Universality of conformal prediction under the assumption of randomness

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 #43

First posted February 26, 2025. Last revised January 17, 2026.

Project web site:
http://alrw.net

Abstract

Conformal predictors provide set or functional predictions that are valid under the assumption of randomness, i.e., under the assumption of independent and identically distributed data. The question asked in this paper is whether there are predictors that are valid in the same sense under the assumption of randomness and that are more efficient than conformal predictors. The answer is that the class of conformal predictors is universal in that only limited gains in predictive efficiency are possible. The previous work in this area has relied on the algorithmic theory of randomness and so involved unspecified constants, whereas this paper's results are much more practical. They are also shown to be optimal in some respects.

Contents

1 Introduction

The main assumption of machine learning is that of *randomness*, i.e., the assumption that the observations are independent and identically distributed (IID). The method of conformal prediction [1, 18] allows us to complement predictions output by standard machine-learning algorithms by some measures of confidence in those predictions that are valid under the assumption of randomness. However, conformal prediction only uses the assumption of exchangeability, which is much weaker than randomness; in particular, the difference between randomness and exchangeability completely dwarfs the differences that are relevant in conformal prediction; see Remark 3 below for details. A natural question is whether conformal prediction loses much by not using the full strength of the assumption of randomness and only using exchangeability. The main message of this paper and the earlier paper by Nouretdinov et al. [9] is that the answer is "no" (we don't lose much), at least in the first approximation. In the terminology of [9], conformal prediction is universal.

The main limitation of the pioneering paper [9] is that it is based on the algorithmic theory of randomness. Because of that, Nouretdinov et al.'s results involve unspecified constants and, therefore, are never applicable to practical machine learning. This paper's results are more precise in several respects, but the main one is that they only involve fully specified constants and, therefore, open up the possibility of quantifying real-world limitations of conformal predictors. Other advances of this paper as compared with [9] are that our results are not restricted to the case of classification, but in the case of classification they are stronger and are complemented by optimality results.

To demonstrate that conformal prediction does not lose much by not using the full strength of the assumption of randomness, this paper introduces the most general class of predictors, "randomness predictors", which produce predictions of the same kind as conformal predictors but are only required to be valid under the assumption of randomness. There are many more randomness predictors than conformal predictors, and our question is whether conformal predictors are as good as arbitrary randomness predictors. The answer given in this paper is a qualified yes: every randomness predictor P can be modelled by a conformal predictor P' so that the predictions output by P' are almost as good as those output by P, unless we are entitled to reject the randomness of the true data sequence. It is unclear whether the difference between conformal and randomness prediction can be usefully exploited at all (cf. [15]).

One simplifying assumption made in this paper is that it concentrates on predictors that are *train-invariant*, i.e., invariant w.r. to permutations of the training sequence. (We will remove this assumption only in Appendix B.) The assumption of train-invariance for predictors is extremely natural under the assumption of randomness for the data. It is reflected in the standard expression for a training sequence being "training set" in machine learning; since "set" implies the lack of order, this expression is only justified for train-invariant predictors (and even in this case it is not justified completely; it would have been more accurate to say "training bag"). In fact, conformal predictors can

be defined as train-invariant predictors that are valid under exchangeability [9, Proposition 1]. This will be discussed in detail in Sect. 2, where we will also see that the requirement of train-invariance is justified by general principles of statistical inference.

Conformal prediction is usually presented as a method of *set prediction* [18, Part I], i.e., as a way of producing prediction sets (rather than point predictions). Another way to look at a conformal predictor is as a way of producing a p-value function (discussed, in a slightly different context, in, e.g., [5]), which is a function mapping each possible label y of a test object to the corresponding conformal p-value. In analogy with "prediction sets", we may call such p-value functions "prediction functions". The prediction set Γ^α corresponding to a prediction function f and a significance level $\alpha \in (0,1)$ (our target probability of error) is the set of all labels y such that $f(y) > \alpha$. A standard property of validity for conformal predictors is that Γ^α makes an error (fails to cover the true label) with probability at most α; it is implied by the conformal p-values being bona fide p-values (under suitable assumptions, such as data exchangeability).

To establish connections between conformal and randomness predictors we will use conformal e-predictors [14], which are obtained by replacing p-values with e-values (for the definition of e-values, see, e.g., [6], [10], [19], or Sect. 2 below). Conformal e-predictors output e-value functions f as their prediction functions. Such functions f can also be represented in terms of the corresponding prediction sets $\Gamma^\alpha := \{y \mid f(y) < \alpha\}$, where $\alpha \in (0,\infty)$ is the significance level (notice that now we exclude the labels with large e-values from the prediction set, which is opposite to what we did for p-values). However, the property of validity of conformal e-predictors is slightly more difficult to state in terms of prediction sets: now validity means that the integral of the probability of error for Γ^α over $\alpha \in (0,\infty)$ does not exceed 1 [14, end of Appendix B]. This implies that the probability of error for Γ^α is at most $1/\alpha$, but this simple derivative property of validity is much weaker.

Conformal e-predictors are not only a useful technical tool, but we can also use them for prediction directly. In Shafer's opinion [11], e-values are even more intuitive than p-values. Because of the importance of e-predictors, in the rest of this paper we will use the word "predictor" in combinations such as "conformal predictor" and "randomness predictor" generically, including both p-predictors (standard predictors based on p-values) and e-predictors (predictors based on e-values); in particular, we will never drop "p-" in "p-predictor". This is a potential source of confusion, and the reader should keep in mind that the usual notion of conformal predictors corresponds to conformal p-predictors in this paper.

We start in Sect. 2 from the main definitions, including those of conformal and randomness predictors. The main classes of predictors that we are interested in in this paper are shown in Figure 1. All four classes are train-invariant. An arrow going from class A to class B means embedding: $A \subseteq B$. Section 3 is devoted to the main results, first for e-predictors in Subsect. 3.1 and then for p-predictors in Subsect. 3.2. A line between classes A and B in Figure 1 means the possibility of transformations from $f \in A$ to $f' \in B$ and vice versa;

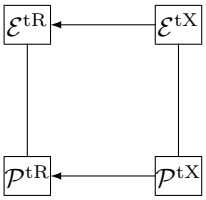

Figure 1: A square representing the main classes of predictors considered in this paper: $\mathcal{E}^{\mathrm{tR}}$ (train-invariant randomness e-predictors), $\mathcal{E}^{\mathrm{tX}}$ (conformal e-predictors), $\mathcal{P}^{\mathrm{tR}}$ (train-invariant randomness p-predictors), and $\mathcal{P}^{\mathrm{tX}}$ (conformal p-predictors).

such transformations, called "calibration", will be discussed at the beginning of Subsect. 3.2 and will serve as a way of deducing results for p-predictors from those for e-predictors. In the two subsections, we establish the predictive efficiency of conformal predictors among randomness predictors in both e- and p-versions. Namely, the prediction functions for conformal predictors turn out, under the proviso of the true data sequence looking IID, to be competitive on average with the prediction functions for any randomness predictors, where "on average" refers to an arbitrary probability measure that can depend on the test example. These results are illustrated on the simple case of binary classification. Section 4 gives applications to multi-class classification and some optimality results, while Sect. 5 concludes listing some limitations of our results and directions of further research.

Our notation for the base of natural logarithms will be $\mathrm{e} \approx 2.72$ (while italic e will often serve as a generic notation for e-values).

2 Definitions

This paper deals with the following prediction problem. We are given a training sequence of *examples* $z_i = (x_i, y_i)$, $i = 1, \dots, n$ for a fixed n, each consisting of an *object* x_i and its *label* y_i, and a new test object x_{n+1}; the task is to predict x_{n+1}'s label y_{n+1}. A potential label y for x_{n+1} is *true* if $y = y_{n+1}$ and *false* otherwise (therefore, while there is only one true label for x_{n+1}, there may be numerous false labels). The objects are drawn from a non-empty measurable space $\mathbf{X}$, the *object space*, and the labels from the *label space* $\mathbf{Y}$, which is assumed to be a non-trivial measurable space (meaning that the σ-algebra on it is different from $\{\emptyset, \mathbf{Y}\}$).

A measurable function $P : \mathbf{Z}^{n+1} \to [0,1]$ is a *randomness p-variable* if, for any probability measure Q on $\mathbf{Z}$ and any *significance level* $\alpha \in (0,1)$, $Q^{n+1}(\{P \le \alpha\}) \le \alpha$. Such a function P is an *exchangeability p-variable* if $R(\{P \le \alpha\}) \le \alpha$ for any exchangeable probability measure R on $\mathbf{Z}^{n+1}$ and any $\alpha \in (0,1)$. (Equivalently, P is an exchangeability p-variable if, for any data sequence in $\mathbf{Z}^{n+1}$ and any threshold $\alpha \in (0,1)$, the fraction of the permutations of

the data sequence at which $P \le \alpha$ does not exceed α.) And a real-valued function P defined on $\mathbf{Z}^{n+1}$ is *train-invariant* if it is invariant w.r. to permutations of the training examples:

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

for each data sequence $(z_1, \dots, z_{n+1}) \in \mathbf{Z}^{n+1}$ and each permutation σ of $\{1, \dots, n\}$. In other words, train-invariant functions should depend on the training examples $z_1, \dots, z_n$ only via the training bag $\wr z_1, \dots, z_n \int$. Finally, P is a *conformal p-variable* if it is a train-invariant exchangeability p-variable.

