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 February 12, 2025.

Project web site:
http://alrw.net

Abstract

This paper continues development of the functional theory of randomness, a modification of the algorithmic theory of randomness getting rid of unspecified additive constants. It introduces new kinds of confidence predictors, including randomness predictors (the most general confidence predictors based on the assumption of IID observations) and exchangeability predictors (the most general confidence predictors based on the assumption of exchangeable observations). The main result implies that both are close to conformal predictors and quantifies the difference between randomness prediction and conformal prediction.

Contents

1 Introduction

The functional theory of randomness was proposed in Vovk [2020] under the name of non-algorithmic theory of randomness. The algorithmic theory of randomness was started by Kolmogorov in the 1960s [Kolmogorov, 1968] and has been developed in numerous papers and books (see, e.g., Shen et al. 2017). It has been a powerful source of intuition, but its weakness is the dependence on the choice of a specific universal partial computable function, which leads to the presence of unspecified additive (sometimes multiplicative) constants in its mathematical results. Kolmogorov [1965, 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 leads to the astronomical multiplicative constant of 2^{100} in the corresponding p-value).

The way of dealing with unspecified constants proposed in Vovk [2020] is to express statements of the algorithmic theory of randomness as relations between various function classes. It will be introduced in Sect. 2. In this paper we will call this approach the functional theory of randomness. While it loses somewhat in intuitive simplicity, it moves closer to practical machine learning and statistics. No formal knowledge of the algorithmic theory of randomness will be assumed on the part of the reader.

In this paper we are interested in applying the functional theory of randomness to prediction. The most standard assumption in machine learning is that of *randomness*: we assume that the observations are generated in the IID fashion (are independent and identically distributed). An *a priori* weaker assumption is that of exchangeability, although for infinite data sequences randomness and exchangeability turn out to be essentially equivalent by the celebrated de Finetti representation theorem. For finite sequences, however, the difference is important, and it will be the subject of our Sect. 3.

We start discussing applications of the functional theory of randomness to prediction in Sect. 2. In it we introduce the notion of a confidence predictor (slightly modifying and generalizing the terminology of Vovk et al. 2022, Sect. 2.1.6). Then we identify eight kinds of confidence predictors based on three dichotomies:

- the assumption about the data-generating mechanism can be randomness (R) or exchangeability (X);
- with each potential label of a test object we can associate its p-value (as in standard statistical hypothesis testing) or e-value (see, e.g., Vovk and Wang 2021 or Grünwald et al. 2024);
- optionally, we can require the invariance (i) of the confidence predictor w.r. to the permutations of the training sequence (the ubiquitous expression "training set" suggests such invariance).

The combination X/p/i corresponds to the conformal predictors [Angelopoulos et al., 2024], while the combinations R/p and R/e correspond to the most general confidence predictors under the assumption of randomness. Our main goal will be to establish the closeness of the X/p/i predictors (i.e., conformal predictors) to the R/p predictors, since confidence predictors based on p-values enjoy a more intuitive property of validity. (The close relation between the conformal predictors and the R/e predictors will be established as a by-product.) Such closeness can be interpreted as the universality of conformal prediction, and it was explored in Nouretdinov et al. [2003], which can be regarded as counterpart of this paper in the algorithmic theory of randomness.

In the rest of this paper we will explore

- the difference between randomness and exchangeability predictors in Sect. 3 (and our results in this section will be summarized in Corollary 5),
- the effect of imposing the requirement of invariance w.r. to the permutations of the training sequence in Sect. 4 (summarized in Corollary 8),
- and the difference between confidence predictors based on p-values and those based on e-values in Sect. 5.

The overall picture will be summarized in our main result, Theorem 10 in Sect. 5. Section 6 concludes with some directions of further research.

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, I will sometimes qualify them using the word "almost".

In this paper e may stand for an e-value; Euler's number (the base of the natural logarithms) is $\mathrm{e} \approx 2.72$. The exponential function is $x \mapsto \exp(x)$ or $x \mapsto \mathrm{e}^x$ (but, e.g., $\mathrm{e}(n+1)$ is a product, the same thing as $(n+1)\mathrm{e}$).

2 Confidence predictors

In this paper we are interested in the following prediction problem. After having observed a training sequence of *examples* $z_i = (x_i, y_i)$, $i = 1, \dots, n$, each consisting of an *object* x_i and its *label* y_i, and given a new test object x_{n+1}, our task is to predict x_{n+1}'s label y_{n+1}. The length n of the training sequence is fixed throughout. We will say that y_{n+1} is the *true label* of x_{n+1} while labels $y \neq y_{n+1}$ are *false*.

Each object is drawn from a measurable space $\mathbf{X}$ (the *object space*), and each label is drawn from a measurable space $\mathbf{Y}$ (the *label space*); both $\mathbf{X}$ and $\mathbf{Y}$ are non-empty. Therefore, the examples are drawn from the Cartesian product $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ (the *example space*). We are mostly interested in the case of

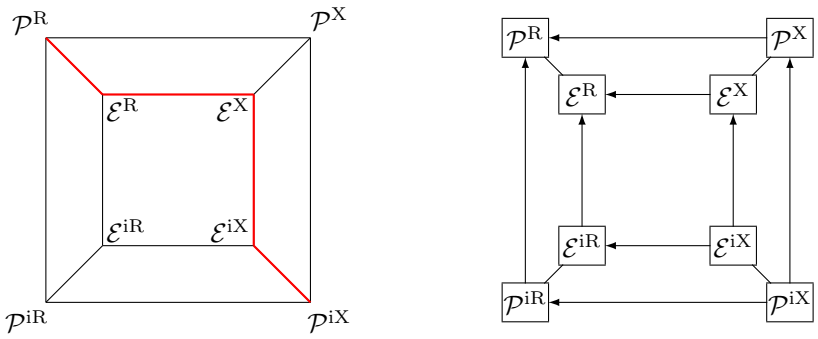

Figure 1: A cube representing eight function classes. The polygonal chain $\mathcal{P}^{\mathrm{R}} \to \mathcal{E}^{\mathrm{R}} \to \mathcal{E}^{\mathrm{X}} \to \mathcal{E}^{\mathrm{iX}} \to \mathcal{P}^{\mathrm{iX}}$ is shown in red in the left panel.

classification, where the label space $\mathbf{Y}$ is finite (equipped with the discrete σ-algebra) with $|\mathbf{Y}| \geq 2$. 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).

Remark 1. In our book Vovk et al. [2022] we referred to prediction algorithms in this setting as "one-off", dropping "one-off" when n was allowed to vary. In this paper, n is always fixed and "one-off" is never used.

We will be interested in two statistical assumptions about the data sequence $z_1, \dots, z_{n+1}$, where $z_{n+1} := (x_{n+1}, y_{n+1})$ is the test example. Under the *assumption of randomness*, the probability measure generating $z_1, \dots, z_{n+1}$ is a *power probability measure* $R = Q^{n+1}$, Q being a probability measure on the example space $\mathbf{Z}$. Under the weaker *assumption of exchangeability*, the probability measure R generating $z_1, \dots, z_{n+1}$ is invariant under permutations of $z_1, \dots, z_{n+1}$.

Figure 1 shows eight function classes that we are interested in in this paper. 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 *randomness p-variables* on $\mathbf{Z}^{n+1}$, i.e., functions $P : \mathbf{Z}^{n+1} \to [0,1]$ such that, for all $\epsilon \in (0,1)$ and all power probability measures R on $\mathbf{Z}^{n+1}$,

$$R(P \leq \epsilon) \leq \epsilon. \tag{1}$$

(By default all functions referred to as "variables" are assumed to be measurable.) The importance of the class $\mathcal{P}^{\mathrm{R}}$ stems from randomness being the standard assumption of machine learning.

In the very informal setting of the naive theory of randomness, the universal randomness p-variable U is the smallest randomness p-variable, and we may call a data sequence $\zeta \in \mathbf{Z}^{n+1}$ random (to a varying degree) if $U(\zeta)$ is not small. We can consider two possible prediction situations:

- if the true data sequence $z_1, \dots, z_{n+1}$ is random, 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 $U(z_1, \dots, z_n, x_{n+1}, y)$ is small for all false labels y;

- if the true data sequence $z_1, \dots, z_{n+1}$ is not random, we are in the situation of "anomaly detection"; in this case all of $U(z_1, \dots, z_n, x_{n+1}, y)$, $y \in \mathbf{Y}$, can be expected to be small.

In this paper we are mainly interested in prediction proper.

