID either. Consider, for simplicity, the case of binary labels. If after flipping the last label in the true data sequence $(z_1, \dots, z_n, x_{n+1}, y_{n+1})$ the bag of its elements becomes non-IID, then either already the original bag $\wr z_1, \dots, z_n, x_{n+1}, y_{n+1} \int$ was non-IID or the last element (x_{n+1}, y_{n+1}) was special in the true data sequence, and in any case already the original data sequence was non-IID. A formal proof is given in [54].

The following asymptotic result says that the $|\mathbf{Y}|$ in the denominator of (14) is in some sense optimal (provided it is large enough).

Theorem 5. *For each constant $c > 1$ the following statement holds true for a sufficiently large $|\mathbf{Y}|$ and a sufficiently large n. There exists an invariant IID e-variable F such that for each IID e-variable G there exist $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$ such that*

$$G(z_1, \dots, z_n, z_{n+1}) < \frac{c}{\mathrm{e}\,|\mathbf{Y}|} F(z_1, \dots, z_n, x_{n+1}, y).$$

Theorem 5 is proved in [54]. The idea of the proof can be explained informally using the algorithmic theory of randomness (or even more informally using the naive theory of randomness): we can make the label y in the bag $\wr z_1, \dots, z_n, x_{n+1}, y \int$ encode the bag $\wr y_1, \dots, y_n \int$ of the other labels; if we also make y easily distinguishable from the other labels, the value $F(z_1, \dots, z_n, x_{n+1}, y)$ of the universal invariant IID e-variable will be large.

Let us now state explicitly a corollary of Theorems 3 and 4 that expresses the universality of weak conformal e-prediction.

Corollary 1. *For each IID e-predictor E there exist an exchangeability e-predictor E' and an IID e-variable G such that, for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$,*

$$E'(z_1, \dots, z_n, x_{n+1}, y) \geq \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G(z_1, \dots, z_n, z_{n+1})}. \tag{15}$$

The informal interpretation of (15) is that, in classification, every false label y for the test object is excluded by an exchangeability e-predictor once it is excluded by an IID e-predictor, unless the true data sequence $(z_1, \dots, z_{n+1})$ is not IID. For this interpretation, there is no need to take E, E', and G universal.

Proof of Corollary 1. Let E be an IID e-variable. By Theorem 3, there exist an exchangeability e-variable E' and an invariant IID e-variable E'' such that

$$E(z_1, \dots, z_n, x_{n+1}, y) = E'(z_1, \dots, z_n, x_{n+1}, y) E''(z_1, \dots, z_n, x_{n+1}, y) \quad (16)$$

for all $z_1, \dots, z_n, x_{n+1}, y$. By Theorem 4 there exists an IID e-variable G such that

$$G(z_1, \dots, z_n, z_{n+1}) \geq \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} E''(z_1, \dots, z_n, x_{n+1}, y) \quad (17)$$

for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$. It remains to combine (16) and (17). □

Remark 9. In Corollary 1 it is possible to have, in principle, 0 in the denominator in (15). Our interpretation of an inequality $A \geq c\frac{B}{C}$, where A, c, B, C are all nonnegative, covering the possibility of $C = 0$ is that it is equivalent, by definition, to $AC \geq cB$. Similar remarks can be made about other results, such as Theorem 6 below.

An important variation on Corollary 1 is where the original IID e-predictor E is already train-invariant. Under the IID assumption, it seems useless to consider predictors that are not train-invariant, and indeed the requirement of train-invariance follows [54, Sect. 2] from fundamental statistical principles, such as the sufficiency principle and the invariance principle. In this case the resulting predictor E' will also be train-invariant and, therefore, a conformal predictor.

Corollary 2. *For each train-invariant IID e-predictor E there exist a conformal e-predictor E' and an IID e-variable G such that, for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$, we have* (15).

Corollary 2 is a statement of universality of conformal e-prediction under train-invariance.

6.3 Train-invariance step

Let us say that $G = G(z_1, \dots, z_n \mid z_{n+1})$ is a *test-conditional exchangeability e-variable* (given the test observation) if

$$\forall (z_1, \dots, z_{n+1}) : \frac{1}{n!} \sum_{\sigma} G\left(z_{\sigma(1)}, \dots, z_{\sigma(n)} \mid z_{n+1}\right) \leq 1,$$

σ ranging over the permutations of $\{1, \dots, n\}$. This property implies $G \in \mathcal{E}^{\mathrm{X}}$. If $G = G(z_1, \dots, z_n \mid z_{n+1})$ is large for the universal test-conditional exchangeability e-variable G, the sequence $z_1, \dots, z_n$ is not exchangeable given z_{n+1}. (And G being large is a stronger property than $z_1, \dots, z_{n+1}$ not being exchangeable.)