We will sometimes refer to the values taken by p-variables as *p-values*, and our notation for the classes of all randomness, train-invariant randomness, exchangeability, and conformal p-variables will be $\mathcal{P}^{\mathrm{R}}$, $\mathcal{P}^{\mathrm{tR}}$, $\mathcal{P}^{\mathrm{X}}$, and $\mathcal{P}^{\mathrm{tX}}$, respectively.

Conformal p-variables can be used for prediction, and we will also refer to them as *conformal p-predictors* (they are usually called simply "conformal predictors"; cf. [1,18]). There are several ways to package the output of conformal p-predictors, as discussed in Sect. 1. One is in terms of set prediction: for each significance level $\alpha \in (0, 1)$, each training sequence $z_1, \dots, z_n$, and each test object x_{n+1}, we can output the *prediction set*

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

By the definition of conformal p-variables, under the assumption of exchangeability, the probability that a conformal p-predictor makes an error at significance level α, i.e., the probability of $y_{n+1} \notin \Gamma^\alpha$, is at most α.

Instead of predicting with one prediction set in the family (1), in this paper we prefer to package our prediction as the *prediction function*

$$f(y) := P(z_1, \dots, z_n, (x_{n+1}, y)), \quad y \in \mathbf{Y}. \tag{2}$$

We may refer to this mode of prediction as *functional prediction*. The step from set prediction to functional prediction is analogous to the step from confidence intervals to p-value functions (see, e.g., [8, Sect. 9] and [4, 5] for the latter).

Remark 1. There are several equivalent definitions of conformal p-predictors, and the definition as train-invariant exchangeability p-variables (first given in [9, Proposition 1]) is one of them. Let us check that it is equivalent to, e.g., the definition of conformal p-values given in [18, (2.20)] for a fixed length n of the training sequence. In one direction, it is obvious that conformal p-values as defined there are train-invariant and valid under exchangeability. On the other hand, given a train-invariant exchangeability p-variable P, we can define the nonconformity measure

$$A(\wr z_1, \dots, z_{n+1} \int, z_i) := 1/P(z_{i+1}, \dots, z_{n+1}, z_1, \dots, z_i),$$

and the resulting conformal p-values will dominate P (dominate in the sense of being less than or equal to).

Remark 2. The term "functional prediction" is a straightforward modification of "set prediction" and "p-value function", but its disadvantage is that it is easy to confuse with function prediction, namely predicting a function (e.g., a biological function, such as that of a protein, or a mathematical function).

Similarly, we can use randomness p-variables for prediction, and then we refer to them as *randomness p-predictors*. By definition, the probability that the prediction set (1) derived from a randomness p-predictor makes an error is at most α, this time under the assumption of randomness. We will use the prediction functions (2) for randomness p-predictors as well. Less important, we call exchangeability p-variables exchangeability p-predictors.

Remark 3. The difference between the assumptions of randomness and exchangeability disappears for infinite data sequences under a mild assumption about the example space $\mathbf{Z}$ (it is required to be a Borel space). This follows from de Finetti's theorem, which represents exchangeable probability measures as integral mixtures of product probability measures Q^∞. The difference becomes very significant for finite data sequences of a given length; this is relevant for our prediction problems where we deal with randomness or exchangeability of the "augmented training sequences" $z_1, \dots, z_n, (x_{n+1}, y)$ of length $N := n+1$ (cf. (1) and (2)). See [17] for a detailed exposition. In particular, it is shown in [17] that, for any N and assuming $\mathbf{Z}$ is rich enough, there exists an event $A \subseteq \mathbf{Z}^N$ such that $R(A) = 1$ for some exchangeable probability measure R on $\mathbf{Z}^N$ while $Q^N(A) \le 2^{-N+1}$ for all probability measures Q on $\mathbf{Z}$. Therefore, assuming $N \gg 1$, a randomness p-variable assigns tiny p-values of at most 2^{-N+1} to each data sequence in A while the event A is perfectly plausible under exchangeability. It is easy to see that the smallest possible p-value (2) output by a conformal p-predictor is $1/N$; this was called the "fundamental limitation of conformal prediction" in [15]. While 2^{-N+1} shrinks exponentially as N grows, $1/N$ only shrinks polynomially fast. In this sense, the difference between randomness and exchangeability is by far more significant than what is attainable in conformal prediction.

Two standard desiderata for conformal, and by extension randomness, predictors are their validity and efficiency. In terms of the prediction function f, validity concerns the value $f(y_{n+1})$ of f at the true label (the typical values should not be too small in p-prediction), and efficiency concerns the values $f(y)$ at the false labels $y \ne y_{n+1}$ (they should be as small as possible in p-prediction). Validity is automatic under randomness (and even under exchangeability for conformal predictors), and in this paper we are interested in the efficiency of conformal predictors relative to other randomness predictors. Later in the paper (Theorem 5 and Corollary 7 below) we will establish efficiency guarantees for conformal prediction in terms of randomness prediction.

A nonnegative measurable function $E : \mathbf{Z}^{n+1} \to [0, \infty]$ is a *randomness e-variable* if $\int E \, \mathrm{d}Q^{n+1} \le 1$ for any probability measure Q on $\mathbf{Z}$. It is an *exchangeability e-variable* if $\int E \, \mathrm{d}R \le 1$ for any exchangeable probability measure R on $\mathbf{Z}^{n+1}$. We will denote the classes of all randomness and exchangeability e-variables by $\mathcal{E}^{\mathrm{R}}$ and $\mathcal{E}^{\mathrm{X}}$, respectively. The class of all measurable functions

$E : \mathbf{Z}^{n+1} \to [0, \infty]$ is denoted by $\mathcal{E}$. It is easy to see that $E \in \mathcal{E}$ belongs to $\mathcal{E}^{\mathrm{X}}$ if and only if, for any data sequence $z_1, \dots, z_{n+1}$,

$$\frac{1}{(n+1)!} \sum_{\pi} E(z_{\pi(1)}, \dots, z_{\pi(n+1)}) \le 1, \tag{3}$$

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

The class $\mathcal{E}^{\mathrm{tX}}$ of *conformal e-variables* consists of all functions $E \in \mathcal{E}^{\mathrm{X}}$ that are train-invariant. We often regard the randomness e-variables $E \in \mathcal{E}^{\mathrm{R}}$ as *randomness e-predictors* and conformal e-variables $E \in \mathcal{E}^{\mathrm{tX}}$ as *conformal e-predictors*. Similarly to (2), they output prediction functions

$$f(y) := E(z_1, \dots, z_n, (x_{n+1}, y)), \quad y \in \mathbf{Y}.$$

Remark 4. Similarly to Remark 1, it is easy to check that a train-invariant exchangeability e-variable is the same thing as a conformal e-predictor as defined in, e.g., [14]. Indeed, the nonconformity e-measure [14, Sect. 2] corresponding to a train-invariant exchangeability e-variable E is

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

The subclass $\mathcal{E}^{\mathrm{tR}} \subseteq \mathcal{E}^{\mathrm{R}}$ of all train-invariant randomness e-predictors is important since under the assumption of randomness it is natural to consider only train-invariant predictors: the requirement of train-invariance is a special case of the principle of sufficiency in statistical inference [2, 2.3.(ii)]. The requirement of train-invariance under the assumption of randomness (the "train-invariance principle") is a special case not only of the sufficiency principle but also of the invariance principle [2, Example 2.35], which makes it even more convincing.

For conformal and randomness e-predictors, validity and efficiency change direction as compared with p-predictors: for validity, typical values $f(y_{n+1})$ should not be too large, and for efficiency typical values $f(y)$ at the false labels $y \ne y_{n+1}$ should be as large as possible. Again validity is automatic under randomness, and Theorem 5 below establishes efficiency guarantees for conformal e-prediction in terms of train-invariant randomness e-prediction. Similarly to $\mathcal{E}^{\mathrm{R}}$ and $\mathcal{E}^{\mathrm{tX}}$, the elements of the class $\mathcal{E}^{\mathrm{X}}$ will be referred to as *exchangeability e-predictors*, but we are not particularly interested in them per se.

