Set and functional prediction: randomness, exchangeability, and conformal

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

February 26, 2025

Project web site:
http://alrw.net

Abstract

This paper continues the study of the efficiency of conformal prediction as compared with more general randomness prediction and exchangeability prediction. It does not restrict itself to the case of classification, and our results will also be applicable to the case of regression. The price to pay is that efficiency will be attained only on average, albeit with respect to a wide range of probability measures on the label space.

Contents

1 Introduction

Conformal prediction is usually presented as a method of *set prediction* [10, 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., [4]), 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 will 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).

The most standard assumption in machine learning is that of *randomness* (i.e., the data are assumed to be produced in the IID fashion). This paper is a development of [8], which introduces the most general class of predictors, "randomness predictors", that produce prediction functions that are valid, in the same sense as conformal predictors, under the assumption of randomness. There are many more randomness predictors than conformal predictors, and an interesting question is whether there are randomness predictors that have significant advantages (e.g., in practice) over conformal predictors. This question was answered (albeit imperfectly) in [8] in the case of classification with few (such as two) classes. In this paper we will be interested in arbitrary label spaces, including the case of regression. The message of this paper is similar to that of [8]: the difference between conformal and randomness prediction is not huge, but it remains an open question whether it can be usefully exploited. This paper strengthens most of the positive results in [8], but its only negative result is much weaker (but also simpler) than the negative results of [8].

A useful technical tool in establishing connections between conformal and randomness predictors is provided by conformal e-predictors [9], which are obtained by replacing p-values with e-values. 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 [9, 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 Glenn Shafer's opinion [7], 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”.

In this paper we will follow the scheme for establishing the closeness of conformal and randomness predictors used in [8, Figure 1]. Namely, we will establish connections between:

(a) conformal p-predictors and conformal e-predictors;

(b) conformal e-predictors and exchangeability e-predictors, where the class of exchangeability e-predictors is intermediate between the conformal e-predictors and the randomness e-predictors, and it consists of randomness e-predictors that are valid under the assumption of exchangeability (which is weaker than the assumption of randomness in the case of finite data sequences);

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

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

Steps (a) and (d) (converting p-values to e-values and back) are standard, and we will mainly concentrate on steps (b) and (c).

We start in Sect. 2 from the main definitions, and Sect. 3 is devoted to the main results. In particular, we establish the efficiency of conformal predictors among randomness predictors in both p- and e-versions. Namely, the prediction functions for conformal predictors turn out 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. Sections 4 and 5 give some applications to classification and regression, respectively, and Sect. 6 concludes.

2 Definitions

This paper is about the following prediction problem (continuing the discussion started in [8]). 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. 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* $\alpha \in (0,1)$, $Q^{n+1}(\{P \le \alpha\}) \le \alpha$. And a measurable $P : \mathbf{Z}^{n+1} \to [0,1]$ is a *conformal p-variable* if

- $R(\{P \le \alpha\}) \le \alpha$ for any exchangeable probability measure R on $\mathbf{Z}^{n+1}$ and any $\alpha \in (0,1)$;

- it is *training-invariant*, i.e., invariant w.r. to permutations of the training examples:

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

  for each data sequence $z_1, \ldots, z_{n+1}$ and each permutation σ of $\{1, \ldots, n\}$ (training-invariant functions were called simply invariant in [8]).

We will sometimes refer to the values taken by p-variables as *p-values*, and our notation for the classes of all randomness and conformal p-variables will be $\mathcal{P}^{\mathrm{R}}$ 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*. Notice that the standard expression "training set" is only justified for predictors P satisfying (1) (and even in this case it is not justified completely; it would be more accurate to say "training bag"). There are several ways to package the output of conformal p-predictors. One is in terms of set prediction: for each significance level $\alpha \in (0, 1)$, each training sequence $z_1, \ldots, z_n$, and each test object x_{n+1}, we can output the *prediction set*

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

By the definition of a conformal p-variable, 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. See [8, Sect. 2] for a more detailed discussion of connections between conformal p-variables and conformal p-prediction.