The p-variable P can be used as a "confidence transducer", in the terminology of Vovk et al. [2022, 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} (where a *p-value* is just a value taken by a p-variable). 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

$$\Gamma^\epsilon(z_1, \dots, z_n, x_{n+1}) := \{y \in \mathbf{Y} : P(z_1, \dots, z_n, x_{n+1}, y) > \epsilon\} \tag{2}$$

by thresholding. 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 Vovk et al. [2022, 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 (2) 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 Vovk et al. [2022]. We will refer to the randomness p-variables $P : \mathbf{Z}^{n+1} \to [0, 1]$ as *randomness predictors* (or, more fully, randomness confidence predictors).

The top right corner in Fig. 1, $\mathcal{P}^{\mathrm{X}}$, 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 predictors*. Naively, a data sequence $\zeta \in \mathbf{Z}^{n+1}$ is exchangeable if $U(\zeta)$ is not small, U being the universal exchangeability p-variable.

The bottom left corner in Fig. 1, $\mathcal{P}^{\mathrm{iR}}$, is the class of the randomness p-variables (elements of $\mathcal{P}^{\mathrm{R}}$) that are invariant w.r. to the permutations of the training sequence. The bottom right corner $\mathcal{P}^{\mathrm{iX}}$ in Fig. 1 is the class of the exchangeability p-variables that are invariant w.r. to the permutations of the training sequence.

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

$$\int E \, \mathrm{d}R \le 1. \tag{3}$$

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 *randomness 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. 1 are defined in a similar way: $\mathcal{E}^{\mathrm{X}}$ consists of all *exchangeability e-variables* on $\mathbf{Z}^{n+1}$, i.e., functions $E : \mathbf{Z}^{n+1} \to [0, 1]$ satisfying (3) for all exchangeable R. Finally, $\mathcal{E}^{\mathrm{iR}}$ and $\mathcal{E}^{\mathrm{iX}}$ consist of all invariant (w.r. to the permutations of the training sequence) functions in $\mathcal{E}^{\mathrm{R}}$ and $\mathcal{E}^{\mathrm{X}}$, respectively. As before, these e-variables may be referred as e-predictors, depending on context. The right panel of Fig. 1 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 (e-)predictors on the right of the cubes in Fig. 1 have names (either existing or trivial modifications of existing) containing the word "conformal":

- $\mathcal{P}^{\mathrm{X}}$ (the exchangeability p-variables) are the weak conformal predictors [Vovk et al., 2022, Sect. 2.2.8 and Proposition 2.9];
- $\mathcal{E}^{\mathrm{X}}$ (the exchangeability e-variables) are the weak conformal e-predictors;
- $\mathcal{P}^{\mathrm{iX}}$ (the invariant exchangeability p-variables) are the conformal predictors [Vovk et al., 2022, Proposition 2.9];
- $\mathcal{E}^{\mathrm{iX}}$ (the invariant exchangeability e-variables) are the conformal e-predictors (see Sect. 4 below).

The equivalence of $\mathcal{P}^{\mathrm{iX}}$ and conformal prediction was first established by Nouretdinov et al. [2003, Proposition 1].

Returning to the very informal language of the naive theory of randomness, our discussion of calibration in Sect. 5 will show that random 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) randomness 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.

We are interested in connecting two opposite vertices of the cube in the left panel of Fig. 1, $\mathcal{P}^{\mathrm{R}}$ (confidence prediction under randomness) and $\mathcal{P}^{\mathrm{iX}}$ (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 $\mathcal{P}^{\mathrm{iX}}$ is understood very well and has been widely implemented (see, e.g., Boström 2024 and Cordier et al. 2023).

A convenient path connecting $\mathcal{P}^{\mathrm{R}}$ and $\mathcal{P}^{\mathrm{iX}}$ is shown as the bold red polygonal chain in the left panel of Fig. 1. It will be used in proving our main result about the closeness of $\mathcal{P}^{\mathrm{R}}$ and $\mathcal{P}^{\mathrm{iX}}$ considered as predictors (Theorem 10 below). Namely, we will prove the closeness for the start and end of 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 randomness) is the calibration step, to be discussed in Sect. 5.
- The step from $\mathcal{E}^{\mathrm{R}}$ to $\mathcal{E}^{\mathrm{X}}$ (from randomness to exchangeability) is the key one; we will call it *Kolmogorov's step*. It is the topic of Sect. 3.

- The step from $\mathcal{E}^{\mathrm{X}}$ to $\mathcal{E}^{\mathrm{iX}}$ (adding invariance) is easier (if we do not worry about its optimality). We will call it the *invariance step*. It is discussed in Sect. 4.
- The step from $\mathcal{E}^{\mathrm{iX}}$ (conformal e-prediction) to $\mathcal{P}^{\mathrm{iX}}$ (conformal prediction) is the e-to-p calibration step, and it is also one of the topics of Sect. 5.

3 Kolmogorov's step

For infinite data sequences and assuming that the example space is standard Borel, each exchangeable probability measure is a convex mixture of power probability measures (see, e.g., Schervish 1995, Theorem 1.49), and so the difference between the assumptions of randomness and exchangeability simply disappears. For finite data sequences the difference exists, but still randomness was formalized as exchangeability by Kolmogorov (see, e.g., Kolmogorov 1968, Sect. 2). The difference was first explored in Vovk [1986, Theorem 1], and this result was translated into the language of the functional theory of randomness in Vovk [2020, Corollary 3]. According to those results, the difference between the randomness and exchangeability for a data sequence $\zeta \in \mathbf{Z}^{n+1}$ lies in the randomness of the *configuration* $\mathrm{conf}(\zeta)$ of ζ, where $\mathrm{conf}(\zeta)$ is the bag (i.e., multiset) obtained from ζ by "forgetting the order of its elements" (if $\zeta = (z_1, \dots, z_N)$, then $\mathrm{conf}(\zeta) = \lbag z_1, \dots, z_N \rbag$, $\lbag \cdots \rbag$ being the notation for bags).

In principle, besides the eight function classes shown in Fig. 1, we are also interested in the following two:

- the class $\mathcal{E}^{\mathrm{cR}}$ consisting of *randomness e-variables for configurations* (or configuration randomness e-variables): $E \in \mathcal{E}^{\mathrm{cR}}$ if $E : \mathbf{Z}^{(n+1)} \to [0, \infty]$ (we are using the notation $\mathbf{Z}^{(N)}$ for the set of all bags of N examples) satisfies
$$\int E(\mathrm{conf}(\zeta))\, R(\mathrm{d}\zeta) \le 1$$
for all power probability measures R on $\mathbf{Z}^{n+1}$;
- the analogous class $\mathcal{P}^{\mathrm{cR}}$ consisting of *randomness p-variables for configurations*, which is the class of $P : \mathbf{Z}^{(n+1)} \to [0, 1]$ satisfying
$$R(\{\zeta \in \mathbf{Z}^{n+1} : P(\mathrm{conf}(\zeta)) \le \epsilon\}) \le \epsilon$$
for all power probability measures R on $\mathbf{Z}^{n+1}$ and all $\epsilon > 0$.

In this paper, however, we will only use $\mathcal{E}^{\mathrm{cR}}$.

Proposition 2. *The class $\mathcal{E}^{\mathrm{R}}$ is the pointwise product of $\mathcal{E}^{\mathrm{X}}$ and another class, namely $\mathcal{E}^{\mathrm{cR}} \circ \mathrm{conf}$:*
$$\mathcal{E}^{\mathrm{R}} = \mathcal{E}^{\mathrm{X}}(\mathcal{E}^{\mathrm{cR}} \circ \mathrm{conf}). \tag{4}$$

The pointwise product of function classes $\mathcal{E}_1$ and $\mathcal{E}_2$ is defined as the class of all products E_1E_2 for $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$, where E_1E_2 is the pointwise product of functions, $(E_1E_2)(\zeta) := E_1(\zeta)E_2(\zeta)$. For a proof of Proposition 2, see Vovk [2020, Corollary 3] or Sect. A.3 below; since this result is derived in Vovk [2020] as a corollary of a much more general statement [Vovk, 2020, Theorem 1] and in order to make the exposition self-contained, I will give a simple independent derivation.

According to Proposition 2, the difference between randomness and exchangeability lies in the randomness of the configuration. Therefore, the following theorem establishes a connection between randomness e-predictors and exchangeability e-predictors.

Theorem 3. *For each randomness e-variable F for configurations there exists a randomness e-variable G such that, for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$,*

$$G(z_1, \dots, z_n, z_{n+1}) \geq \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} F(\{z_1, \dots, z_n, x_{n+1}, y\}). \tag{5}$$

In applications of this theorem later in the paper, $z_1, \dots, z_n, z_{n+1}$ will be the true data sequence with z_{n+1} being the test example; y will be a false label, and ideally it should be excluded by our confidence predictor.

Let me give an informal argument why nonrandomness of $\{z_1, \dots, z_n, x_{n+1}, y\}$ for $y \neq y_{n+1}$ implies nonrandomness of the true data sequence

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

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 nonrandom, then either the original bag $\{z_1, \dots, z_{n+1}\}$ was nonrandom or the last element z_{n+1} was special, and in any case already the original data sequence was nonrandom. A formal proof is given in Appendix A (Sect. A.1).