We will also need another, even narrower, subclass of $\mathcal{E}^{\mathrm{R}}$, $\mathcal{E}^{\mathrm{iR}} \subseteq \mathcal{E}^{\mathrm{tR}}$. The class $\mathcal{E}^{\mathrm{iR}}$ consists of all e-variables $E \in \mathcal{E}^{\mathrm{R}}$ that are invariant w.r. to all permutations:

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

for each permutation π of $\{1, \dots, n+1\}$; let us call such randomness e-variables *invariant*. An equivalent definition is to say that $E(z_1, \dots, z_{n+1})$ depends on the data sequence $z_1, \dots, z_{n+1}$ only via the bag $\wr z_1, \dots, z_{n+1} \int$ of its elements. The interpretation is that the value $E(z_1, \dots, z_{n+1})$ for $E \in \mathcal{E}^{\mathrm{iR}}$ is the degree

Table 1: Function classes defined in Sect. 2

notation	description	defined on
$\mathcal{P}^{\mathrm{R}}$	randomness p-variables/p-predictors	page 4
$\mathcal{P}^{\mathrm{X}}$	exchangeability p-variables/p-predictors	page 4
$\mathcal{P}^{\mathrm{tR}}$	train-invariant randomness p-variables/p-predictors	page 4
$\mathcal{P}^{\mathrm{tX}}$	conformal (= train-invariant exchangeability) p-variables/p-predictors	page 4
$\mathcal{E}^{\mathrm{R}}$	randomness e-variables/e-predictors	page 5
$\mathcal{E}^{\mathrm{X}}$	exchangeability e-variables/e-predictors	page 5
$\mathcal{E}^{\mathrm{tR}}$	train-invariant randomness e-variables/e-predictors	page 6
$\mathcal{E}^{\mathrm{tX}}$	conformal (= train-invariant exchangeability) e-variables/e-predictors	page 6
$\mathcal{E}$	all $[0,\infty]$-valued random variables	page 6
$\mathcal{E}^{\mathrm{iR}}$	invariant randomness e-variables	page 6

to which we can reject the hypothesis that the bag $\wr z_1,\dots,z_{n+1}\int$ resulted from its elements being generated in the IID fashion.

The function classes introduced in this section are listed in Table 1. It might appear that there is some redundancy in our terminology, since, e.g., we refer to the same mathematical objects as both exchangeability p-variables and exchangeability p-predictors. However, these terms are applied in different contexts and are not interchangeable in informal discussions. Namely, exchangeability p-variables test the assumption of exchangeability for given data, whereas exchangeability p-predictors accept exchangeability and use it for prediction (rejecting future observations that would lead to rejection of exchangeability). The same remark is applicable to the assumption of randomness and to e-variables and e-predictors.

A big advantage of e-variables over p-variables is that the average of e-variables is again an e-variable. This allows us to define, given an e-variable $E\in\mathcal{E}^{\mathrm{R}}$, two derivative e-variables:

$$E^{\mathrm{i}}(z_1,\dots,z_{n+1}) := \frac{1}{(n+1)!}\sum_{\pi} E(z_{\pi(1)},\dots,z_{\pi(n+1)}), \tag{4}$$

$$E^{\mathrm{X}}(z_1,\dots,z_{n+1}) := \frac{E(z_1,\dots,z_{n+1})}{E^{\mathrm{i}}(z_1,\dots,z_{n+1})}, \tag{5}$$

with π ranging over the permutations of $\{1,\dots,n+1\}$ and $0/0$ interpreted as 1 (the last convention may be relevant to (5)). More generally, we can allow any $E\in\mathcal{E}$ in (4) and (5). It is clear that $E^{\mathrm{i}}\in\mathcal{E}^{\mathrm{iR}}$ whenever $E\in\mathcal{E}^{\mathrm{R}}$ and that $E^{\mathrm{X}}\in\mathcal{E}^{\mathrm{X}}$ for all $E\in\mathcal{E}$.

To see the intuitive meaning of (4) and (5), let us start from a generic randomness e-variable $E\in\mathcal{E}^{\mathrm{R}}$. It does not have to be an exchangeability e-variable, but we would like to make an exchangeability e-variable out of it, i.e., we want the average of the values of the new e-variable over all permutations

of the input data sequence to be at most 1 (see (3)). Then (5) is simply the result of normalizing E: we divide E by its average over all permutations. Therefore, E^{X} is an exchangeability e-variable. The denominator, E^{i}, is what distinguishes randomness e-variables from exchangeability e-variables. If E is already an exchangeability e-variable, E^{i} will be at most 1, but randomness e-variables E can play with their invariant component E^{i}.

In the most important for us case (transforming train-invariant randomness e-predictors to conformal e-predictors), the operator (5) is polynomially computable: namely, if $E \in \mathcal{E}^{\mathrm{tR}}$ is efficiently computable, the extra time when computing $E^{\mathrm{X}} \in \mathcal{E}^{\mathrm{tX}}$ on top of computing E is linear, $O(n)$. This follows from

$$E^{\mathrm{X}}(z_1, \dots, z_{n+1}) = \frac{E(z_1, \dots, z_{n+1})}{\frac{1}{n+1} \sum_{i=1}^{n+1} E(z_{i+1}, \dots, z_{n+1}, z_1, \dots, z_i)}$$

for $E \in \mathcal{E}^{\mathrm{tR}}$.

3 Main results

Let B be a Markov kernel with source $\mathbf{Z}$ and target $\mathbf{Y}$, which we will write in the form $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ (as in [18, Sect. A.4]). We will write $B(A \mid z)$ for its value on $z \in \mathbf{Z}$ and $A \subseteq \mathbf{Y}$ (where A is measurable) and write $\int f(y) B(\mathrm{d}y \mid z)$ for the integral of a function f on $\mathbf{Y}$ w.r. to the measure $A \mapsto B(A \mid z)$. We will show that the efficiency of the conformal predictor derived from a train-invariant randomness predictor E is not much worse than the efficiency of the original randomness predictor E on average, and B will define the meaning of "on average".

3.1 From train-invariant randomness e-prediction to conformal e-prediction

The following statement shows that efficiency does not suffer much on average when we move from randomness e-prediction to exchangeability e-prediction. It is more general than what we need at the moment, since it also covers randomness e-predictors that are not train-invariant.

Theorem 5. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel. For each randomness e-predictor E,*

$$G(z_1, \dots, z_n, z_{n+1}) := \mathrm{e}^{-1} \int \frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, y))} B(\mathrm{d}y \mid z_{n+1}) \tag{6}$$

(with $0/0$ set to 0 and z_{n+1} represented as (x_{n+1}, y_{n+1})) is a randomness e-variable.

We can interpret (6) as a statement that E^{X} is almost as efficient as E unless the true data sequence $z_1, \dots, z_{n+1}$ does not look IID. The "unless" clause makes sense in view of our assumption of randomness. According to (6) and assuming

that $B(\cdot \mid z_{n+1})$ is concentrated on $\mathbf{Y} \setminus \{y_{n+1}\}$, the mean ratio of the degree to which E rejects a false label y to the degree to which E^{X} rejects y is not large under any probability measure that may depend on the test example unless we can reject the assumption of randomness for the true data sequence. The gap $\mathrm{e}^{-1} \approx 0.37$ between the mean ratio and the e-value at which we reject the assumption of randomness is optimal (Theorem 6 below).

From now on, until the end of the main paper, let us assume that E is train-invariant in Theorem 5, i.e., E is a train-invariant randomness e-predictor (the only exceptions are Corollary 7 and the line following its statement). Then E^{X} in (6) is a conformal e-predictor.

A full proof of Theorem 5 will be given in Sect. A.1, but we will demonstrate the idea of the proof on a simple special case, which also makes the statement of the theorem more tangible. In the case of binary classification, $\mathbf{Y} := \{-1, 1\}$, the most natural choice of B is $B(\{y\} \mid (x, y)) := 0$, so that the Markov kernel sends every example (x, y) to the other label $-y$. We can then rewrite (6) as

$$G(z_1, \dots, z_n, z_{n+1}) := \mathrm{e}^{-1} \frac{E(z_1, \dots, z_n, (x_{n+1}, -y_{n+1}))}{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, -y_{n+1}))}, \tag{7}$$

which does not involve any averaging. We can interpret (7) as the conformal e-predictor E^{X} being almost as efficient as the original train-invariant randomness e-predictor E, where efficiency is measured by the degree to which we reject the false label $-y_{n+1}$. For example, for a small positive constant ϵ, $G \geq 1/\epsilon$ with probability at most ϵ, and so

$$E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, -y_{n+1})) > \mathrm{e}^{-1} \epsilon E(z_1, \dots, z_n, (x_{n+1}, -y_{n+1})) \tag{8}$$

with probability at least $1 - \epsilon$.