Instead of predicting with one prediction set in the family (2), we can package our prediction as the *prediction function*

$$f(y) := P(z_1, \ldots, z_n, x_{n+1}, y), \quad y \in \mathbf{Y}. \tag{3}$$

We may refer to this mode of prediction as *functional prediction*. The step from set prediction to functional prediction is analogous from the step from confidence intervals to p-value functions (see, e.g., [6, Sect. 9] and [3–5] for the latter).

Remark 1. 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 (2) 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 (3) for randomness p-predictors as well.

Two important desiderata for conformal and randomness predictors are their validity and efficiency. In terms of the prediction function f, validity concerns the value $f(y_{n+1})$ at the true label (its typical values should not be too small in p-prediction), and efficiency concerns the values $f(y)$ at the false labels $y \neq y_{n+1}$ (they should be as small as possible in p-prediction). Validity is automatic under randomness (and even 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 (Theorems 7 and 8 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 $E : \mathbf{Z} \to [0,\infty$ is denoted by $\mathcal{E}$.

The class $\mathcal{E}^{\mathrm{tX}}$ of *conformal e-variables* consists of all functions $E \in \mathcal{E}^{\mathrm{X}}$ that are training-invariant:

$$E(z_{\sigma(1)},\dots,z_{\sigma(n)},z_{n+1}) = E(z_1,\dots,z_n,z_{n+1}) \tag{4}$$

for each data sequence $z_1,\dots,z_{n+1}$ and each permutation σ of $\{1,\dots,n\}$. 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 (3), they output prediction functions

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

For conformal and randomness e-predictors, validity and efficiency change direction: for validity, typical values $f(y_{n+1})$ should not be too large, and for efficiency typical values $f(y)$ at the false labels $y \neq y_{n+1}$ should be as large as possible. Again validity is automatic under randomness, and Theorem 7 below establishes efficiency guarantees for conformal e-prediction in terms of randomness e-prediction.

We will also need two important subclasses of $\mathcal{E}^{\mathrm{R}}$. The subclass $\mathcal{E}^{\mathrm{tR}}$ of $\mathcal{E}^{\mathrm{R}}$ consists of all functions $E \in \mathcal{E}^{\mathrm{R}}$ that are training-invariant (i.e., satisfy (4)). The subclass $\mathcal{E}^{\mathrm{iR}}$ of $\mathcal{E}^{\mathrm{R}}$ consists of all functions $E \in \mathcal{E}^{\mathrm{R}}$ 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* (this is almost the same thing as *configuration randomness e-variables* in [8]).

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 $E \in \mathcal{E}^{\mathrm{R}}$, three derivative e-variables:

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

$$E^{\mathrm{X}}(z_1,\dots,z_{n+1}) := \frac{E(z_1,\dots,z_{n+1})}{E^{\mathrm{i}}(z_1,\dots,z_{n+1})}, \tag{6}$$

$$E^{\mathrm{t}}(z_1,\dots,z_{n+1}) := \frac{1}{n!}\sum_{\sigma} E(z_{\sigma(1)},\dots,z_{\sigma(n)},z_{n+1}), \tag{7}$$

π ranging over the permutations of $\{1,\dots,n+1\}$ and σ ranging over the permutations of $\{1,\dots,n\}$. It is clear that $E^{\mathrm{i}} \in \mathcal{E}^{\mathrm{iR}}$ whenever $E \in \mathcal{E}^{\mathrm{R}}$, that $E^{\mathrm{X}} \in \mathcal{E}^{\mathrm{X}}$ for all $E \in \mathcal{E}$, and that $E^{\mathrm{t}} \in \mathcal{E}^{\mathrm{tX}}$ whenever $E \in \mathcal{E}^{\mathrm{X}}$. The operators (5) and (7) 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.