The following asymptotic result says that the $|\mathbf{Y}|$ in the denominator of (5) is in some sense optimal.

Theorem 4. *For each constant $c > 1$ the following statement holds true for a sufficiently large $|\mathbf{Y}|$ and a sufficiently large n. There exists a randomness e-variable F for configurations such that for each randomness 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*

$$G(z_1, \dots, z_n, z_{n+1}) < \frac{c}{\mathrm{e}\,|\mathbf{Y}|} F(\{z_1, \dots, z_n, x_{n+1}, y\}). \tag{6}$$

When we say "for a sufficiently large $|\mathbf{Y}|$" in Theorem 4, the lower bound on $|\mathbf{Y}|$ is allowed to depend on c, and when we say "and a sufficiently large n", the lower bound on n is allowed to depend on c and $|\mathbf{Y}|$.

Theorem 4 will be proved in Appendix B. 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 $F(\wr z_1, \dots, z_n, x_{n+1}, y\int)$ of the universal randomness e-variable for configurations will be large.

In conclusion of this section, let us now state explicitly the corollary of Proposition 2 and Theorem 3 that we will need later in the paper.

Corollary 5. *For each randomness e-variable E there exist an exchangeability e-variable E' and a randomness e-variable G such that, for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$,*

$$E'(z_1, \dots, z_n, x_{n+1}, y) \geq \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G(z_1, \dots, z_n, z_{n+1})}. \tag{7}$$

The informal interpretation of (7) is that, in classification, every false label y for the test object is excluded by an exchangeability e-predictor once it is excluded by a randomness e-predictor, unless the true data sequence $(z_1, \dots, z_{n+1})$ is not random.

Proof of Corollary 5. Let E be a randomness e-variable. By Proposition 2, there exist an exchangeability e-variable E' and a randomness e-variable E'' for configurations such that

$$E(z_1, \dots, z_n, x_{n+1}, y) = E'(z_1, \dots, z_n, x_{n+1}, y) E''(\wr z_1, \dots, z_n, x_{n+1}, y\int) \tag{8}$$

for all $z_1, \dots, z_n, x_{n+1}, y$. By Theorem 3 there exists a randomness e-variable G such that

$$G(z_1, \dots, z_n, z_{n+1}) \geq \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} E''(\wr z_1, \dots, z_n, x_{n+1}, y\int) \tag{9}$$

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 (8) and (9). □

Remark 6. In Corollary 5 it is possible to have, in principle, 0 in the denominator in (7). 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 Theorem 7, Corollary 8, and Theorem 9 below (with "$\geq$" replaced by "$<$" in the case of Theorem 9).

4 Invariance step

Let us say that an exchangeability e-variable $E = E(z_1, \dots, z_n, z_{n+1})$ is *invariant* (or training invariant) if

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

for all permutations π of $\{1,\dots,n\}$. It is easy to check that an invariant exchangeability e-variable is the same thing as a conformal e-predictor as defined in Vovk [2024]; indeed, the nonconformity e-measure [Vovk, 2024, Sect. 2] corresponding to an invariant exchangeability e-variable E is

$$A(z_1,\dots,z_{n+1}) := \big(E(z_2,\dots,z_{n+1},z_1), E(z_3,\dots,z_{n+1},z_1,z_2),\dots, E(z_1,z_2,\dots,z_{n+1})\big).$$

Perhaps a more convenient definition of a nonconformity e-measure in this context would have been

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

(in the spirit of Vovk et al. 2022, Sect. 2.9.3).

We will also be interested in *conditional training exchangeability e-variables* $G = G(z_1,\dots,z_n \mid z_{n+1})$ (given the test observation). The defining property of such an e-variable is

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

π 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, 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 invariant exchangeability e-predictor $\bar{E}$ by

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

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

$$\bar{E}(z_1,\dots,z_n,x_{n+1},y) \ge \frac{1}{|\mathbf{Y}|-1}\frac{E(z_1,\dots,z_n,x_{n+1},y)}{G(z_1,\dots,z_n \mid z_{n+1})}. \tag{10}$$

For a simple proof, see Appendix A (Sect. A.2). Since each conditional training exchangeability e-variable is a randomness e-variable, we have the following corollary of Theorem 7 (which is what we will need in the next section).

Corollary 8. *For each exchangeability e-variable E there exist an invariant exchangeability e-variable E' and a randomness e-variable G such that, for all $z_1,\dots,z_n$, $z_{n+1} = (x_{n+1},y_{n+1})$, and $y \ne y_{n+1}$,*

$$E'(z_1,\dots,z_n,x_{n+1},y) \ge \frac{1}{|\mathbf{Y}|-1}\frac{E(z_1,\dots,z_n,x_{n+1},y)}{G(z_1,\dots,z_n,z_{n+1})}. \tag{11}$$

Informally, (11) says that, in classification problems, if an exchangeability e-predictor E excludes a false label y as the label of the test object, then an invariant exchangeability e-predictor E' will exclude it as well unless the true data sequence is not random.

The following result, to be proved and discussed in Appendix C, concerns the optimality of Corollary 8.

Theorem 9. *For each constant $c > 1$ the following statement holds true for a sufficiently large $|\mathbf{Y}|$ and a sufficiently large n. There exists an exchangeability e-variable E such that for all invariant exchangeability e-variables E' and all randomness e-variables G there are $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$ such that*

$$E'(z_1, \dots, z_n, x_{n+1}, y) < \frac{c\sqrt{\pi n/2}}{|\mathbf{Y}|} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G(z_1, \dots, z_n, z_{n+1})}. \tag{12}$$

This inverse statement is imperfect even asymptotically because of the factor $\sqrt{\pi n/2}$ in (12), albeit $\sqrt{\pi n/2}$ can grow much more slowly than $|\mathbf{Y}|$ (the proof in Appendix C will show that $|\mathbf{Y}|$ can grow exponentially fast in n). Inequality (12) can be made, however, asymptotically perfect if we replace "all randomness e-variables G" by "all conditional training exchangeability e-variables G", as in Theorem 7. Namely, in this case we will be able to remove the factor $\sqrt{\pi n/2}$ in (12), as explained in Sect. C.4.

5 Other steps

In previous sections we discussed the two interior red steps shown in the left panel of Fig. 1. Here we will discuss the two other red steps and summarize the overall picture.

Conversion from p-values to e-values (*calibration*) and vice versa (*e-to-p calibration*) is understood very well: see, e.g., Vovk and Wang [2021, Sect. 2]. E-to-p calibration is particularly simple: there is one optimal e-to-p-calibrator, $e \mapsto \min(1/e, 1)$ [Vovk and Wang, 2021, 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 $\int_0^1 f \leq 1$ [Vovk and Wang, 2021, Proposition 2.1]. We will use the calibrator

$$f(p) := \delta p^{\delta - 1} \tag{13}$$

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 Vovk and Wang 2021, Appendix B; this calibrator is even closer to $1/p$ than (13) with a small δ) and Shafer's [2021, 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. 1 stand for the possibility of calibration or e-to-p calibration.

The following result combines all the previous statements in this paper and can be regarded as its main result.

Theorem 10. *Let $\delta \in (0,1)$. For all $P \in \mathcal{P}^{\mathrm{R}}$ there exist $P' \in \mathcal{P}^{\mathrm{iX}}$ 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 $y \neq y_{n+1}$,*

$$P'(z_1, \dots, z_n, x_{n+1}, y) \le (\mathrm{e}/\delta)(|\mathbf{Y}| - 1)^2 \\ G(z_1, \dots, z_{n+1})^2 P(z_1, \dots, z_n, x_{n+1}, y)^{1-\delta}. \quad (14)$$

Theorem 10 reduces (as usual, imperfectly) randomness predictors to conformal predictors. It says that in classification problems every false label excluded by a randomness predictor is excluded by a conformal predictor (perhaps less strongly) unless the true data sequence is nonrandom.