For any exchangeability e-predictor E, define the corresponding train-invariant exchangeability e-predictor $\bar{E}$ by

$$\bar{E}(z_1, \dots, z_n, z_{n+1}) := \frac{1}{n!} \sum_{\sigma} E\left(z_{\sigma(1)}, \dots, z_{\sigma(n)}, z_{n+1}\right),$$

σ again ranging over the permutations of $\{1, \dots, n\}$. This is the train-invariance step. The following theorem says that $\bar{E}$ is almost as good as E in our prediction problem unless $z_1, \dots, z_{n+1}$ is not exchangeable.

Theorem 6. *For each exchangeability e-predictor E there exists a test-conditional exchangeability e-variable G such that, for all $z_1, \dots, z_n$, all $z_{n+1} = (x_{n+1}, y_{n+1})$, and all $y \neq y_{n+1}$,*

$$\bar{E}(z_1, \dots, z_n, x_{n+1}, y) \geq \frac{1}{|\mathbf{Y}| - 1} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G(z_1, \dots, z_n \mid z_{n+1})}.$$

The intuition behind Theorem 6 is that each exchangeability e-predictor can be made train-invariant without significant loss of efficiency in classification proper. For a simple proof, see [54].

6.4 Other steps

In previous sections we discussed the two interior red steps shown in the left panel of Fig. 3. Here we will discuss the two other red steps and summarize the overall picture obtaining a version of [29, Theorem 2] in the functional theory of randomness. These two steps are analogues of the inequalities (5) in the functional theory of randomness.

Conversion from p-values to e-values (*calibration*) and vice versa (*e-to-p calibration*) is understood very well: see, e.g., [62, Sect. 2]. E-to-p calibration is particularly simple: there is one optimal e-to-p-calibrator, $e \mapsto \min(1/e, 1)$ [62, Proposition 2.2]. As for calibration, a decreasing function $f : [0, 1] \to [0, \infty]$ is a *calibrator* (transforms p-values into e-values) if and only if $\int_0^1 f \leq 1$ [62, Proposition 2.1]. We will use the calibrator

$$f(p) := \delta p^{\delta - 1} \tag{18}$$

for a fixed value $\delta \in (0, 1)$. If δ is small, $f(p)$ will be close to $1/p$ if we ignore the multiplicative constant (as customary in the algorithmic theory of randomness). Other popular calibrators are

$$f(p) := \begin{cases} \infty & \text{if } p = 0 \\ \kappa(1 + \kappa)^\kappa p^{-1} (-\ln p)^{-1-\kappa} & \text{if } p \in (0, \exp(-1 - \kappa)] \\ 0 & \text{if } p \in (\exp(-1 - \kappa), 1] \end{cases}$$

for a constant $\kappa > 0$ (see [62, Appendix B]; this calibrator is even closer to $1/p$ than (18) with a small δ) and Shafer's [34, Sect. 3, (6)] calibrator

$$f(p) := p^{-1/2} - 1.$$

The lines between the corresponding $\mathcal{P}$ and $\mathcal{E}$ vertices in the right panel of Fig. 3 stand for the possibility of calibration or e-to-p calibration (similarly to the double-headed arrow in Fig. 2).

The following result combines all the previous statements in this section.

Corollary 3. *Let $\delta \in (0,1)$. For all $P \in \mathcal{P}^{\mathrm{R}}$ there exist $P' \in \mathcal{P}^{\mathrm{tX}}$ and $G \in \mathcal{E}^{\mathrm{R}}$ such that, for all observations $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and labels $y \neq y_{n+1}$,*

$$P'(z_1, \dots, z_n, x_{n+1}, y) \leq \frac{\mathrm{e}(|\mathbf{Y}| - 1)^2}{\delta} G(z_1, \dots, z_{n+1})^2 P(z_1, \dots, z_n, x_{n+1}, y)^{1-\delta}. \quad (19)$$

Corollary 3 reduces (as usual, imperfectly) IID p-predictors to conformal predictors. It says that in classification problems every false label excluded by an IID p-predictor is excluded by a conformal predictor (perhaps less strongly) unless the true data sequence is non-IID. It is an analogue of [29, Theorem 2].