Using two of these three operators, 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 2. *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:

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

It is easy to see that $E \in \mathcal{E}$ belongs to $\mathcal{E}^{\mathrm{X}}$ if and only if, for any data sequence $z_1,\dots,z_{n+1}$,

$$\frac{1}{(n+1)!}\sum_{\pi} E(z_{\pi(1)},\dots,z_{\pi(n+1)}) \le 1, \tag{8}$$

π ranging over the permutations of $\{1,\dots,n+1\}$. Let us say that such an E is *admissible* if (8) always holds with "=" in place of "≤". (This agrees with the standard notion of admissibility in statistical decision theory.)

The intuition (which can be formalized easily) behind the operators that we have just introduced 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 $\mathcal{E}^{\mathrm{tX}}$;
- $^{\mathrm{tX}}$ projects $\mathcal{E}$ onto the admissible part of $\mathcal{E}^{\mathrm{tX}}$.

In particular, these operators are idempotent:

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

Despite these operators being projections, we cannot claim that these ways of moving between different function classes are always optimal.

Lemma 2 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:

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

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 [10, 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 we will write $\int f(y) B(\mathrm{d}y \mid z)$ for the integral of a function f on $\mathbf{Y}$ w.r.t. to the measure $A \mapsto B(A \mid z)$. We will show that the efficiency of various predictors (such as the conformal predictor) derived from a 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 Kolmogorov's step

The following statement shows that the efficiency does not suffer much on average when we move from randomness e-prediction to exchangeability e-prediction. It is a counterpart of Corollary 5 in [8] (with a similar proof).

Theorem 3. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel. For each randomness e-predictor E,*

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

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

We can interpret (9) as a statement that E^{X} is almost as efficient as E: the mean ratio of the degree to which E rejects a false label y to the degree to which E^{X} rejects y is bounded by e under any probability measure that may depend on the test example. This will be further discussed after we state our main result, Theorem 8.

Proof of Theorem 3. 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, $G_1(z_1, \ldots, z_{n+1})$ is obtained by randomly (with equal probabilities) choosing an example z_i in the data sequence $z_1, \ldots, z_{n+1}$, replacing its label y_i by a random label $y \sim B(\cdot \mid z_i)$, and finding the expectation of E^{i} on $z_1, \ldots, z_{n+1}$ 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, $i = 1, \ldots, n+1$, 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

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

Combining all these statements, we get

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

By the definition (6), this is equivalent to (9). □

Notice that Theorem 3 does not assume the homoscedasticity of the labels. The simplest informative examples, however, are indeed homoscedastic: in them, for each $z = (x, y) \in \mathbf{Z}$, $B(\cdot \mid z)$ is the distribution of $y + \xi$ for a given random variable ξ. In general, however, the distribution of ξ may depend on the object x.

The following result is a simple inverse to Theorem 3.

Theorem 4. *The constant* e^{-1} *in Theorem 3 cannot be replaced by a larger one.*

Proof. In this proof we follow the example in [8, Sect. B.1] (the example in [8] is informal, and here we formalize it). Without loss of generality we assume $|\mathbf{X}| = 1$ (so that the objects become uninformative and we can omit them from our notation) and $\mathbf{Y} = \{0, 1\}$ (with the discrete σ-algebra). Define a randomness e-variable E by

$$E(y_1, \ldots, y_{n+1}) := \begin{cases} \left(1 - \frac{1}{n+1}\right)^{-n} & \text{if } k = 1 \\ 0 & \text{if not,} \end{cases} \tag{10}$$

where k is the number of 1s in $y_1, \ldots, y_{n+1}$. This is indeed a randomness e-variable, since the maximum probability of $k = 1$ in the Bernoulli model, $(n +$

$1)p(1-p)^n \to \max$, is attained at $p = \frac{1}{n+1}$. The corresponding exchangeability e-variable is

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

Let B just flip the label: $B(\{1-y\} \mid y) = 1$. Suppose Theorem 3 holds with the e^{-1} in (9) 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 5. Whereas the randomness e-variable E defined by (10) is all we need to prove Theorem 4, it is not useful for prediction. A variation on (10) 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 $E(0, \dots, 0, 1) \sim en$, under exchangeability we can only achieve $E^{\mathrm{X}}(0, \dots, 0, 1) = n + 1 \sim n$.