To get an idea of the proof, suppose the true data sequence $z_1, \dots, z_{n+1}$, generated in the IID fashion, is such that the ratio in (7) is large. By definition, the ratio $E^{\mathrm{i}} = E/E^{\mathrm{X}}$ is an invariant randomness e-variable, and so only depends on the bag of its input examples and measures the degree to which that bag does not look IID. Flipping the label of a randomly chosen z_i, $i \in \{1, \dots, n+1\}$, in the training bag $\{z_1, \dots, z_{n+1}\}$ leads to a data bag that is still compatible with the assumption of randomness, and if flipping the label y_{n+1} in the training bag leads to a large value of E/E^{X}, this means that the example z_{n+1} was unusual in the training bag, which makes the original data sequence $z_1, \dots, z_{n+1}$, in which the unusual example is also the last one, not compatible with the assumption of randomness.

The following result is a simple statement of optimality for Theorem 5.

Theorem 6. *The constant* e^{-1} *in Theorem 5 cannot be replaced by a larger one, even if we assume* $E \in \mathcal{E}^{\mathrm{tR}}$.

In principle, Theorem 6 follows from a later result (namely, Theorem 8 below), but in Sect. A.2 we will give a simple independent proof. The origin of the factor e^{-1} is the difference between the assumptions of randomness and

exchangeability: while flipping the label of a randomly chosen example z_i, $i \in \{1, \dots, n+1\}$, keeps the exchangeability of the original IID data sequence, the new sequence ceases to be generated in the IID fashion. Therefore, we approximate flipping the label of one randomly chosen example by flipping the label of each example z_i, $i \in \{1, \dots, n+1\}$, with a small probability; e^{-1} is the largest probability (which is attainable) that exactly one label will be flipped. Full details are given in Sect. A.1.

3.2 From train-invariant randomness p-prediction to conformal p-prediction

It is known that, for any $\delta \in (0,1)$, the function $p \mapsto \delta p^{\delta-1}$ transforms p-values to e-values and that the function $e \mapsto e^{-1}$ transforms e-values to p-values. See, e.g., [19, Propositions 2.1 and 2.2]. More generally, any function $f : [0,1] \to [0,\infty]$ integrating to 1 transforms p-values to e-values (is a *p-to-e calibrator*), while $e \mapsto \min(e^{-1}, 1)$ is the optimal way of transforming e-values to p-values (it is an optimal *e-to-p calibrator*). This allows us to adapt Theorem 5 to p-predictors.

Corollary 7. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel and let $\delta \in (0,1)$. For each randomness p-predictor P there exists an exchangeability p-predictor P' such that*

$$G(z_1, \dots, z_n, z_{n+1}) := \frac{\delta}{\mathrm{e}} \int \frac{P'(z_1, \dots, z_n, (x_{n+1}, y))}{P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, y))} \, B(\mathrm{d}y \mid z_{n+1}) \tag{9}$$

is a randomness e-variable.

The proof is obvious (calibrate P to get $E \in \mathcal{E}^{\mathrm{X}}$ and then calibrate E^{X} to get $P' \in \mathcal{P}^{\mathrm{X}}$), but it is still spelled out in Sect. A.3. As before, we concentrate on the case where P is train-invariant, and in this case P' can be chosen as conformal p-predictor.

The interpretation of (9) is much simpler in the binary case $\mathbf{Y} = \{-1, 1\}$ with the same Markov kernel as before. In this case (9) becomes

$$G(z_1, \dots, z_n, z_{n+1}) := \frac{\delta}{\mathrm{e}} \frac{P'(z_1, \dots, z_n, (x_{n+1}, -y_{n+1}))}{P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, -y_{n+1}))}.$$

Therefore,

$$P'(z_1, \dots, z_n, (x_{n+1}, -y_{n+1})) < \frac{\mathrm{e}}{\delta\epsilon} P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, -y_{n+1}))$$

with probability at least $1 - \epsilon$.

In conclusion of this section, let us discuss (9) in general. The interpretation of (9) is that, under the randomness of the true data sequence, $P'(z_1, \dots, z_n, (x_{n+1}, y))$ is typically small (perhaps not to the same degree) when $P(z_1, \dots, z_n, (x_{n+1}, y))$ is small; i.e., we do not lose much in efficiency when converting train-invariant randomness p-predictors to conformal p-predictors. To

see this, fix small $\epsilon_1, \epsilon_2 \in (0,1)$. Then we will have $G(z_1, \dots, z_n, z_{n+1}) < 1/\epsilon_1$ for the true data sequence $z_1, \dots, z_n, z_{n+1}$ unless a rare event (of probability at most ϵ_1) happens. For the vast majority of the potential labels $y \in \mathbf{Y}$ we will then have

$$\frac{\delta}{\mathrm{e}} \frac{P'(z_1, \dots, z_n, (x_{n+1}, y))}{P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, y))} < \frac{1}{\epsilon_1 \epsilon_2}, \tag{10}$$

where "the vast majority" means that the $B(\cdot \mid z_{n+1})$ measure of the y satisfying (10) is at least $1 - \epsilon_2$. We can rewrite (10) as

$$P'(z_1, \dots, z_n, (x_{n+1}, y)) < \frac{\mathrm{e}}{\delta \epsilon_1 \epsilon_2} P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, y)), \tag{11}$$

so that $P'(z_1, \dots, z_n, (x_{n+1}, y)) \to 0$ as $P(z_1, \dots, z_n, (x_{n+1}, y)) \to 0$. In terms of prediction sets (see (1)), (11) means that the conformal p-predictor P' produces prediction sets that are as precise as, or more precise than, the prediction sets produced by the train-invariant randomness p-predictor P if we relax the significance level, assuming that the assumption of randomness is not rejected for the true data sequence and ignoring labels in a set of a small $B(\cdot \mid z_{n+1})$ measure. This is true, of course, for any Markov kernel B.

4 Applications to multi-class classification and optimality results

In this section we discuss the case of classification, $|\mathbf{Y}| < \infty$, but now we are interested in the non-binary case $|\mathbf{Y}| > 2$ (the σ-algebra on $\mathbf{Y}$ is discrete, as usual). Let us only discuss reduction of train-invariant randomness e-predictors to conformal e-predictors. Reduction of train-invariant randomness p-predictors to conformal p-predictors is completely analogous; it just uses (9) instead of (6).

In the case of multi-class classification, $2 < |\mathbf{Y}| < \infty$, the most natural Markov kernel B is perhaps the one for which $B(\cdot \mid (x,y))$ is the uniform probability measure on $\mathbf{Y} \setminus \{y\}$. In this case we can rewrite (6) as

$$G(z_1, \dots, z_n, z_{n+1}) := \frac{\mathrm{e}^{-1}}{|\mathbf{Y}| - 1} \sum_{y \in \mathbf{Y} \setminus \{y_{n+1}\}} \frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, y))}. \tag{12}$$

The interpretation of (12) is that the conformal e-predictor E^{X} is almost as efficient as the original train-invariant randomness e-predictor E on average; as before, efficiency is measured by the degree to which we reject the false labels $y \neq y_{n+1}$. Roughly, on average, we lose at most a factor of e in the e-values of false labels when we replace E by E^{X}.

Of course, we can avoid "on average" by making (12) cruder and replacing it by the existence of $G \in \mathcal{E}^{\mathrm{R}}$ satisfying

$$\forall (z_1, \dots, z_n) \in \mathbf{Z}^n \; \forall x_{n+1} \in \mathbf{X} \; \forall y \in \mathbf{Y} \setminus \{y_{n+1}\} :$$

$$G(z_1,\dots,z_n,z_{n+1}) \geq \frac{\mathrm{e}^{-1}}{|\mathbf{Y}|-1}\frac{E(z_1,\dots,z_n,(x_{n+1},y))}{E^{\mathrm{X}}(z_1,\dots,z_n,(x_{n+1},y))}, \quad (13)$$

where $z_{n+1} := (x_{n+1}, y_{n+1})$. For a small positive constant ϵ, we can then claim that, with probability at least $1-\epsilon$ over the true data sequence,

$$\forall y \in \mathbf{Y} \setminus \{y_{n+1}\} : E^{\mathrm{X}}(z_1,\dots,z_n,(x_{n+1},y)) > \frac{\mathrm{e}^{-1}\epsilon}{|\mathbf{Y}|-1}E(z_1,\dots,z_n,(x_{n+1},y)). \quad (14)$$

An interesting variation of (12), corresponding to the Markov kernel B for which $B(\cdot \mid (x,y))$ is the uniform probability measure on $\mathbf{Y}$, is

$$G(z_1,\dots,z_n,z_{n+1}) := \frac{\mathrm{e}^{-1}}{|\mathbf{Y}|}\sum_{y\in\mathbf{Y}}\frac{E(z_1,\dots,z_n,(x_{n+1},y))}{E^{\mathrm{X}}(z_1,\dots,z_n,(x_{n+1},y))}. \quad (15)$$

Under this definition, the randomness e-variable G does not depend on y_{n+1}.

Let us check that the denominators, $|\mathbf{Y}|-1$ or $|\mathbf{Y}|$, in (12), (13), and (15) are asymptotically optimal.