Proof of Corollary 3. Let $P \in \mathcal{P}^{\mathrm{R}}$ and $\delta \in (0,1)$. We will construct $P' \in \mathcal{P}^{\mathrm{tX}}$ and $G \in \mathcal{E}^{\mathrm{R}}$ satisfying (19) in several steps. Since (18) is a calibrator, there is $E \in \mathcal{E}^{\mathrm{R}}$ satisfying

$$E(z_1, \dots, z_n, x_{n+1}, y) \geq \delta P(z_1, \dots, z_n, x_{n+1}, y)^{\delta - 1} \quad (20)$$

(in fact, with "=" in place of "≥"); here and in the rest of the proof we will leave "for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$" implicit. By Corollary 1, there exist $E' \in \mathcal{E}^{\mathrm{X}}$ and $G_1 \in \mathcal{E}^{\mathrm{R}}$ such that

$$E'(z_1, \dots, z_n, x_{n+1}, y) \geq \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G_1(z_1, \dots, z_n, z_{n+1})}. \quad (21)$$

By Theorem 6, there exist $E'' \in \mathcal{E}^{\mathrm{tX}}$ and $G_2 \in \mathcal{E}^{\mathrm{R}}$ such that

$$E''(z_1, \dots, z_n, x_{n+1}, y) \geq \frac{1}{|\mathbf{Y}| - 1} \frac{E'(z_1, \dots, z_n, x_{n+1}, y)}{G_2(z_1, \dots, z_n, z_{n+1})}. \quad (22)$$

Finally, since $e \mapsto \min(1/e, 1)$ is an e-to-p calibrator, there is $P' \in \mathcal{P}^{\mathrm{tX}}$ satisfying

$$P'(z_1, \dots, z_n, x_{n+1}, y) \leq 1/E''(z_1, \dots, z_n, x_{n+1}, y). \quad (23)$$

It remains to combine (20)–(23) and set $G := \sqrt{G_1 G_2}$. (By the inequality between the geometric and arithmetic means, $G \in \mathcal{E}^{\mathrm{R}}$.) □

Of course, Corollary 3 continues to hold if the condition $P \in \mathcal{P}^{\mathrm{R}}$ is replaced by $P \in \mathcal{P}^{\mathrm{X}}$. In this case, however, we can drop step (21) and replace (19) by

$$P'(z_1, \dots, z_n, x_{n+1}, y) \leq \frac{|\mathbf{Y}| - 1}{\delta} G(z_1, \dots, z_{n+1}) P(z_1, \dots, z_n, x_{n+1}, y)^{1-\delta}.$$

The most important case, however, is where $P \in \mathcal{P}^{\mathrm{tR}}$. Now we can drop step (22), as spelled out in the following corollary.

Corollary 4. *Let $\delta \in (0,1)$. For each $P \in \mathcal{P}^{\mathrm{tR}}$ there exist $P' \in \mathcal{P}^{\mathrm{tX}}$ and $G \in \mathcal{E}^{\mathrm{R}}$ such that, for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$,*

$$P'(z_1, \dots, z_n, x_{n+1}, y) \leq \frac{\mathrm{e}(|\mathbf{Y}| - 1)}{\delta} G(z_1, \dots, z_{n+1}) P(z_1, \dots, z_n, x_{n+1}, y)^{1-\delta}.$$

Corollary 4 reduces train-invariant IID p-prediction to conformal prediction without significant loss in efficiency. Since the condition of train-invariance is so natural under the IID assumption, Corollaries 2 and 4 may be the most useful statements in this section.

7 Conclusion

This paper further develops the functional theory of randomness in the direction of Nouretdinov, V'yugin, and Gammerman's work on universality of conformal prediction. While the statements of the functional theory of randomness may be less intuitive than those of the algorithmic theory of randomness, they are more precise avoiding unspecified constants and are simpler in an important respect: e.g., the analogue of (13) in the algorithmic theory of randomness, (4), involves the condition "$| \; D^{\mathrm{eR}}(⌈\zeta⌋)$", which disappears in (13). In the naive theory of randomness we could write, instead,

$$D^{\mathrm{eR}}(\zeta) = D^{\mathrm{eX}}(\zeta) + D^{\mathrm{eR}}(⌈\zeta⌋).$$

For the reader familiar with the algorithmic theory of complexity, the condition "$| \; D^{\mathrm{eR}}(⌈\zeta⌋)$" is analogous to the second entry of "$K(x)$" in the Levin–Chaitin formula

$$K(x, y) =^+ K(x) + K(y \mid x, K(x))$$

[38, Theorem 67]. Conditions of this type can be removed in the functional theory of randomness and functional theory of complexity (the latter introduced in [50, Appendix]).