3.2 Training-invariance step

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 $\mathcal{E}^{\mathrm{X}}$.

Theorem 6. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel. For each exchangeability e-predictor E,*

$$G(z_1, \dots, z_n, z_{n+1}) := \int \frac{E(z_1, \dots, z_n, x_{n+1}, y)}{E^{\mathrm{t}}(z_1, \dots, z_n, x_{n+1}, y)} B(\mathrm{d}y \mid z_{n+1}) \qquad (11)$$

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

The interpretation of (11) is similar to that of (9).

Proof of Theorem 6. It suffices to check that the right-hand side of (11) is a test-conditional exchangeability e-variable. We have:

$$\begin{aligned}
&\frac{1}{n!}\sum_{\sigma} G(z_{\sigma(1)},\dots,z_{\sigma(n)},z_{n+1}) \\
&\qquad = \frac{1}{n!}\sum_{\sigma}\int \frac{E(z_{\sigma(1)},\dots,z_{\sigma(n)},x_{n+1},y)}{E^{\mathrm{t}}(z_1,\dots,z_n,x_{n+1},y)} B(\mathrm{d}y \mid z_{n+1}) \\
&\qquad\qquad = \int \frac{E^{\mathrm{t}}(z_1,\dots,z_n,x_{n+1},y)}{E^{\mathrm{t}}(z_1,\dots,z_n,x_{n+1},y)} B(\mathrm{d}y \mid z_{n+1}) \le 1. \quad \square
\end{aligned}$$

3.3 Putting everything together

The following theorem combines Theorems 3 and 6 and establishes a connection between randomness and conformal e-predictors. Remember that the conformal e-predictor $\mathcal{E}^{\mathrm{tX}}$ derived from a randomness e-predictor E is obtained by combining the operators (6) and (7), i.e., as

$$E^{\mathrm{tX}}(z_1,\dots,z_{n+1}) := (n+1)\frac{\sum_{\sigma} E(z_{\sigma(1)},\dots,z_{\sigma(n)},z_{n+1})}{\sum_{\pi} E(z_{\pi(1)},\dots,z_{\pi(n+1)})}, \tag{12}$$

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

Theorem 7. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel. For each randomness e-predictor E,*

$$G(z_1,\dots,z_n,z_{n+1}) := \mathrm{e}^{-1/2}\int \sqrt{\frac{E(z_1,\dots,z_n,x_{n+1},y)}{E^{\mathrm{tX}}(z_1,\dots,z_n,x_{n+1},y)}}\, B(\mathrm{d}y \mid z_{n+1}) \tag{13}$$

is a randomness e-variable.

Theorem 7 is the main result of this paper for e-predictors. Its main weakness 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 $G_1, G_2, G \in \mathcal{E}^{\mathrm{R}}$,

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

$$
\times \sqrt{\int \frac{E^{\mathrm{X}}(z_1,\dots,z_n,x_{n+1},y)}{E^{\mathrm{tX}}(z_1,\dots,z_n,x_{n+1},y)} B(\mathrm{d}y \mid z_{n+1})}
$$
$$
= \sqrt{G_1(z_1,\dots,z_{n+1}) G_2(z_1,\dots,z_{n+1})} \le G(z_1,\dots,z_{n+1})
$$

(the existence of G_1 and G_2 follows from Theorems 3 and 6, respectively). □

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., [11, Propositions 2.1 and 2,2]). This allows us to adapt Theorem 8 to p-predictors.