Theorem 8. *For each constant $c > 1$ the following statement holds true for a sufficiently large $|\mathbf{Y}|$ and a sufficiently large n. There exists a train-invariant randomness e-predictor E 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*

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

When we say "for a sufficiently large $|\mathbf{Y}|$" in Theorem 8, 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 $|\mathbf{Y}|$.

A complete proof of Theorem 8 is given in Sect. A.4, but the informal idea of the proof is that we can ignore the objects and make a false test label $y \neq y_{n+1}$ encode the bag $\{\!|y_1,\dots,y_n|\!\}$ of the training labels. We have $E^{\mathrm{i}} := E/E^{\mathrm{X}} \in \mathcal{E}^{\mathrm{iR}}$. If we also make y easily distinguishable from the training labels $y_1,\dots,y_n$, the value $E^{\mathrm{i}}(z_1,\dots,z_n,(x_{n+1},y))$ (even if depending only on $\{\!|z_1,\dots,z_n,(x_{n+1},y)|\!\}$) can be made large, even for a true data sequence generated in the IID fashion.

Formally, Theorem 8 is an inverse to (13) and (15), but it has several weaknesses in this role:

- It shows that E^{X} is not competitive with E in some cases, but can there be another conformal e-predictor $E' \in \mathcal{E}^{\mathrm{tX}}$ that is better than E^{X} in this respect?
- Even if the ratio $E(z_1,\dots,z_n,(x_{n+1},y))/E'(z_1,\dots,z_n,(x_{n+1},y))$ for $E' \in \mathcal{E}^{\mathrm{tX}}$ is very large, it is not so interesting if already $E'(z_1,\dots,z_n,(x_{n+1},y))$ is very large.
- Even if the ratio $E(z_1,\dots,z_n,(x_{n+1},y))/E'(z_1,\dots,z_n,(x_{n+1},y))$ for $E' \in \mathcal{E}^{\mathrm{tX}}$ is very large, it is not so interesting if $G(z_1,\dots,z_n,z_{n+1})$ is very large for the true data sequence.

The following result overcomes these weaknesses.

Theorem 9. *For each constant $c \in (0,1)$ and each label space $\mathbf{Y}$, $1 < |\mathbf{Y}| < \infty$, the following statement holds true for sufficiently large n. There exists a train-invariant randomness e-predictor E such that for each conformal e-predictor 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*

$$G(z_1, \dots, z_n, z_{n+1}) \le 2, \tag{16}$$

$$E'(z_1, \dots, z_n, (x_{n+1}, y)) \le 2.01, \tag{17}$$

$$E(z_1, \dots, z_n, (x_{n+1}, y)) \ge c\,|\mathbf{Y}|\,. \tag{18}$$

To compare Theorem 9 with Theorem 8, notice that (16)–(18) imply

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

Therefore, Theorem 9 implies Theorem 8, but only if we ignore the constant factor $4.02\mathrm{e} < 11$.

Proof sketch of Theorem 9. 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 $Y \equiv -Y_1 - \dots - Y_n \pmod{|\mathbf{Y}|}$; capital letters are used to emphasize that the labels (elements of $\mathbf{Y}$) are random. The idea is to prove that the three events

$$G(Y_1, \dots, Y_n, Y_{n+1}) \le 2, \tag{19}$$

$$E'(Y_1, \dots, Y_n, Y) \le 2.01, \tag{20}$$

$$E(Y_1, \dots, Y_n, Y) \ge c\,|\mathbf{Y}| \tag{21}$$

(cf. (16)–(18), 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.

Since $Y_1, \dots, Y_{n+1}$ are IID and $Y_1, \dots, Y_n, Y$ are exchangeable, the probabilities of (19) and (20) are bounded by Markov's inequality. And we can define E to ensure that (21) holds with a small probability since $Y_1, \dots, Y_n, Y$ are not IID: were they IID, we would typically expect their sum to be divisible by $|\mathbf{Y}|$ with probability close to $1/|\mathbf{Y}|$ for a large n.

Further details are given in Sect. A.5. □

5 Conclusion

This paper gives explicit statements, not involving any unspecified constants, of universality of conformal predictors. Namely, for each train-invariant randomness predictor there is a conformal predictor that is competitive with it. Some constants is these statements have been shown to be optimal.

These are some directions of further research:

- This paper shows that the attainable improvement on conformal prediction under the assumption of randomness is limited, such as a factor of e in the e-values for false labels (see, e.g., (7)). Can we develop practically useful predictors exploiting such potential improvements? (For a toy example, see Remark 10 in Sect. A.2. Some experimental results are reported in [15].)
- Can we connect $\mathcal{P}^{\mathrm{tR}}$ and $\mathcal{P}^{\mathrm{tX}}$ (in the spirit of Corollary 7) directly, without a detour via e-values?
- A related remark is that all our optimality results (Theorems 6, 8, and 9) cover only e-prediction. It would be ideal to have similar optimality results for Corollary 7 or its stronger versions.
- The assumption of randomness is very strong, and there has been extensive work devoted to relaxing this assumption; see, e.g., [13] and [1, Chap. 7]. To what degree do the results of this paper carry over to more general settings?

Acknowledgments

In October 2024 Vladimir V'yugin, one of my teachers in the algorithmic theory of randomness, passed away 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.

Anonymous reviewers' comments were helpful in improving the presentation.

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, December 2025. Prepublication version of a book to be published by Cambridge University Press.

[2] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

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

[4] Donald A. S. Fraser. p-values: the insight to modern statistical inference. *Annual Review of Statistics and its Application*, 4:1–14, 2017.

[5] Donald A. S. Fraser. The p-value function and statistical inference. *American Statistician*, 73sup1:135–147, 2019.

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

[7] 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.

[8] Olli S. Miettinen. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. Wiley, New York, 1985.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, 166:111674, 2025.

[15] Vladimir Vovk. Inductive randomness predictors: beyond conformal. *Proceedings of Machine Learning Research*, 266:6–33, 2025. COPA 2025.

[16] Vladimir Vovk. Conditionality principle under unconstrained randomness. *Statistical Science*, 41:218–221, 2026.

[17] Vladimir Vovk. Exchangeability and randomness for infinite and finite sequences. Technical Report arXiv:2512.22162 [math.ST], arXiv.org e-Print archive, January 2026.

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

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

A Proofs

This appendix gives detailed proofs of all results stated in the main paper. The *Bernoulli model* is defined as the statistical model $(B_\theta^{n+1} : \theta \in [0,1])$, where B_θ is the *Bernoulli measure* on $\{0,1\}$, defined by $B_\theta(\{1\}) := \theta \in [0,1]$.

A.1 Proof of Theorem 5

We will define G as $G_2 G_3$, where $G_2 \in \mathcal{E}^{\mathrm{iR}}$ and $G_3 \in \mathcal{E}^{\mathrm{X}}$ (it is obvious that these two inclusions will imply $G \in \mathcal{E}^{\mathrm{R}}$). First we define an approximation G_1 to G_2 as

$$G_1(z_1, \dots, z_{n+1}) := \frac{1}{n+1} \sum_{i=1}^{n+1} \int E^{\mathrm{i}}(z_1, \dots, z_{i-1}, (x_i, y), z_{i+1}, \dots, z_{n+1}) B(\mathrm{d}y \mid z_i).$$

In other words, $G_1(z_1, \dots, z_{n+1})$ is obtained by randomly (with equal probabilities) choosing an example z_i in the data sequence $z_1, \dots, z_{n+1}$, replacing its label y_i by a random label $y \sim B(\cdot \mid z_i)$, and finding the expectation of E^{i} on $z_1, \dots, z_{n+1}$ modified in this way. We can see that G_1 is invariant, but it does not have to be in $\mathcal{E}^{\mathrm{iR}}$. The invariant randomness e-variable G_2 is defined similarly, except that now we replace each label y_i, $i = 1, \dots, n+1$, by a random label $y \sim B(\cdot \mid z_i)$ with probability $\frac{1}{n+1}$ (all independently). The key observation is that $G_2/G_1 \geq 1/\mathrm{e}$, which follows from the probability that exactly one label will be changed in the construction of G_2 being

$$(n+1)\frac{1}{n+1}\left(\frac{n}{n+1}\right)^n \geq 1/\mathrm{e}.$$

Finally, $G_3 \in \mathcal{E}^{\mathrm{X}}$ is defined by

$$G_3(z_1, \dots, z_{n+1}) := \frac{\int E^{\mathrm{i}}(z_1, \dots, z_n, (x_{n+1}, y)) B(\mathrm{d}y \mid z_{n+1})}{G_1(z_1, \dots, z_{n+1})}.$$

Combining all these statements, we get the following randomness e-variable G':

$$\begin{aligned} G'(z_1, \dots, z_{n+1}) &:= G_2(z_1, \dots, z_{n+1}) G_3(z_1, \dots, z_{n+1}) \\ &\geq \mathrm{e}^{-1} G_1(z_1, \dots, z_{n+1}) G_3(z_1, \dots, z_{n+1}) \\ &= \mathrm{e}^{-1} \int E^{\mathrm{i}}(z_1, \dots, z_n, (x_{n+1}, y)) B(\mathrm{d}y \mid z_{n+1}). \end{aligned} \tag{22}$$

By the definition (5), (6) is equal to the expression in (22), and so $G' \in \mathcal{E}^{\mathrm{R}}$ implies that (6) is also a randomness e-variable.