Proof of Theorem 10. Let $P \in \mathcal{P}^{\mathrm{R}}$ and $\delta \in (0,1)$. We will construct $P' \in \mathcal{P}^{\mathrm{iX}}$ and $G \in \mathcal{E}^{\mathrm{R}}$ satisfying (14) in several steps. Since (13) is a calibrator, there is $E \in \mathcal{E}^{\mathrm{R}}$ satisfying

$$E(z_1, \dots, z_n, x_{n+1}, y) \ge \delta P(z_1, \dots, z_n, x_{n+1}, y)^{\delta - 1} \quad (15)$$

(in fact, with "=" in place of "$\ge$"); 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 5, there exist $E' \in \mathcal{E}^{\mathrm{X}}$ and $G_1 \in \mathcal{E}^{\mathrm{R}}$ such that

$$E'(z_1, \dots, z_n, x_{n+1}, y) \ge \frac{1}{\mathrm{e}(|\mathbf{Y}| - 1)} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G_1(z_1, \dots, z_n, z_{n+1})}. \quad (16)$$

By Corollary 8, there exist $E'' \in \mathcal{E}^{\mathrm{iX}}$ and $G_2 \in \mathcal{E}^{\mathrm{R}}$ such that

$$E''(z_1, \dots, z_n, x_{n+1}, y) \ge \frac{1}{|\mathbf{Y}| - 1} \frac{E'(z_1, \dots, z_n, x_{n+1}, y)}{G_2(z_1, \dots, z_n, z_{n+1})}. \quad (17)$$

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

$$P'(z_1, \dots, z_n, x_{n+1}, y) \le 1/E''(z_1, \dots, z_n, x_{n+1}, y). \quad (18)$$

It remains to combine (15)–(18) 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, Theorem 10 continues to hold if the condition $P \in \mathcal{P}^{\mathrm{R}}$ is replaced by $P \in \mathcal{P}^{\mathrm{X}}$ (thus justifying one of the claims made in the abstract). In this case, however, we can drop step (16) and replace (14) by

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

6 Conclusion

This paper continues study of the functional theory of randomness started in Vovk [2020]. While its statements are somewhat less intuitive than those of the algorithmic theory of randomness, they are more precise (do not involve unspecified constants) and simpler in an important respect: e.g., the analogue of (4) in the algorithmic theory of randomness is

$$d^{\mathrm{R}}(\zeta) = d^{\mathrm{X}}(\zeta \mid d^{\mathrm{cR}}(\mathrm{conf}(\zeta))) + d^{\mathrm{cR}}(\mathrm{conf}(\zeta)) + O(1)$$

(see Vovk 1986, Theorem 1), where d^{R} is deficiency of randomness of data sequences (defined in terms of prefix rather than plain Kolmogorov complexity), d^{X} is deficiency of exchangeability of data sequences, and d^{cR} is deficiency of randomness of configurations of data sequences. The condition "$\mid d^{\mathrm{cR}}(\mathrm{conf}(\zeta))$" disappears in (4).

These are the most obvious directions of further research:

- Our optimality results, such as Theorems 4 and 9, are asymptotic, and in this asymptotic setting, e.g., the difference between $|\mathbf{Y}| - 1$ (used in Theorems 3 and 7) and $|\mathbf{Y}|$ (used in Theorems 4 and 9) disappears. It would be interesting to derive optimality results in the form of explicit inequalities.

- Besides, the optimality of individual steps shown in red in Fig. 1 does not mean that the whole red path is optimal. Deriving optimal versions of the inequality (14) in Theorem 10 is another interesting direction of further research.

- It would be interesting to explore the class $\mathcal{P}^{\mathrm{iR}}$ as a possible alternative to the class $\mathcal{P}^{\mathrm{iX}}$ of conformal predictors. Are there any predictors in $\mathcal{P}^{\mathrm{iR}} \setminus \mathcal{P}^{\mathrm{iX}}$ that are more efficient, in some important respects, than any conformal predictor? Theorem 10 says that the difference is not huge, but it can still be important.

- Finally, can we prove similar universality results for conformal prediction in the case of regression ($\mathbf{Y} = \mathbb{R}$)? Our negative results suggest that in this case conformal predictors may be inefficient at some individual false labels, but it might be possible to prove interesting results about their efficiency on average or for the vast majority of false labels.

Acknowledgments

In October 2024 Vladimir V'yugin, one of my teachers in the algorithmic theory of randomness, died at the age of 75. I am deeply grateful to him for his support and encouragement from the time when I first met him, in 1980, until the end of his life. The research reported in this paper is dedicated to his memory.

Many thanks to Ruodu Wang for his comments. As always, the Stack Exchange TEX–LATEX community have been ready to help.

References

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, November 2024. Pre-publication version of a book to be published by Cambridge University Press.

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

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.

Aryeh Dvoretzky and Jacob Wolfowitz. Sums of random integers reduced modulo m. *Duke Mathematical Journal*, 18:501–507, 1951.

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

H. Burke Horton and R. Tynes Smith, III. A direct method for producing random digits in any number system. *Annals of Mathematical Statistics*, 20: 82–90, 1949.

Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.

Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, IT-14:662–664, 1968.

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.

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

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.

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

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

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

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.

Vladimir Vovk. Conformal e-prediction. Technical Report arXiv:2001.05989 [stat.ME], arXiv.org e-Print archive, November 2024.

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

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

A Proofs of Theorems 3 and 7

In this appendix we prove the main positive results of this paper. The appendix ends with an easy self-contained proof of Proposition 2.

A.1 Proof of Theorem 3

We will construct the two components of $G \in \mathcal{E}^{\mathrm{R}}$ (the exchangeability component and the configuration randomness component, as per Proposition 2) separately. To obtain an approximation G_1 to the configuration randomness component for $(z_1, \dots, z_{n+1})$, choose randomly (with equal probabilities) $i \in \{1, \dots, n+1\}$ and change the label y_i to $y \in \mathbf{Y} \setminus \{y_i\}$ randomly (with equal probabilities; y_i is the label in z_i). Then

$$G_1(\lbag z_1, \dots, z_{n+1} \rbag) := \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{|\mathbf{Y}| - 1} \sum_{y \in \mathbf{Y} \setminus \{y_i\}} F(\lbag z_1, \dots, z_{i-1}, x_i, y, z_{i+1}, \dots, z_{n+1} \rbag)$$

will be the expectation of the resulting value

$$F(\lbag z_1, \dots, z_{i-1}, x_i, y, z_{i+1}, \dots, z_{n+1} \rbag)$$

over the random choice of i and y. The configuration randomness component G_2 is the IID analogue of G_1: for each $i \in \{1, \dots, n+1\}$, with probability $1/(n+1)$ choose a random $y \neq y_i$ and change y_i to y; define $G_2(\lbag z_1, \dots, z_{n+1} \rbag)$ to be

the expectation of the value of F on the resulting bag. The exchangeability component is

$$G_3(z_1,\dots,z_{n+1}) := \frac{1}{|\mathbf{Y}|-1}\sum_{y\in\mathbf{Y}\setminus\{y_{n+1}\}}\frac{F(\lbag z_1,\dots,z_n,x_{n+1},y\rbag)}{G_1(\lbag z_1,\dots,z_{n+1}\rbag)}.$$

Define $G := (G_2\circ\mathrm{conf})G_3$; then $G\in\mathcal{E}^{\mathrm{R}}$ by Proposition 2.

Let us first check that G_2 and G_3 are really a configuration randomness e-variable and an exchangeability e-variable, respectively. To check that $G_2\in\mathcal{E}^{\mathrm{cR}}$, we need to prove, for a given power probability measure on $z_1,\dots,z_{n+1}$, that the expectation of $G_2\circ\mathrm{conf}$ does not exceed 1. It suffices to notice that if $z_1,\dots,z_{n+1}$ are generated by a power probability measure, randomly changing the label of each z_i to a different label leads to a random sequence that is still IID. And to check that $G_3\in\mathcal{E}^{\mathrm{X}}$, notice that the average of G_3 over all permutations of its arguments is 1 by the definition of G_1:

$$\begin{aligned}&\frac{1}{(n+1)!}\sum_{\pi}G_3(z_{\pi(1)},\dots,z_{\pi(n+1)})\\ &=\frac{1}{(n+1)!(|\mathbf{Y}|-1)}\sum_{i=1}^{n+1}\sum_{y\in\mathbf{Y}\setminus\{y_i\}}n!\frac{F(\lbag z_1,\dots,z_{i-1},z_{i+1},\dots,z_{n+1},x_i,y\rbag)}{G_1(\lbag z_1,\dots,z_{n+1}\rbag)}=1,\end{aligned}$$

π ranging over the permutations of $\{1,\dots,n+1\}$.

Let us now check that $G_2\ge G_1/\mathrm{e}$. The probability that randomly changing the label of each z_i to a different label leads to a change in just one label is

$$\left(1-\frac{1}{n+1}\right)^n\ge\frac{1}{\mathrm{e}}$$

(by substitution $t:=1/n$ this inequality reduces to $(1+t)\mathrm{e}^{-t}\le 1$, which can be checked by differentiation).