Theorem 8. *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*

$$
G(z_1,\dots,z_n,z_{n+1}) := (\delta/\mathrm{e})^{1/2} \times \int \sqrt{P(z_1,\dots,z_n,x_{n+1},y)^{\delta-1} P'(z_1,\dots,z_n,x_{n+1},y)}\, B(\mathrm{d}y \mid z_{n+1}) \quad (14)
$$

is a randomness e-variable.

The interpretation of (14) is that $P'(z_1,\dots,z_n,x_{n+1},y)$ 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 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 have

$$
(\delta/\mathrm{e})^{1/2} \sqrt{P(z_1,\dots,z_n,x_{n+1},y)^{\delta-1} P'(z_1,\dots,z_n,x_{n+1},y)} < \frac{1}{\epsilon_1 \epsilon_2}, \quad (15)
$$

where "the vast majority" means that the $B(\cdot \mid z_{n+1})$ measure of the y satisfying (15) is at least $1 - \epsilon_2$. We can rewrite (15) as

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

so that $P'(z_1,\dots,z_n,x_{n+1},y) \to 0$ as $P(z_1,\dots,z_n,x_{n+1},y) \to 0$. This is, of course, true for any Markov kernel B.

Proof of Theorem 8. Fix $\delta \in (0,1)$ and $P \in \mathcal{P}^{\mathrm{R}}$. Set $E := \delta P^{\delta-1}$ and $P' := 1/E^{\mathrm{tX}}$, so that $E \in \mathcal{E}^{\mathrm{R}}$ and $P' \in \mathcal{P}^{\mathrm{tX}}$. According to (13),

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

is a randomness e-variable. □

4 Applications to classification

In this and next sections we will discuss some interesting examples of Markov kernels B as used in Theorems 7 and 8. In this section we discuss the case of classification, $|\mathbf{Y}| < \infty$, which was also discussed earlier in [8].

Let us start from binary classification, $\mathbf{Y} := \{0, 1\}$. In this case, 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 $1 - y$. We can rewrite (13) as

$$G(z_1, \dots, z_n, z_{n+1}) := \mathrm{e}^{-1/2} \sqrt{\frac{E(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1})}{E^{\mathrm{tX}}(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1})}}, \qquad (16)$$

which does not involve any averaging. We can interpret (16) as the conformal e-predictor E^{tX} being almost as efficient as the original randomness e-predictor E, where efficiency is measured by the degree to which we reject the false label $1 - y_{n+1}$. For example, for a small positive constant ϵ, $G \ge 1/\epsilon$ with probability at most ϵ, and so

$$E^{\mathrm{tX}}(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1}) > \mathrm{e}^{-1} \epsilon^2 E(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1})$$

with probability at least $1 - \epsilon$.

In the case of reduction of a randomness p-predictor, we rewrite (14) as

$$\begin{aligned} G(z_1, \dots, z_n, z_{n+1}) &= (\delta/\mathrm{e})^{1/2} \\ &\quad \times \sqrt{P(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1})^{\delta - 1} P'(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1})}. \end{aligned}$$

Therefore,

$$P'(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1}) < \mathrm{e} \delta^{-1} \epsilon^{-2} P(z_1, \dots, z_n, x_{n+1}, 1 - y_{n+1})^{1 - \delta}$$

with probability at least $1 - \epsilon$. The interpretation is similar to that of the e-version.

In the rest of this section, let us only discuss reduction of randomness e-predictors to conformal e-predictors. Reduction of randomness p-predictors to conformal p-predictors is completely analogous; it just uses (14) instead of (13).