A.2 Proof of Theorem 6

Without loss of generality we assume $|\mathbf{X}| = 1$ (so that the objects become uninformative and we can omit them from our notation) and $\mathbf{Y} = \{-1, 1\}$

(with the discrete σ-algebra). Define a randomness e-variable E by

$$E(y_1, \dots, y_{n+1}) := \begin{cases} \left(1 - \frac{1}{n+1}\right)^{-n} & \text{if } k = 1 \\ 0 & \text{if not,} \end{cases} \tag{23}$$

where k is the number of 1s among $y_1, \dots, y_{n+1}$. This is indeed a randomness e-variable, since the maximum probability of $k = 1$ in the Bernoulli model, $(n+1)\theta(1-\theta)^n \to \max$, is attained at $\theta = \frac{1}{n+1}$. The corresponding exchangeability e-variable is

$$E^{\mathrm{X}}(y_1, \dots, y_{n+1}) = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{if not.} \end{cases}.$$

Both E and E^{X} are train-invariant (and even invariant). Let B just flip the label: $B(\{-y\} \mid y) = 1$. Suppose Theorem 5 holds with the e^{-1} in (6) replaced by $c > \mathrm{e}^{-1}$. Then the randomness e-variable G satisfies

$$G(0, \dots, 0) = c\left(1 - \frac{1}{n+1}\right)^{-n} \sim c\mathrm{e} > 1,$$

which is impossible for a large enough n (since the probability measure concentrated on $(0, \dots, 0)$ is of the form Q^{n+1}).

Remark 10. Whereas the randomness e-variable E defined by (23) is all we need to prove Theorem 6, it is not useful for prediction. A variation on (23) that can be used in prediction is

$$E(y_1, \dots, y_{n+1}) := \begin{cases} (n+1)\left(1 - \frac{1}{n+1}\right)^{-n} & \text{if } (y_1, \dots, y_n, y_{n+1}) = (0, \dots, 0, 1) \\ 0 & \text{if not.} \end{cases}$$

According to this randomness e-predictor, after observing n 0s in a row, we are likely to see 0 rather than 1. This is a version of Laplace's rule of succession. While under randomness we have $E(0, \dots, 0, 1) \sim \mathrm{e}n$, under exchangeability we can only achieve $E^{\mathrm{X}}(0, \dots, 0, 1) = n + 1 \sim n$.

A.3 Proof of Corollary 7

Fix $\delta \in (0, 1)$ and $P \in \mathcal{P}^{\mathrm{R}}$. Set $E := \delta P^{\delta - 1}$ and $P' := 1/E^{\mathrm{X}}$, so that $E \in \mathcal{E}^{\mathrm{R}}$ and $P' \in \mathcal{P}^{\mathrm{X}}$. According to (6),

$$\begin{aligned} G(z_1, \dots, z_n, z_{n+1}) &:= \mathrm{e}^{-1} \int \frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, y))} \, B(\mathrm{d}y \mid z_{n+1}) \\ &= \mathrm{e}^{-1} \int \frac{\delta P'(z_1, \dots, z_n, (x_{n+1}, y))}{P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, y))} \, B(\mathrm{d}y \mid z_{n+1}) \end{aligned}$$

is a randomness e-variable.

A.4 Proof of Theorem 8

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 fixed object $x_0 \in \mathbf{X}$, 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 disjoint 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 (24) 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 $E \in \mathcal{E}^{\mathrm{tR}}$ (which ignores the objects) as follows:

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

$$\sum_{i=1}^{n} y_i - n/2 = y, \tag{24}$$

  E takes value $ae\sqrt{\pi/2}n^{3/2}$, where $a < 1$ is a positive constant (it will be taken close to 1 later in the proof);

- E takes value 0 on all other sequences in $\mathbf{Z}^{n+1}$.

In (24), y is determined by $y_1 + \dots + y_n$ and vice versa. This agrees with the idea of the proof of Theorem 8 given after its statement: since $y_1, \dots, y_n$ are binary, $y_1 + \dots + y_n$ carries the same information as $\wr y_1, \dots, y_n \int$.

First we check that E is indeed a train-invariant randomness e-variable. Let the underlying probability space be $\mathbf{Z}^{n+1}$ equipped with a probability measure $R = Q^{n+1}$, so that individual examples are generated independently from Q. We will use the notation Z_i, $i = 1, \dots, n+1$, for z_i considered as a random example (formally, Z_i is the random element defined by $Z_i(z_1, \dots, z_{n+1}) := z_i$) and the notation Y_i, $i = 1, \dots, n+1$, for the label of Z_i. The maximum probability of the event $E(Z_1, \dots, Z_{n+1}) > 0$ is attained for Q giving maximum probabilities to the following two events:

1. The random labels $Y_1, \dots, Y_n$ take values in $\{0', 1'\}$, and the remaining random label Y_{n+1} takes value in $\{-k, \dots, k\}$.

2. Conditionally on the first event, we have (24), where y is the value taken by Y_{n+1} and $y_1, \dots, y_n$ are the values taken by $Y_1, \dots, Y_n$.

The maximum probability of the first event (in item 1) is

$$\left(1 - \frac{1}{n+1}\right)^n \frac{1}{n+1} \sim \frac{1}{en},$$

which is obtained by maximizing $(1-\theta)^n\theta$ over θ and where the asymptotic equivalence is as $n \to \infty$. The maximum probability of the second event (in item 2) conditional on both the first event and $Y_{n+1} = y \in \{-k, \dots, k\}$ is asymptotically equivalent, by the local limit theorem [12, 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 $Y_1, \dots, Y_n$ being distributed as B^n_θ, and the maximum over θ being attained at $\theta = 1/2 + y/n$. Since $a < 1$, $E \in \mathcal{E}^{\mathrm{tR}}$ for a sufficiently large n. Notice that our argument in this paragraph only uses $k \ll n$, since in our application of the local limit theorem the exponential term $\exp(\dots)$ was 1.

Let us now generate $Y_1, \dots, Y_{n+1}$ from the Bernoulli model and find the maximum probability that

$$\sum_{i=1}^{n} Y_i - n/2 \in \{-k, \dots, k\} \tag{25}$$

(so that this condition does not involve Z_{n+1}). Again by the local limit theorem, the maximum probability 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 $\theta := 1/2$. 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. Therefore, for any $G \in \mathcal{E}^{\mathrm{R}}$, there are $y_1, \dots, y_{n+1} \in \{0', 1'\}$ such that

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

(remember that we are omitting x_0), where $b > 1$ is a constant (to be chosen close to 1 later), and

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

(cf. (25)). Fix such $y_1, \dots, y_{n+1}$ and set

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

(cf. (24)). Taking a and b sufficiently close to 1, we obtain

$$\frac{G(y_1, \dots, y_{n+1})}{E^{\mathrm{i}}(y_1, \dots, y_n, y)} = (n+1)\frac{G(y_1, \dots, y_{n+1})}{E(y_1, \dots, y_n, y)} \le \frac{b\sqrt{\pi n}(n+1)}{2\sqrt{2}k a \mathrm{e}\sqrt{\pi/2}n^{3/2}} < \frac{c}{\mathrm{e}\,|\mathbf{Y}|} \tag{26}$$

for a sufficiently large k. The equality "=" in (26) follows from $\{0', 1'\}$ and $\{-k, \dots, k\}$ being disjoint sets, the inequality "$\leq$" follows from the definitions of E and G, and the inequality "$<$" follows from $|\mathbf{Y}| = 2k + 3$. This proves Theorem 8 since $E^{\mathrm{i}} = E/E^{\mathrm{X}}$.

Finally, the formal meaning of the condition $1 \ll k \ll \sqrt{n}$ is that the first sentence in the statement of Theorem 8 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 $\sqrt{n} \geq C\,|\mathbf{Y}|$."

A.5 Proof of Theorem 9

This section spells out details omitted in the proof sketch given in Sect. 4. As mentioned there, the probability of (19) being at least 0.5 and the probability of (20) being at least 0.502 follow immediately from Markov's inequality. The notation $Y_1, \dots, Y_{n+1}$ was introduced formally in Sect. A.4 (and used informally already in the proof sketch of Theorem 9).