Combining everything said so far in this proof, we obtain, for each $y\in\mathbf{Y}\setminus\{y_{n+1}\}$,

$$\begin{aligned}G(z_1,\dots,z_{n+1}) &= G_2(\lbag z_1,\dots,z_{n+1}\rbag)G_3(z_1,\dots,z_{n+1})\\ &\ge G_2(\lbag z_1,\dots,z_{n+1}\rbag)\frac{1}{|\mathbf{Y}|-1}\frac{F(\lbag z_1,\dots,z_n,x_{n+1},y\rbag)}{G_1(\lbag z_1,\dots,z_{n+1}\rbag)}\\ &\ge\frac{1}{\mathrm{e}(|\mathbf{Y}|-1)}F(\lbag z_1,\dots,z_n,x_{n+1},y\rbag).\end{aligned}$$

A.2 Proof of Theorem 7

Let E be an exchangeability e-variable. Define a conditional training exchangeability e-variable G by

$$G(z_1,\dots,z_n\mid z_{n+1}) := \frac{1}{|\mathbf{Y}|-1}\sum_{y\in\mathbf{Y}\setminus\{y_{n+1}\}}\frac{E(z_1,\dots,z_n,x_{n+1},y)}{\bar{E}(z_1,\dots,z_n,x_{n+1},y)},\tag{19}$$

where x_{n+1} and y_{n+1} are the components of $z_{n+1} = (x_{n+1}, y_{n+1})$. Since (19) implies (10), we only need to check that (19) is a conditional training exchangeability e-variable:

$$\begin{aligned}
&\frac{1}{n!}\sum_{\pi} G(z_{\pi(1)}, \dots, z_{\pi(n)} \mid z_{n+1}) \\
&= \frac{1}{|\mathbf{Y}| - 1} \sum_{y \in \mathbf{Y} \setminus \{y_{n+1}\}} \frac{1}{n!} \sum_{\pi} \frac{E(z_{\pi(1)}, \dots, z_{\pi(n)}, x_{n+1}, y)}{\bar{E}(z_1, \dots, z_n, x_{n+1}, y)} \\
&= \frac{1}{|\mathbf{Y}| - 1} \sum_{y \in \mathbf{Y} \setminus \{y_{n+1}\}} 1 = 1,
\end{aligned} \tag{20}$$

π ranging over the permutations of $\{1, \dots, n\}$.

A.3 Proof of Proposition 2

We consider the probability space $\mathbf{Z}^{n+1}$ equipped with a power probability measure R and consider Z_i, $i = 1, \dots, n+1$, to be z_i regarded as a random example (formally, Z_i is the random element on that probability space defined by $Z_i(z_1, \dots, z_{n+1}) := z_i$).

To prove the inclusion "$\subseteq$" in (4), let $E \in \mathcal{E}^{\mathrm{R}}$. Set

$$\begin{aligned}
F(\lbag z_1, \dots, z_{n+1} \rbag) &:= \frac{1}{(n+1)!} \sum_{\pi} E(z_{\pi(1)}, \dots, z_{\pi(n+1)}) \\
E'(z_1, \dots, z_{n+1}) &:= \frac{E(z_1, \dots, z_{n+1})}{F(\lbag z_1, \dots, z_{n+1} \rbag)},
\end{aligned}$$

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 $F \in \mathcal{E}^{\mathrm{cR}}$:

$$\begin{aligned}
\mathbb{E}(F(\lbag Z_1, \dots, Z_{n+1} \rbag)) &= \frac{1}{(n+1)!} \sum_{\pi} \mathbb{E}(E(Z_{\pi(1)}, \dots, Z_{\pi(n+1)})) \\
&\le \frac{1}{(n+1)!} \sum_{\pi} 1 = 1
\end{aligned}$$

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

To prove the inclusion "$\supseteq$" in (4), let $E \in \mathcal{E}^{\mathrm{X}}$ and $F \in \mathcal{E}^{\mathrm{cR}}$. Let us check that their product is in $\mathcal{E}^{\mathrm{R}}$:

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

where $\mathcal{G}$ is the bag σ-algebra as defined in Vovk et al. [2022, Sect. A.5.2]; the first inequality follows from Vovk et al. [2022, Lemma A.3].

B Proof and discussion of Theorem 4

This appendix is mainly devoted to the proof of Theorem 4, but let me start (in Sect. B.1) from a simple informal example showing that the constant e in (5) is optimal (although its optimality is also implied by Theorem 4). Let us fix an object $x_0 \in \mathbf{X}$; it will often be sufficient in this and next appendixes to have x_0 as the only object in our data sequences (and then it will often be omitted, after a warning).

B.1 An informal example

The following informal example shows that the constant e in Theorem 3 cannot be improved. It is much simpler than the argument in the proof of Theorem 4 given in the next section.

Suppose the training set is such that

$$(y_1, \dots, y_n) = (0, \dots, 0)$$

(all the objects are x_0 and omitted, and we assume, without loss of generality, $\{0,1\} \subseteq \mathbf{Y}$); we are asked to predict y_{n+1}. The natural (universal) exchangeability e-value for the potential label $y_{n+1} = 1$ will be $n+1$ under the exchangeability model, and the natural randomness e-value for the potential label $y_{n+1} = 1$ will be close to $\mathrm{e}(n+1)$ (for large n): indeed, the maximum probability of the event

$$(y_1, \dots, y_n, y_{n+1}) = (0, \dots, 0, 1)$$

under a power probability measure is

$$\left(1 - \frac{1}{n+1}\right)^n \frac{1}{n+1} \sim \frac{1}{\mathrm{e}(n+1)}. \tag{21}$$

B.2 Proof of Theorem 4

The examples $z_1, \dots, z_{n+1}$ whose existence is asserted in the statement of the theorem and which will be constructed in this proof will all share the same object x_0, and we will sometimes omit "x_0" in our notation. Suppose, without loss of generality, that the label set $\mathbf{Y}$ is the disjoint union of $\{0,1\}$ and the set $\{-k, \dots, k\}$ for some positive integer k; to distinguish between the 0s and the 1s in these two sets, let us write $0'$ and $1'$ for the elements of the first set, $\{0,1\} = \{0', 1'\}$. (The primes are ignored, of course, when $0'$ and $1'$ are used as inputs to arithmetic operations, as in (22) below. If $|\mathbf{Y}|$ is an even number, we can leave one of its elements unused.) To avoid trivial complications, let n be an even number. This proof will assume $1 \ll k \ll \sqrt{n}$; the formal meaning of this assumption will be summarized at the end of the proof.

Define $F \in \mathcal{E}^{\mathrm{cR}}$ (which ignores the objects) as follows:

- on the bags in $\mathbf{Z}^{(n+1)}$ of the form $\wr x_1, y_1, \dots, x_n, y_n, x_{n+1}, y\int$, where $y_1, \dots, y_n \in \{0', 1'\}$, $y \in \{-k, \dots, k\}$, and

$$\sum_{i=1}^{n} y_i - n/2 = y, \tag{22}$$

  F takes value $ae\sqrt{\pi n/2}$, where $a < 1$ is a positive constant (it will be taken close to 1 later in the proof);

- F takes value 0 on all other bags in $\mathbf{Z}^{(n+1)}$.

Let us check that F is indeed a randomness e-variable for configurations. We will use the random elements $Z_1, \dots, Z_{n+1}$ on the probability space $(\mathbf{Z}^{n+1}, R)$ introduced in Sect. A.3; let $R = Q^{n+1}$ (so that individual examples are generated by Q). We will also use the notation Y_i, $i = 1, \dots, n+1$, for the label of Z_i. The maximum probability of the event $F(\wr Z_1, \dots, Z_{n+1}\int) > 0$ is attained for Q giving maximum probabilities to the following two events:

1. Exactly n of the random variables $Y_1, \dots, Y_{n+1}$ take values in $\{0', 1'\}$, and the remaining random variable Y takes value in $\{-k, \dots, k\}$.

2. Conditionally on the first event, we have (22), where y is the value taken by Y and $y_1, \dots, y_n$ are the values taken by the other n random variables.

The *Bernoulli model* is defined as $(B_\theta^{n+1} : \theta \in [0,1])$, where B_θ is the *Bernoulli measure* on $\{0,1\}$, defined by $B_\theta(\{1\}) := \theta \in [0,1]$. The maximum probability of the first event (in item 1) tends to $1/\mathrm{e}$ as $n \to \infty$ (cf. (21)); indeed, the maximum probability of the event that only one 1 is generated under the Bernoulli model is attained for $\theta := 1/(n+1)$, which is obtained by maximizing $(n+1)\theta(1-\theta)^n$ over θ. The maximum probability (conditional) of the second event (in item 2) for a given $y \in \{-k, \dots, k\}$ is asymptotically equivalent, by the local limit theorem [Shiryaev, 2016, Sect. 1.6], to

$$\frac{1}{\sqrt{2\pi n\left(\frac{1}{2} + \frac{y}{n}\right)\left(\frac{1}{2} - \frac{y}{n}\right)}} = \frac{1}{\sqrt{2\pi n\left(\frac{1}{4} - \frac{y^2}{n^2}\right)}} \sim \sqrt{\frac{2}{\pi n}};$$

this follows from the random variables $Y_1, \dots, Y_{n+1}$ with Y excluded being distributed as B_θ (conditionally on Y and its index), and the maximum over θ being attained at $\theta = 1/2 + y/n$. Since $a < 1$, $F \in \mathcal{E}^{\mathrm{cR}}$ for a sufficiently large n. Notice that our argument in this paragraph only uses $k \ll n$.