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 (13) as

$$G(z_1, \dots, z_n, z_{n+1}) := \frac{\mathrm{e}^{-1/2}}{|\mathbf{Y}| - 1} \sum_{y \in \mathbf{Y} \setminus \{y_{n+1}\}} \sqrt{\frac{E(z_1, \dots, z_n, x_{n+1}, y)}{E^{\mathrm{tX}}(z_1, \dots, z_n, x_{n+1}, y)}}. \qquad (17)$$

The interpretation of (17) is that the conformal e-predictor E^{tX} is almost as efficient as the original randomness e-predictor E on average; as before, efficiency is measured by the degree to which we reject the false labels $y \ne y_{n+1}$. Of course, we can avoid “on average” by making (17) cruder and replacing it by the existence of $G \in \mathcal{E}^{\mathrm{R}}$ satisfying

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

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 $1-\epsilon$,

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

An interesting variation of (17), corresponding to the Markov kernel B for which $B(\cdot \mid (x,y))$ is the uniform probability measure on $\mathbf{Y}$, is

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

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

5 Applications to regression

In this section we set $\mathbf{Y} := \mathbb{R}$; therefore, we consider the problem of regression. In applied regression problems, we are often interested in prediction intervals rather than arbitrary prediction sets. This calls for an investigation of the regularity of the derived conformal predictor, which we will start in this section.

We will be interested only in upper prediction limits, as in [2, Sect. 7.2(i)]. Once we deal with those, we can treat lower prediction limits in the same way, and then a prediction interval can be formed as the intersection of the rays defined by upper and lower prediction limits. (In several respects, this is much more convenient than finding prediction intervals directly, as shown in the context of conformal regression in [1] and discussed in [10, Sect. 2.3.2].) For simplicity, let us concentrate on e-prediction.

Given a randomness e-predictor E, the derived conformal e-predictor E^{tX} is based on repeated averaging: see (12). Therefore, we can expect E^{tX} to be more regular. We start from checking the regularity of E^{tX} in a simple case.

Let us say that a randomness e-predictor E is *monotonic* if $E(z_1,\dots,z_{n+1})$ is increasing in y_{n+1} but decreasing in y_i for each $i \in \{1,\dots,n\}$. (This is analogous to monotonic conformity measures as discussed in [10, Sect. 7.2.3].)

Lemma 9. *If a randomness e-predictor $E \in \mathcal{E}^{\mathrm{R}}$ is monotonic, then its conformal version $E^{\mathrm{tX}} \in \mathcal{E}^{\mathrm{tX}}$, defined by* (12)*, is increasing in y_{n+1}.*

Lemma 9 says that the prediction function output by E^{tX} is increasing. It is clear that this lemma is also applicable to conformal p-predictors, in the following sense: if $P \in \mathcal{P}^{\mathrm{R}}$ is monotonic (decreasing in y_{n+1} and increasing in y_i, $i \neq n+1$), then its conformalized version $1/(\delta P^{\delta-1})^{\mathrm{tX}}$ outputs a decreasing prediction function. Therefore, the conformalized version will output rays as its prediction sets.

Proof of Lemma 9. The denominator of the fraction in (12) is the sum of its numerator, which is increasing in y_{n+1}, and addends that are decreasing in y_{n+1}. It remains to notice that the fraction

$$\frac{f(y)}{f(y)+g(y)} = \frac{1}{1+g(y)/f(y)},$$

where f and g are nonnegative functions that are increasing and decreasing, respectively, is increasing in y. □

The next proposition gives a lower bound on the conformalized version of a particularly simple prediction set output by a monotonic randomness e-predictor. Namely, we consider an "upper prediction ray", a prediction function of the form $f(y) = D1_{\{y \geq b\}}$, where $b \in \mathbb{R}$ is the upper prediction limit and $D > 0$ reflects the confidence in this prediction.