Now let us define E:

$$E(y_1, \dots, y_{n+1}) := \begin{cases} c\,|\mathbf{Y}| & \text{if } S \equiv 0 \pmod{|\mathbf{Y}|} \text{ and } \forall y \in \mathbf{Y} : \left|k_y - n/|\mathbf{Y}|\right| \leq 0.1n/|\mathbf{Y}| \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

where S is the sum of $y_1, \dots, y_{n+1}$ and $k_y := |\{i \in \{1, \dots, n+1\} \mid y_i = y\}|$ is the number of times $y \in \mathbf{Y}$ occurs in the sequence $y_1, \dots, y_{n+1}$. We are required to prove two statements for a large enough n: first, that $E \in \mathcal{E}^{\mathrm{R}}$, and second, that the probability of $E(Y_1, \dots, Y_n, Y) > 0$ is at least 0.999 (see (21)).

The second statement is easier. Let $y_1, \dots, y_{n+1}$ be the values taken by $Y_1, \dots, Y_n, Y$, respectively. Then the first condition

$$S \equiv 0 \pmod{|\mathbf{Y}|} \tag{28}$$

in (27) holds automatically, and the second condition

$$\forall y \in \mathbf{Y} : \left|k_y - n/|\mathbf{Y}|\right| \leq 0.1n/|\mathbf{Y}| \tag{29}$$

holds with probability that tends to 1 as $n \to \infty$ for each $y \in \mathbf{Y}$, by the central limit theorem. Since $\mathbf{Y}$ is a finite set, the second statement has been proved.

Now let us prove the first statement. Let Q be a probability measure on $\mathbf{Y}$ and set $R := Q^{n+1}$. We will denote by A the event given after "if" in (27), namely the conjunction of (28) and (29). It suffices to prove that asymptotically as $n \to \infty$ the R-probability of A does not exceed $c^{-1/2}/|\mathbf{Y}|$, uniformly in Q. Consider two cases:

- If $Q(\{y\}) \geq 0.1/|\mathbf{Y}|$ for all $y \in \mathbf{Y}$, the convergence $R(A_1) \to 1/|\mathbf{Y}|$ as $n \to \infty$ for the superset A_1 of A given by (28) follows from the asymptotic uniformity result of Horton and Smith [7, Theorem (5.01)] (proved independently by Dvoretzky and Wolfowitz [3, Theorem 2]). The convergence is uniform on this compact set of Q.

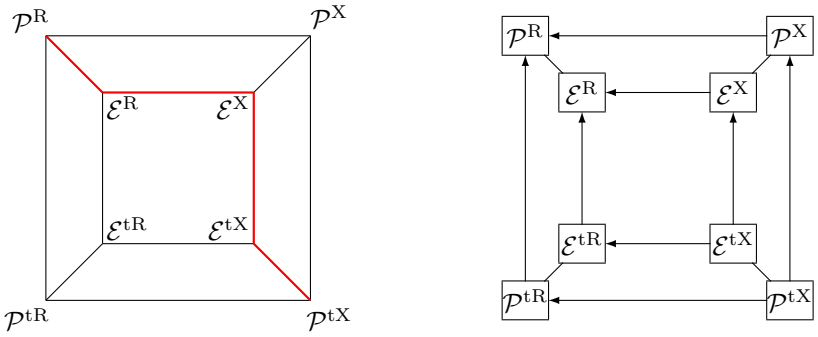

Figure 2: A cube representing the main classes of predictors considered in Appendix B. The right panel shows embeddings as arrows, as in Figure 1.

- If $Q(\{y\}) \le 0.1/|\mathbf{Y}|$ for some $y \in \mathbf{Y}$, the convergence $R(A_2) \to 0$ as $n \to \infty$ for the superset A_2 of A given by (29) follows from, e.g., the central limit theorem combined with the Bonferroni correction for the multiplicity of y. The convergence is again uniform on this compact set of Q.

Combining the two cases, we get $E \in \mathcal{E}^{\mathrm{R}}$ (and so, obviously, $E \in \mathcal{E}^{\mathrm{tR}}$).

B Reducing randomness predictors to conformal

In the main paper we assumed the train-invariance of randomness predictors. In this appendix we drop this assumption, since statistical principles may go astray (see, e.g., the criticism [16] of the conditionality principle from the vantage point of conformal prediction). Besides, in some cases we may have convincing reasons to violate train-invariance, such as using inductive conformal predictors [18, Sect. 4.2] for the purpose of computational efficiency.

The connections between various classes of predictors (introduced in the main paper) that we use or explore in this appendix are shown in red in the left-hand panel of Figure 2. Namely, we are interested in connections between:

(a) randomness e-predictors and exchangeability e-predictors;

(b) exchangeability e-predictors and conformal e-predictors;

(c) randomness p-predictors and randomness e-predictors;

(d) conformal e-predictors and conformal p-predictors.

Theorem 5 stated in Subsect. 3.1 is general enough to cover connection (a). Connection (b) will be discussed in Sect. B.2. Connections (c) and (d) result

from the possibility of converting p-values to e-values and back, as discussed in Subsect. 3.2.

In Sect. B.3 we will see how the connections shown in red in Figure 2 can be combined to demonstrate the universality of conformal prediction without accepting the train-invariance principle. Section B.4 applies this to classification. But we start in Sect. B.1 by introducing an operator making randomness predictors (in particular, exchangeability predictors) train-invariant.

B.1 Another operator

In addition to the operators (4) and (5), we will need another operator,

$$E^{\mathrm{t}}(z_1,\dots,z_{n+1}) := \frac{1}{n!}\sum_{\sigma} E(z_{\sigma(1)},\dots,z_{\sigma(n)},z_{n+1}), \tag{30}$$

σ ranging over the permutations of $\{1,\dots,n\}$. Now we have $E^{\mathrm{t}} \in \mathcal{E}^{\mathrm{tX}}$ (resp. $E^{\mathrm{t}} \in \mathcal{E}^{\mathrm{tR}}$) whenever $E \in \mathcal{E}^{\mathrm{X}}$ (resp. $E^{\mathrm{t}} \in \mathcal{E}^{\mathrm{R}}$). Both operators (4) and (30) are kinds of averaging: while $E \mapsto E^{\mathrm{i}}$ averages over all permutations of an input data sequence (including both training and test examples), $E \mapsto E^{\mathrm{t}}$ averages over the permutations of the training sequence only. And while $E \mapsto E^{\mathrm{i}}$ is polynomially computable for train-invariant E, $E \mapsto E^{\mathrm{t}}$ requires exponential computation time in general.

Using two of the three operators (4), (5), and (30), we can turn any randomness e-variable E to an exchangeability e-variable E^{X} to a conformal e-variable $(E^{\mathrm{X}})^{\mathrm{t}}$. The following lemma shows that the order in which the last two operators are applied does not matter.

Lemma 11. *The operators* $^{\mathrm{t}}$ *and* $^{\mathrm{X}}$ *commute: for any* $E \in \mathcal{E}$, $(E^{\mathrm{t}})^{\mathrm{X}} = (E^{\mathrm{X}})^{\mathrm{t}}$.

Proof. Let us fix a data sequence $z_1,\dots,z_{n+1}$ and check

$$(E^{\mathrm{t}})^{\mathrm{X}}(z_1,\dots,z_{n+1}) = (E^{\mathrm{X}})^{\mathrm{t}}(z_1,\dots,z_{n+1}).$$

As functions of a permutation of $z_1,\dots,z_{n+1}$, E and E^{X} are proportional to each other, and therefore, E^{t} and $(E^{\mathrm{X}})^{\mathrm{t}}$ are also proportional to each other. This implies $(E^{\mathrm{t}})^{\mathrm{X}} = (E^{\mathrm{X}})^{\mathrm{t}}$ on the permutations of $z_1,\dots,z_{n+1}$. And this is true for each $(z_1,\dots,z_{n+1})$. □

We will let $^{\mathrm{tX}}$ stand for the composition of the two operators:

$$E^{\mathrm{tX}} := (E^{\mathrm{t}})^{\mathrm{X}} = (E^{\mathrm{X}})^{\mathrm{t}}.$$

Let us say that $E \in \mathcal{E}^{\mathrm{X}}$ is *admissible* if (3) always holds with "=" in place of "$\le$". (This agrees with the standard notion of admissibility in statistical decision theory.) The intuition (which can be formalized easily) behind the four operators that we have introduced in this paper is that:

- $^{\mathrm{i}}$ projects $\mathcal{E}^{\mathrm{R}}$ onto $\mathcal{E}^{\mathrm{iR}}$; it also projects the admissible part of $\mathcal{E}^{\mathrm{X}}$ onto the identical 1;

- $^{\mathrm{X}}$ projects $\mathcal{E}$ onto the admissible part of $\mathcal{E}^{\mathrm{X}}$;
- $^{\mathrm{t}}$ projects $\mathcal{E}^{\mathrm{X}}$ onto $\mathcal{E}^{\mathrm{tX}}$ and $\mathcal{E}^{\mathrm{R}}$ onto $\mathcal{E}^{\mathrm{tR}}$;
- $^{\mathrm{tX}}$ projects $\mathcal{E}$ onto the admissible part of $\mathcal{E}^{\mathrm{tX}}$.