Let us now find the maximum probability that

$$\sum_{i=1}^{n} Y_i - n/2 \in \{-k, \dots, k\} \tag{23}$$

(so that this condition does not involve Z_{n+1}). Since $Y_1, \dots, Y_{n+1}$ are distributed according to the Bernoulli model, the maximum probability, again by

the local limit theorem, is asymptotically equivalent to

$$\frac{2k+1}{\sqrt{2\pi n/4}} \sim \frac{2\sqrt{2}k}{\sqrt{\pi n}};$$

it is attained at $\theta := 1/2$. Notice that now the assumption $1 \ll k \ll \sqrt{n}$ is essential in order for the exponential term in the local limit theorem to go away (the exponential term was 1 in the previous paragraph). Therefore, for any $G \in \mathcal{E}^{\mathrm{R}}$, there are $y_1, \dots, y_{n+1} \in \mathbf{Y}$ such that

$$G(y_1, \dots, y_{n+1}) \le \frac{b\sqrt{\pi n}}{2\sqrt{2}k} \tag{24}$$

(remember that we are omitting x_0), where $b > 1$ is arbitrarily close to 1, and

$$\sum_{i=1}^{n} y_i - n/2 \in \{-k, \dots, k\}$$

(cf. (23)). Fix such $y_1, \dots, y_{n+1}$ and set

$$y := \sum_{i=1}^{n} y_i - n/2$$

(cf. (22)). Taking a and b sufficiently close to 1, we obtain

$$\frac{G(y_1, \dots, y_{n+1})}{F(\wr y_1, \dots, y_n, y\int)} \le \frac{b\sqrt{\pi n}}{2\sqrt{2}ka\mathrm{e}\sqrt{\pi n/2}} < \frac{c}{\mathrm{e}\,|\mathbf{Y}|}$$

for a sufficiently large k (remember that $|\mathbf{Y}| = 2k + 3$).

Finally, the formal meaning of the condition $1 \ll k \ll \sqrt{n}$ is that the first sentence in the statement of Theorem 4 can be replaced by "For each constant $c > 1$ there is $C > 0$ such that the following statement holds true assuming $|\mathbf{Y}| \ge C$ and $\sqrt{n} \ge C\,|\mathbf{Y}|$."

B.3 Discussion of Theorem 4, its proof, and Corollary 5

One weakness of Theorem 4 is its asymptotic nature. Moreover, in the proof we used the condition $1 \ll |\mathbf{Y}| \ll \sqrt{n}$, and this condition was given in a much cruder form in the theorem's statement.

What is even more important, the construction in the proof of Theorem 4 is only applicable in the context of Corollary 5 when we would like to attain extremely high levels of confidence in our predictions. In the framework of the naive theory of randomness, let F and G in (6) be the universal randomness e-variable for configurations and the universal randomness e-variable, respectively. The construction in the proof makes y a distinguished label in the bag $\wr y_1, \dots, y_n, y\int$ (as the only label in $\{-k, \dots, k\}$). Therefore, already the universal exchangeability e-variable will take a value of almost n on the data

sequence $(y_1, \dots, y_n, y)$. Since we assume $n \gg 1$, the potential value of y for y_{n+1} will be confidently excluded even under exchangeability. Under randomness, the confidence will be even higher, and G will take a value of almost $Fn \approx \mathrm{e}\sqrt{\pi n/2}n \geq n^{3/2}$ on $(y_1, \dots, y_n, y)$. An interesting question is whether it is possible for y to be typical under exchangeability but confidently excluded under randomness. In the following section we will see that the answer is positive.

Another disadvantage of the construction in the proof of Theorem 4 is that G may take a high value, of the order of magnitude of $\sqrt{n}/k$ (see (24)), even on the true data sequence. This means that we are in a situation of anomaly detection rather than prediction proper (in the terminology of Sect. 2). This will be rectified in the following section.

B.4 An inverse to Corollary 5

Let us state an inverse to Corollary 5 in a strong form that answers the questions asked in the previous section.

Theorem 11. *For each constant $c \in (0,1)$ the following statement holds true for a sufficiently large $|\mathbf{Y}|$ and a sufficiently large n. There exists a randomness e-variable E such that for each exchangeability e-variable E' and each randomness 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*

$$G(z_1, \dots, z_n, z_{n+1}) \leq 2, \tag{25}$$

$$E'(z_1, \dots, z_n, x_{n+1}, y) \leq 2.01, \tag{26}$$

$$E(z_1, \dots, z_n, x_{n+1}, y) \geq c\,|\mathbf{Y}|\,. \tag{27}$$

In the naive interpretation of Theorem 11, the randomness e-variable E is the universal (largest) one. The worst-case choice of E' and G are also the universal elements of $\mathcal{E}^{\mathrm{X}}$ and $\mathcal{E}^{\mathrm{R}}$, respectively. Then we can see that randomness prediction can be superior to exchangeability prediction: the former can confidently exclude a label y that is not excluded by the latter, and this happens in a situation of prediction proper.

To compare Theorem 11 with Corollary 5, notice that (25)–(27) implies

$$E'(z_1, \dots, z_n, x_{n+1}, y) < \frac{4.02}{c\,|\mathbf{Y}|} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G(z_1, \dots, z_n, z_{n+1})}. \tag{28}$$

This is an inverse to (7) if we ignore the constant $4.02\mathrm{e} < 11$.

Proof of Theorem 11. Let us ignore the objects and set, without loss of generality (assuming our prediction problem is classification), $\mathbf{Y} := \{0, \dots, |\mathbf{Y}| - 1\}$. Generate labels $Y_1, \dots, Y_{n+1}$ randomly (independently from the uniform distribution on $\mathbf{Y}$), and set $Y \equiv -Y_1 - \dots - Y_n \pmod{|\mathbf{Y}|}$. Our plan is to prove that the three events

$$G(Y_1, \dots, Y_n, Y_{n+1}) \leq 2, \tag{29}$$

$$E'(Y_1,\dots,Y_n,Y) \le 2.01, \tag{30}$$
$$E(Y_1,\dots,Y_n,Y) \ge c\,|\mathbf{Y}| \tag{31}$$

(cf. (25)–(27), respectively) hold with probability at least 0.5, 0.502, and 0.999, respectively. This will imply the statement of the theorem as their intersection will be nonempty.

The probability of (29) being at least 0.5 follows immediately from Markov's inequality applied to $G \in \mathcal{E}^{\mathrm{R}}$.

Since $E' \in \mathcal{E}^{\mathrm{X}}$, the average of E' over all permutations of a given data sequence in $\mathbf{Y}^{n+1}$ is at most 1. The distribution of the random sequence $(Y_1,\dots,Y_n,Y)$ is uniform on the set of all permutations of a given data sequence in $\mathbf{Y}^{n+1}$ (since the probability of each permutation is $|\mathbf{Y}|^{-n}$ provided the labels in the data sequence sum to 0), and this implies that $\mathbb{E}(E'(Y_1,\dots,Y_n,Y)) \le 1$. The probability of (30) is at least 0.502 again by Markov's inequality.

Now let us define E:

$$E(y_1,\dots,y_{n+1}) := \begin{cases} c\,|\mathbf{Y}| & \text{if } k \equiv 0 \pmod{|\mathbf{Y}|} \text{ and } |k - n/2| \le n/6 \\ 0 & \text{otherwise,} \end{cases} \tag{32}$$

where $k := y_1 + \dots + y_{n+1}$. By the definition of E, we only need to prove that $E \in \mathcal{E}^{\mathrm{R}}$ (for a large enough n).

Let Q be a probability measure on $\mathbf{Y}$ and set $R := Q^{n+1}$. We will denote by E_1 the event given after "if" in (32), namely the conjunction of $k \equiv 0 \pmod{|\mathbf{Y}|}$ and $|k - n/2| \le n/6$. It suffices to prove that asymptotically the R-probability of E_1 does not exceed $|\mathbf{Y}|$, uniformly in Q. Consider two cases:

- If $Q(\{y\}) \in [1/6, 5/6]$ for all $y \in \mathbf{Y}$, the convergence $R(E_2) \to 1/|\mathbf{Y}|$ as $n \to \infty$ for the superset $E_2 := \{k \equiv 0 \pmod{|\mathbf{Y}|}\}$ of E_1 follows from the asymptotic uniformity result of Horton and Smith [1949, Theorem (5.01)] (proved independently by Dvoretzky and Wolfowitz 1951, Theorem 2). The convergence is uniform on this compact set.