In Sect. 2 I emphasized the narrowness of both IID and exchangeability assumptions, whereas this paper concentrates on what can be achieved under IID. There have been many developments in conformal prediction beyond exchangeability, such as those in [3, 6, 19, 40]; see also the review [1, Chap. 7]. For all of them, exchangeability serves as a starting point.

Acknowledgments

Many thanks to Alex Gammerman for sharing his recollections. I am also grateful to Ruodu Wang, Irina Shevtsova, Nicolo Colombo, and Alexander Shen for their advice. As always, the Stack Exchange TEX–LATEX community have been ready to help.

References

[1] Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. Technical Report arXiv:2411.11824 [math.ST], arXiv.org e-Print archive, June 2025. Pre-publication version of a book to be published by Cambridge University Press.

[2] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

[3] Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

[4] Eugene A. Asarin. Some properties of Kolmogorov Δ-random finite sequences. *Theory of Probability and its Applications*, 32:507–508, 1987. Russian original: О некоторых свойствах Δ-случайных по Колмогорову конечных последовательностей.

[5] Eugene A. Asarin. On some properties of finite objects random in the algorithmic sense. *Soviet Mathematics Doklady*, 36:109–112, 1988. Russian original: О некоторых свойствах случайных в алгоритмическом смысле конечных объектов, published in 1987.

[6] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51:816–845, 2023.

[7] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713.

[8] Henrik Boström. Conformal prediction in Python with crepes. *Proceedings of Machine Learning Research*, 230:236–249, 2024. COPA 2024.

[9] British Standards Institution. *Quantities and units, Part 2: Mathematical signs and symbols to be used in the natural sciences and technology*, 2010. BS ISO 80000-2:2009.

[10] Thibault Cordier, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, and Nicolas Brunel. Flexible and systematic uncertainty estimation with conformal prediction via the MAPIE library. *Proceedings of Machine Learning Research*, 204:549–581, 2023. COPA 2023.

[11] Andrew I. Dale. A study of some early investigations into exchangeability. *Historia Mathematica*, 12:323–336, 1985.

[12] A. Philip Dawid and Vladimir Vovk. Prequential probability: Principles and properties. *Bernoulli*, 5:125–162, 1999.

[13] Persi Diaconis. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36:271–281, 1977.

[14] Persi Diaconis and David A. Freedman. Finite exchangeable sequences. *Annals of Probability*, 8:745–764, 1980.

[15] Peter Gács. Exact expressions for some randomness tests. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 26:385–394, 1980.

[16] Alex Gammerman, Vladimir Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 148–155, San Francisco, CA, 1998. Morgan Kaufmann.

[17] Alexander Gammerman. *Machine Learning: Progress and Prospects. An Inaugural Lecture by Alexander Gammerman, Professor of Computer Science. Presented at Royal Holloway, University of London, on 11th December 1996*. Royal Holloway, University of London, Egham, Surrey, 1997.

[18] Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning (with discussion). *Computer Journal*, 50:151–177, 2007.

[19] Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[20] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing (with discussion). *Journal of the Royal Statistical Society B*, 86:1091–1171, 2024.

[21] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. English translation: *Foundations of the Theory of Probability*. Chelsea, New York, 1950.

[22] Andrei N. Kolmogorov. Теория вероятностей. In *Математика, ее содержание, методы и значение*, volume 2, pages 252–284. Издательство АН СССР, Moscow, 1956.

[23] Andrei N. Kolmogorov. On tables of random numbers. *Sankhyā. Indian Journal of Statistics A*, 25:369–376, 1963.

[24] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965. Russian original: Три подхода к определению понятия "количество информации".

[25] Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, IT-14:662–664, 1968. Russian original: К логическим основам теории информации и теории вероятностей, published in Проблемы передачи информации, 1969.

[26] Andrei N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983. Russian original: Комбинаторные основания теории информации и исчисления вероятностей.

[27] Andrei N. Kolmogorov and Vladimir A. Uspensky. Algorithms and randomness. *Theory of Probability and Its Applications*, 32:389–412, 1987. Russian original: Алгоритмы и случайность.

[28] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[29] Ilia Nouretdinov, Vladimir V'yugin, and Alex Gammerman. Transductive Confidence Machine is universal. In Ricard Gavaldà, Klaus P. Jantke, and Eiji Takimoto, editors, *Proceedings of the Fourteenth International Conference on Algorithmic Learning Theory*, volume 2842 of *Lecture Notes in Artificial Intelligence*, pages 283–297, Berlin, 2003. Springer.