In particular, these operators are idempotent:

$$(E^{\mathrm{i}})^{\mathrm{i}} = E^{\mathrm{i}}, \quad (E^{\mathrm{X}})^{\mathrm{X}} = E^{\mathrm{X}}, \quad (E^{\mathrm{t}})^{\mathrm{t}} = E^{\mathrm{t}}, \quad (E^{\mathrm{tX}})^{\mathrm{tX}} = E^{\mathrm{tX}}$$

(and we can allow any $E \in \mathcal{E}$ here). Despite these operators being projections, we cannot claim that these ways of moving between different classes of predictors are always optimal.

Lemma 11 lists the only two cases where the combination of two of our three basic operators ($^{\mathrm{i}}$, $^{\mathrm{X}}$, and $^{\mathrm{t}}$) gives something interesting. The other four cases are:

$$(E^{\mathrm{X}})^{\mathrm{i}} = (E^{\mathrm{i}})^{\mathrm{X}} = 1, \quad (E^{\mathrm{i}})^{\mathrm{t}} = (E^{\mathrm{t}})^{\mathrm{i}} = E^{\mathrm{i}}.$$

B.2 Efficiency of making predictors train-invariant

To state our result in its strongest form, we define a *test-conditional exchangeability e-variable* $G = G(z_1, \dots, z_n, z_{n+1})$ as an element of $\mathcal{E}$ satisfying

$$\forall (z_1, \dots, z_{n+1}) \; \forall \sigma : \frac{1}{n!} \sum_{\sigma} G(z_{\sigma(1)}, \dots, z_{\sigma(n)}, z_{n+1}) \le 1,$$

σ ranging over the permutations of $\{1, \dots, n\}$. Such G form a subclass of $\mathcal{E}^{\mathrm{X}}$ (and therefore, of $\mathcal{E}^{\mathrm{R}}$).

Theorem 12. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel. For each exchangeability e-predictor E,*

$$G(z_1, \dots, z_n, z_{n+1}) := \int \frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{t}}(z_1, \dots, z_n, (x_{n+1}, y))} \, B(\mathrm{d}y \mid z_{n+1}) \tag{31}$$

(with $0/0$ interpreted as 0) is a test-conditional exchangeability e-variable.

The interpretation of (31) is similar to that of (6). It will be clear from the proof that we can allow E to be any randomness e-predictor, or even any element of $\mathcal{E}$.

Proof of Theorem 12. Let us check that the right-hand side of (31) is a test-conditional exchangeability e-variable:

$$\frac{1}{n!} \sum_{\sigma} G(z_{\sigma(1)}, \dots, z_{\sigma(n)}, z_{n+1})$$
$$= \frac{1}{n!} \sum_{\sigma} \int \frac{E(z_{\sigma(1)}, \dots, z_{\sigma(n)}, (x_{n+1}, y))}{E^{\mathrm{t}}(z_1, \dots, z_n, (x_{n+1}, y))} B(\mathrm{d}y \mid z_{n+1})$$

$$= \int \frac{E^{\mathrm{t}}(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{t}}(z_1, \dots, z_n, (x_{n+1}, y))} B(\mathrm{d}y \mid z_{n+1}) \le 1$$

(where we have " ≤ 1" rather than " $= 1$" because of our interpretation of $0/0$). □

B.3 Putting everything together

The following theorem combines Theorems 5 and 12 and establishes a connection between randomness and conformal e-predictors, without accepting the train-invariance principle. Remember that the conformal e-predictor $\mathcal{E}^{\mathrm{tX}}$ derived from a randomness e-predictor E is obtained by combining the operators (5) and (30), i.e., as

$$E^{\mathrm{tX}}(z_1, \dots, z_{n+1}) := (n+1) \frac{\sum_\sigma E(z_{\sigma(1)}, \dots, z_{\sigma(n)}, z_{n+1})}{\sum_\pi E(z_{\pi(1)}, \dots, z_{\pi(n+1)})},$$

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

Corollary 13. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel. For each randomness e-predictor E,*

$$G(z_1, \dots, z_n, z_{n+1}) := \mathrm{e}^{-1/2} \int \sqrt{\frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{tX}}(z_1, \dots, z_n, (x_{n+1}, y))}} \, B(\mathrm{d}y \mid z_{n+1}) \quad (32)$$

is a randomness e-variable.

The main weakness of Corollary 13 is the presence of the term $\mathrm{e}^{-1/2}$, but it might be inevitable.

Proof. Applying the Cauchy–Schwarz inequality, we have, for some $G_1, G_2, G_3 \in \mathcal{E}^{\mathrm{R}}$,

$$\begin{aligned}
&\mathrm{e}^{-1/2} \int \sqrt{\frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{tX}}(z_1, \dots, z_n, (x_{n+1}, y))}} \, B(\mathrm{d}y \mid z_{n+1}) \\
&= \mathrm{e}^{-1/2} \int \sqrt{\frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, y))}} \sqrt{\frac{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{tX}}(z_1, \dots, z_n, (x_{n+1}, y))}} \, B(\mathrm{d}y \mid z_{n+1}) \\
&\le \sqrt{\mathrm{e}^{-1} \int \frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, y))} \, B(\mathrm{d}y \mid z_{n+1})} \\
&\quad \times \sqrt{\int \frac{E^{\mathrm{X}}(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{tX}}(z_1, \dots, z_n, (x_{n+1}, y))} B(\mathrm{d}y \mid z_{n+1})} \\
&= \sqrt{G_1(z_1, \dots, z_{n+1}) G_2(z_1, \dots, z_{n+1})} \le G_3(z_1, \dots, z_{n+1})
\end{aligned}$$

(the existence of G_1 and G_2 follows from Theorems 5 and 12, respectively, and G_3 can be set, e.g., to the arithmetic average of G_1 and G_2). □

Similarly to Corollary 7, we can adapt Corollary 13 to p-predictors.

Corollary 14. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel and let $\delta \in (0,1)$. For each randomness p-predictor P there exists a conformal p-predictor P' such that*

$$G(z_1, \dots, z_n, z_{n+1}) := \sqrt{\frac{\delta}{\mathrm{e}}} \int \sqrt{\frac{P'(z_1, \dots, z_n, (x_{n+1}, y))}{P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, y))}}\, B(\mathrm{d}y \mid z_{n+1}) \quad (33)$$

is a randomness e-variable.

The interpretation of (33) is similar to that of (9): $P'(z_1, \dots, z_n, (x_{n+1}, y))$ is typically small when $P(z_1, \dots, z_n, (x_{n+1}, y))$ is small. Instead of (10) we will have

$$\sqrt{\frac{\delta}{\mathrm{e}}} \sqrt{\frac{P'(z_1, \dots, z_n, (x_{n+1}, y))}{P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, y))}} < \frac{1}{\epsilon_1 \epsilon_2},$$

which can be rewritten as

$$P'(z_1, \dots, z_n, (x_{n+1}, y)) < \frac{\mathrm{e}}{\delta \epsilon_1^2 \epsilon_2^2} P^{1-\delta}(z_1, \dots, z_n, (x_{n+1}, y))$$

in place of (11).

Proof of Corollary 14. We proceed as in the proof of Corollary 7 except for replacing (6) by (32). □

B.4 Applications to classification

As compared with Sect. 4, we get similar but weaker performance guarantees for the derived conformal predictors without accepting the train-invariance principle. In the binary case $\mathbf{Y} = \{-1, 1\}$, we apply (32) to obtain

$$G(z_1, \dots, z_n, z_{n+1}) := \mathrm{e}^{-1/2} \sqrt{\frac{E(z_1, \dots, z_n, (x_{n+1}, -y_{n+1}))}{E^{\mathrm{tX}}(z_1, \dots, z_n, (x_{n+1}, -y_{n+1}))}}$$

in place of (7). This implies (8) with $\mathcal{E}^{\mathrm{tX}}$ in place of $\mathcal{E}^{\mathrm{X}}$ and ϵ^2 in place of ϵ.

Instead of (12) now we have

$$G(z_1, \dots, z_n, z_{n+1}) := \frac{\mathrm{e}^{-1/2}}{|\mathbf{Y}| - 1} \sum_{y \in \mathbf{Y} \setminus \{y_{n+1}\}} \sqrt{\frac{E(z_1, \dots, z_n, (x_{n+1}, y))}{E^{\mathrm{tX}}(z_1, \dots, z_n, (x_{n+1}, y))}},$$

and instead of (14) we have

$$\forall y \in \mathbf{Y} \setminus \{y_{n+1}\} : E^{\mathrm{tX}}(z_1, \dots, z_n, (x_{n+1}, y)) > \frac{\mathrm{e}^{-1} \epsilon^2}{(|\mathbf{Y}| - 1)^2} E(z_1, \dots, z_n, (x_{n+1}, y)).$$