- If $Q(\{y\}) \in [0, 1/6] \cup [5/6, 1]$ for some $y \in \mathbf{Y}$, the convergence $R(E_3) \to 0$ as $n \to \infty$ for the superset $E_3 := \{|k - n/2| \le n/6\}$ of E_1 follows from, e.g., the central limit theorem combined with the Bonferroni correction. The convergence is uniform on this compact set.

Combining the two cases, we get $E \in \mathcal{E}^{\mathrm{R}}$. □

C Optimality in the invariance step

The main part of this appendix is the proof of Theorem 9 in Sect. C.2.

C.1 Optimality of Theorem 7

We start from noticing that an inverse statement to Theorem 7 is almost trivial but still much stronger than typical inverse statements, at least in the binary

case $|\mathbf{Y}| = 2$. Let us state such an inverse to the following corollary of Theorem 7 (weakening Theorem 7 makes its inverse stronger).

Corollary 12. *For each exchangeability e-variable E there exists a randomness e-variable G such that, for all $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$,*

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

A significant weakness of the following inverse to Corollary 12 in the binary case will be discussed after its proof.

Proposition 13. *Let $|\mathbf{Y}| = 2$. For each exchangeability e-variable E and each randomness e-variable G, there are $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \neq y_{n+1}$ satisfying*

$$\bar{E}(z_1, \dots, z_n, x_{n+1}, y) \leq \frac{1}{|\mathbf{Y}| - 1} \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{G(z_1, \dots, z_n, z_{n+1})}. \tag{33}$$

Notice that Proposition 13 is an inverse to Corollary 12 for each E separately (a more standard inverse statement would start from "there exists an exchangeability e-variable E such that", analogously to Theorems 4 and 9).

Without loss of generality, we set $\mathbf{Y} := \{0, 1\}$. We can then remove "and $y \neq y_{n+1}$" in the statement of the theorem replacing y by $1 - y_{n+1}$ in (33). The fraction $\frac{1}{|\mathbf{Y}|-1}$ in (33) can also be removed, of course.

Proof of Proposition 13. Let $E \in \mathcal{E}^{\mathrm{X}}$ and $G \in \mathcal{E}^{\mathrm{R}}$. Define a randomness e-variable G^* as the right-hand side of (19). Since the expectation of G^* is 1 under any power probability measure (cf. (20)), we cannot increase G^* everywhere in such a way that it still remains a randomness e-variable. Therefore, there exists $(z_1, \dots, z_n, z_{n+1})$ on which $G \leq G^*$. This is equivalent to (33). □

The main weakness of Proposition 13 is that it postulates $\bar{E}$ as the invariant replacement for E. This is why we are more interested in inverse statements to Corollary 8, such as Theorem 9.

C.2 Proof of Theorem 9

Fix $c > 1$. The proof is based on similar ideas to the proof of Theorem 4. As there, our construction will depend very little on the objects (we will usually use x_0), and we will set, without loss of generality, $\mathbf{Y} \subseteq \{0', 1', 0, \dots, 2^n - 1\}$, with $0'$ and $1'$ used in the same sense as in Sect. B.2. What will be actually used in the proof is $1 \ll |\mathbf{Y}| \ll 2^n$, which we assume; in particular, $|\mathbf{Y}| > 2$. Each binary sequence $y_1, \dots, y_n$ in $\{0', 1'\}^n$ is encoded as a label in $\mathbf{Y} \setminus \{0', 1'\}$, namely

$$\overline{y_1 \dots y_n} := \sum_{i=1}^{n} y_i 2^{n-i},$$

so that the sequence $y_1, \dots, y_n$ can be decoded as the binary representation of the label $\overline{y_1 \dots y_n}$.

Define $E \in \mathcal{E}^{\mathrm{X}}$ as

$$E(x_1, y_1, \dots, x_n, y_n, x_{n+1}, y) := \begin{cases} (n+1)\binom{n}{k} & \text{if } y_1, \dots, y_n \in \{0', 1'\} \text{ and } \overline{y_1 \dots y_n} = y \\ 0 & \text{otherwise,} \end{cases}$$

where $k := y_1 + \dots + y_n$. It is easy to see that this is indeed an exchangeability e-variable.

We include in $\mathbf{Y}$ the bits $0'$ and $1'$, and the rest of its elements are the codes of different binary sequences $y_1, \dots, y_n$ in $\{0', 1'\}^n$ that have $y_1 + \dots + y_n$ as close to $n/2$ as possible. Let $\mathbf{Y}'$ be the set of those binary sequences (so that $|\mathbf{Y}'| = |\mathbf{Y}| - 2$). We can construct $\mathbf{Y}'$ sequentially starting from the empty set and at each step adding to it a new $(y_1, \dots, y_n)$ with the smallest $y_1 + \dots + y_n$ (and with ties broken arbitrarily).

By the central limit theorem, $|\mathbf{Y}'| = o(2^n)$ implies that

$$\max_{(y_1, \dots, y_n) \in \mathbf{Y}'} |y_1 + \dots + y_n - n/2| = o(\sqrt{n}),$$

and so by the local limit theorem [Shiryaev, 2016, Sect. 1.6],

$$\binom{n}{k} \sim 2^n B^n_{1/2}\left(\{(y_1, \dots, y_n) \in \{0,1\}^n : y_1 + \dots + y_n = k\}\right) \sim 2^n / \sqrt{\pi n/2},$$

provided $k = y_1 + \dots + y_n$ for some $(y_1, \dots, y_n) \in \mathbf{Y}'$. Therefore, the value of E on $\mathbf{Y}'$ will be asymptotically equivalent to $n2^n/\sqrt{\pi n/2}$.

Choose any $E' \in \mathcal{E}^{\mathrm{iX}}$ and $G \in \mathcal{E}^{\mathrm{R}}$. It is always true that $E' \le n+1$ for $E' \in \mathcal{E}^{\mathrm{iX}}$. To show this, we can argue indirectly. Suppose $E'(z'_1, \dots, z'_{n+1}) > n+1$. Averaging $E'(z'_{\pi(1)}, \dots, z'_{\pi(n+1)})$ over the permutations of $\{1, \dots, n+1\}$, we will obtain more than $(n+1)n!/(n+1)! = 1$, which is impossible.

The maximum B^n_θ-probability of the set $\mathbf{Y}'$ will be asymptotically equivalent to $|\mathbf{Y}|\, 2^{-n}$ (which is also asymptotically equivalent to $B^n_{1/2}(\mathbf{Y}')$). The analogous statement will also be true about the set $\mathbf{Y}' \times \{0', 1'\}$ of one-bit extensions of the elements of $\mathbf{Y}'$:

$$\max_\theta B^{n+1}_\theta(\mathbf{Y}' \times \{0', 1'\}) \sim |\mathbf{Y}|\, 2^{-n}.$$

Therefore, there exists a data sequence $(y_1, \dots, y_{n+1})$ (with all objects being x_0) on which $G \le a2^n/|\mathbf{Y}|$ for a given constant $a > 1$ and whose sequence of the first n labels $(y_1, \dots, y_n)$ is in $\mathbf{Y}'$. Fix such a sequence $(y_1, \dots, y_{n+1})$, and set $y := \overline{y_1 \dots y_n}$.

Taking $a > 1$ and $b < 1$ sufficiently close to 1 and combining statements made earlier in the proof,

$$\frac{c\sqrt{\pi n/2}}{|\mathbf{Y}|}\frac{E(y_1,\dots,y_n,y)}{G(y_1,\dots,y_n,y_{n+1})} > \frac{c\sqrt{\pi n/2}}{|\mathbf{Y}|}\frac{bn2^n/\sqrt{\pi n/2}}{a2^n/|\mathbf{Y}|} = c\frac{b}{a}n$$
$$> n+1 \geq E'(y_1,\dots,y_n,y),$$

we obtain (12) holding for sufficiently large $|\mathbf{Y}|$ and sufficiently large n. The actual condition on $|\mathbf{Y}|$ and n that we used was $1 \ll k \ll 2^n$; formally, the first sentence in the statement of Theorem 9 can be replaced by "For each constant $c > 1$ there is $C > 0$ such that the following statement holds true assuming $|\mathbf{Y}| \geq C$ and $2^n \geq C\,|\mathbf{Y}|$."

C.3 Discussion of Theorem 9

In addition to the presence of the factor $\sqrt{\pi n/2}$ in (12) (discussed earlier, after the statement of the theorem), there are two further obvious weaknesses of Theorem 9. One is the requirement $1 \ll |\mathbf{Y}| \ll 2^n$ (again stated in a much cruder form in the theorem). The other is that the inequality (12) in Theorem 9 was shown to hold in a situation where both E and E' take very large values (the maximum possible value in the case of E' in the proof of Theorem 9). These two weaknesses parallel two weaknesses of Theorem 4 pointed out in Appendix B (Sect. B.3).