[30] Gleb Novikov. Relations between randomness deficiencies. Technical Report arXiv:1608.08246 [math.LO], arXiv.org e-Print archive, August 2016. Published in *Lecture Notes in Computer Science* 10307:338–350 (2017).

[31] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. Technical Report arXiv:2410.23614 [math.ST], arXiv.org e-Print archive, May 2025. Book draft.

[32] Craig Saunders, Alex Gammerman, and Vladimir Vovk. Transduction with confidence and credibility. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 722–726, San Francisco, CA, 1999. Morgan Kaufmann.

[33] Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

[34] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society A*, 184:407–478, 2021.

[35] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[36] Glenn Shafer and Vladimir Vovk. The sources of Kolmogorov's *Grundbegriffe*. *Statistical Science*, 21:70–98, 2006. Extended version: arXiv:1802.06071 [math.HO].

[37] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ, 2019.

[38] Alexander Shen, Vladimir A. Uspensky, and Nikolai Vereshchagin. *Kolmogorov Complexity and Algorithmic Randomness*. American Mathematical Society, Providence, RI, 2017.

[39] Albert N. Shiryaev. *Probability-1*. Springer, New York, third edition, 2016.

[40] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

[41] Vladimir A. Uspensky and Alexei L. Semenov. *Algorithms: Main Ideas and Applications*. Kluwer, Dordrecht, 1993.

[42] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[43] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.

[44] Richard von Mises. *Wahrscheinlichkeit, Statistik, und Wahrheit*. Springer, Berlin, 1928. English translation: *Probability, Statistics and Truth*. William Hodge, London (1939).

[45] Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986. Russian original: О понятии бернуллиевости. Another English translation with proofs: arXiv:1612.08859 (math.ST).

[46] Vladimir Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society B*, 55:317–351, 1993.

[47] Vladimir Vovk. Minimum description length estimators under the optimal coding scheme. In Paul Vitányi, editor, *Computational Learning Theory*, volume 904 of *Lecture Notes in Computer Science*, pages 237–251, Berlin, 1995. Springer. EuroCOLT 1995.

[48] Vladimir Vovk. Learning about the parameter of the Bernoulli model. *Journal of Computer and System Sciences*, 55:96–104, 1997. EuroCOLT 1995 Special Issue. This paper is the journal version of [47].

[49] Vladimir Vovk. Kolmogorov's complexity conception of probability. In Vincent F. Hendricks, Stig Andur Pedersen, and Klaus Frovin Jørgensen, editors, *Probability Theory: Philosophy, Recent History and Relations to Science*, pages 51–69. Kluwer, Dordrecht, 2001.

[50] Vladimir Vovk. Non-algorithmic theory of randomness. In Andreas Blass, Patrick Cégielski, Nachum Dershowitz, Manfred Droste, and Berndt Finkbeiner, editors, *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*, volume 12180 of *Lecture Notes in Computer Science*, pages 323–340, Cham, 2020. Springer.

[51] Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, 166:article 111674, 2025. Special Issue on Conformal Prediction and Distribution-Free Uncertainty Quantification.

[52] Vladimir Vovk. Inductive randomness predictors: beyond randomness. Technical Report arXiv:2503.02803 [cs.LG], arXiv.org e-Print archive, July 2025. Conference version: COPA 2025.

[53] Vladimir Vovk. Testing exchangeability in the batch mode with e-values and Markov alternatives. *Machine Learning*, 114:article 99, 2025. Special Issue on Conformal Prediction and Distribution-Free Uncertainty Quantification.

[54] Vladimir Vovk. Universality of conformal prediction under the assumption of randomness. Technical Report arXiv:2502.19254 [cs.LG], arXiv.org e-Print archive, June 2025.

[55] Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, San Francisco, CA, 1999. Morgan Kaufmann.

[56] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.

[57] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *Annals of Statistics*, 37:1566–1590, 2009.

[58] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Conformal e-testing. *Pattern Recognition*, 168:article 111841, 2025. Special Issue on Conformal Prediction and Distribution-Free Uncertainty Quantification.

[59] Vladimir Vovk and Glenn Shafer. Kolmogorov's contributions to the foundations of probability. *Problems of Information Transmission*, 39:21–31, 2003.

[60] Vladimir Vovk and Glenn Shafer. A conversation with A. Philip Dawid. *Statistical Science*, 40:148–166, 2025.

[61] Vladimir Vovk and Vladimir V. V'yugin. On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society B*, 55:253–266, 1993.

[62] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49:1736–1754, 2021.

[63] Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *Statistical Science*, 38:329–354, 2023.