C.4 Variations on Theorem 9

The factor $\sqrt{\pi n/2}$ in (12) comes, essentially, from the requirement that G be a randomness e-variable in Theorem 9. In Theorem 7 the requirement on G is stronger; in particular, it is required to be an exchangeability e-variable. Let us replace "randomness" by "exchangeability" in Theorem 9. We will check that in this case we have

$$E'(z_1,\dots,z_n,x_{n+1},y) < \frac{2c}{|\mathbf{Y}|}\frac{E(z_1,\dots,z_n,x_{n+1},y)}{G(z_1,\dots,z_n,z_{n+1})}. \tag{34}$$

in place of (12) under the condition $|\mathbf{Y}| \leq 2^n/\sqrt{\pi n/2}$. We still have an extra factor of 2 in (34), but now it is a small constant.

Let us check that we can achieve (34). When constructing $\mathbf{Y}'$ sequentially starting from the empty set, we will now break ties more systematically; namely, we prefer adding to $\mathbf{Y}'$ a new $(y_1,\dots,y_n)$ for which there is already an existing sequence $(y'_1,\dots,y'_n)$ in $\mathbf{Y}'$ with $y'_1+\dots+y'_n = y_1+\dots+y_n$. The *kth level* of $\{0,1\}^n$ is the set $\{(y_1,\dots,y_n) : y_1+\dots+y_n\}$; we will also use this term with $n+1$ in place of n. By the local limit theorem, the condition $|\mathbf{Y}| \leq 2^n/\sqrt{\pi n/2}$ implies that either all elements of $\mathbf{Y}'$ will be at the same level or one level will be completely covered by $\mathbf{Y}'$ (and the rest of $\mathbf{Y}'$ will be negligible). Let it be the kth level. We have $B^n_{1/2}(\mathbf{Y}') \geq b\,|\mathbf{Y}|\,2^{-n}$ for a given constant $b < 1$ asymptotically (i.e., for sufficiently large $|\mathbf{Y}|$ and n). Therefore, $B^{n+1}_{1/2}(\mathbf{Y}'\times\{j\}) \geq b\,|\mathbf{Y}|\,2^{-n-1}$ for $j = 0$ or $j = 1$. By the local limit theorem and $|k - n/2| \leq 1/2$, the maximum exchangeable probability of $\mathbf{Y}'\times\{j\}$ (which is essentially a subset of the $(k+j)$th

level) will be at least $b^2 |\mathbf{Y}| 2^{-n-1}\sqrt{\pi n/2}$ asymptotically. Therefore, we have $G \le b^{-2}2^{n+1}/(\sqrt{\pi n/2}\,|\mathbf{Y}|)$ on some data sequence with labels $(y_1, \dots, y_n) \in \mathbf{Y}'$ and j (and with all objects being x_0). Replacing the old lower bound on G by this new one (with b sufficiently close to 1), we will obtain (34).

To get rid of the extra factor of 2 in (34), we can further require that G be a conditional training exchangeability e-variable, as in Theorem 7. Since the maximum exchangeable probability of $\mathbf{Y}' \subseteq \{0', 1'\}^n$ is asymptotically equivalent to $\sqrt{\pi n/2}\,|\mathbf{Y}|\,2^{-n}$, under the same condition $|\mathbf{Y}| \le 2^n/\sqrt{\pi n/2}$, we can choose y_{n+1} arbitrarily and then take $(y_1, \dots, y_n) \in \mathbf{Y}'$ such that

$$G(x_0, y_1, \dots, x_0, y_{n+1}) \le \frac{a2^n}{\sqrt{\pi n/2}\,|\mathbf{Y}|},$$

for a given constant $a > 1$. This will give us (34) without the factor of 2.

C.5 Another inverse to Corollary 12

The topic of this section is the following analogue of Theorem 11, which addresses a weakness of Theorem 9 mentioned in Sect. C.3.

Theorem 14. *For each constant $c \in (0, 1)$ the following statement holds true for a sufficiently large $|\mathbf{Y}|$ and a sufficiently large n. There exists an exchangeability e-variable E such that for each invariant exchangeability e-variable E' and each randomness e-variable G there exist $z_1, \dots, z_n$, $z_{n+1} = (x_{n+1}, y_{n+1})$, and $y \ne y_{n+1}$ satisfying* (25)–(27).

The naive interpretation of Theorem 14 is completely analogous to that of Theorem 11 and demonstrates that exchangeability e-prediction proper can be superior to conformal e-prediction proper for a large label space. To compare Theorem 14 with Corollary 8, we can still use (28). This inequality falls short of being a perfect inverse to (11) by a smaller constant than in Kolmogorov's step, namely 4.02.

Proof of Theorem 14. Let us again ignore the objects. Generate labels $Y_1, \dots, Y_{n+1}$ randomly (independently from the uniform distribution on $\mathbf{Y}$), and set $Y := Y_1$. As in the proof of Theorem 11, we will prove that the probabilities of the events (29), (30), and (31) are at least 0.5, 0.502, and 0.999, respectively; this will imply the statement of the theorem. For (29), the argument is the same as before.

Define E as

$$E(y_1, \dots, y_{n+1}) := \begin{cases} \frac{(n+1)n}{\sum_{y \in \mathbf{Y}} k_y(k_y - 1)} & \text{if } y_1 = y_{n+1} \\ 0 & \text{otherwise,} \end{cases} \tag{35}$$

where k_y, $y \in \mathbf{Y}$, is the number of times that y occurs in $y_1, \dots, y_{n+1}$, $k_y := \sum_{i=1}^{n+1} 1_{\{y_i = y\}}$. Let us check that (35) is indeed an exchangeability e-variable, i.e., that the average of E over all permutations of $(y_1, \dots, y_{n+1})$ does

not exceed 1. The probability that a random permutation of $(y_1, \dots, y_{n+1})$ has $y_{\pi(1)} = y_{\pi(n+1)}$ is equal to the probability that $y_i = y_j$ for two random positions $i, j \in \{1, \dots, n+1\}$, and the last probability is

$$\frac{\sum_{y \in \mathbf{Y}} k_y (k_y - 1)/2}{(n+1)n/2};$$

therefore, the average is 1.

As $n \to \infty$ (while $|\mathbf{Y}|$ is kept fixed), $E(Y_1, \dots, Y_n, Y)$ will converge to

$$\frac{n^2}{(n/|\mathbf{Y}|)^2 |\mathbf{Y}|} = |\mathbf{Y}|$$

in probability by the law of large numbers. This implies that the probability of (31) can be made arbitrarily close to 1.

To bound the probability of (30), let us find the largest possible $\mathbb{E}(E'(Y_1, \dots, Y_n, Y))$. We will find optimal values of E' on all permutations of a given data sequence $\zeta \in \mathbf{Z}^{n+1}$ for each ζ separately (for a given ζ these values should average to at most 1). Let k_y be the number of occurrences of $y \in \mathbf{Y}$ in ζ and set $k^* := \max_y k_y$. The probability of $(Y_1, \dots, Y_n, Y) = \zeta'$ for a permutation ζ' of ζ will be $|\mathbf{Y}|^{-n}$ if the first and last elements of ζ' coincide and 0 if they are different. The conditional probability of ζ' having a given $y \in \mathbf{Y}$ as its first and last elements (conditional on ζ' being a random permutation of ζ) will be $k_y(k_y - 1)/((n+1)n)$. The conditional expectation of E' given the bag $\mathrm{conf}(\zeta)$ is, therefore,

$$\sum_{y \in \mathbf{Y}} E'(y) \frac{k_y (k_y - 1)}{(n+1)n} |\mathbf{Y}|; \tag{36}$$

this expression uses the shorthand $E'(y)$ for $E'(\zeta \setminus \{y\}, y)$, where $\zeta \setminus \{y\}$ is ζ with one of the entries of y crossed out (it does not matter which entry because of the invariance of E'). Maximizing (36) under the constraint

$$\sum_{y \in \mathbf{Y}} E'(y) \frac{k_y}{n+1} = 1$$

is a Neyman–Pearson-type optimization problem, and its solution is attained on E' concentrated on $\arg\max_y k_y$. The solution itself is $(k^* - 1) |\mathbf{Y}| / n$. By the law of large numbers, $(k^* - 1) |\mathbf{Y}| / n$ converges to 1 as $n \to \infty$ (since $k_y / n \to 1/|\mathbf{Y}|$ for each $y \in \mathbf{Y}$). Therefore, $\mathbb{E}(E'(Y_1, \dots, Y_n, Y)) \to 1$, and it remains to apply Markov's inequality to get our statement about the probability of (30). □