Proposition 10. *Let $B : \mathbf{Z} \hookrightarrow \mathbf{Y}$ be a Markov kernel. Suppose a monotonic randomness e-predictor E, given a training sequence $z_1, \dots, z_n$ and test object x_{n+1}, outputs a set prediction $f(y) := D1_{\{y \geq b\}}$ for the label y_{n+1}, where $D > 0$ and $b \in \mathbb{R}$. Let F be the distribution function of $B(\cdot \mid z_{n+1})$. Then the function $h : \mathbb{R} \to \mathbb{R}$ defined by*

$$h(y) := \frac{D(F(y)-F(b))^2}{\mathrm{e}g^2} 1_{\{y \geq b\}}, \tag{18}$$

where $g := G(z_1, \dots, z_{n+1})$ and G is the randomness e-variable defined in Theorem 7, is a lower bound on the prediction function output by E^{tX}.

The parameter b of the prediction function $D1_{\{y \geq b\}}$ of E, reflecting the precision of the upper prediction limit, should ideally be greater than but close to y_{n+1}, and then we could interpret $b - y_{n+1}$ as the precision. The prediction function of E^{tX} is increasing by Lemma 9. The bound (18) is very weak, which can be seen from $h(\infty) \leq D/(\mathrm{e}g^2)$. However, this is the best that can be derived from Theorem 7: being competitive on average does not mean being competitive at each individual label y.

Proof of Proposition 10. It suffices to consider only $y > b$ in (18). Let A be the value of the prediction function output by E^{tX} at some point $a > b$. Then the right-hand side of (13) is smallest if the prediction function $E^{\mathrm{tX}}(z_1, \dots, z_n, x_{n+1}, \cdot)$ is A at and to the left of a and is ∞ to the right of a. Therefore, it is impossible to have

$$\mathrm{e}^{-1/2} \int_b^a \sqrt{\frac{D}{A}}\, \mathrm{d}F > g,$$

i.e.,

$$A < \frac{D(F(a) - F(b))^2}{\mathrm{e}g^2},$$

which gives the lower bound (18). □

Let us see what the lower bound (18) becomes for specific distributions, e.g., the exponential ones centred at y_{n+1},

$$F(y) := \begin{cases} 1 - \exp(-\lambda(y - y_{n+1})) & \text{if } y > y_{n+1} \\ 0 & \text{otherwise,} \end{cases}$$

where λ is a positive constant. (Remember that F is allowed to depend on z_{n+1}.) The lower bound becomes

$$h(y) = \frac{D(\exp(-\lambda(b - y_{n+1})) - \exp(-\lambda(y - y_{n+1})))^2}{\mathrm{e}g^2}$$

for $y > b$ and $b > y_{n+1}$. In the homoscedastic, or nearly homoscedastic, regular case we could choose the parameter λ close to the typical values of $1/(b - y_{n+1})$. Then $h(2b - y_{n+1})$ would have the order of magnitude at least Dg^{-2} (the geometric interpretation of $2b - y_{n+1}$ is that b is half-way between y_{n+1} and $2b - y_{n+1}$).

6 Conclusion

These are some directions of further research:

- Can we connect any two of the classes $\mathcal{P}^{\mathrm{R}}$, $\mathcal{P}^{\mathrm{X}}$, and $\mathcal{P}^{\mathrm{tX}}$ directly (in the spirit of Theorem 8), without a detour via e-values?
- Our only optimality result, Theorem 4, covers Kolmogorov's step only. It would be ideal to have optimality results related to Theorems 7 and 8.

References

[1] Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. *JMLR: Workshop and Conference Proceedings*, 35:605–622, 2014. COLT 2014.

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

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

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

[5] Denis Infanger and Arno Schmidt-Trucksäss. *P* value functions: An underused method to present research results and to promote quantitative reasoning. *Statistics in Medicine*, 38:4189–4197, 2019.

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

[7] 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.

[8] Vladimir Vovk. Randomness, exchangeability, and conformal prediction. Technical Report arXiv:2501.11689 [cs.LG], arXiv.org e-Print archive, February 2025.

[9] Vladimir Vovk. Conformal e-prediction, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 26, February 2025. A slightly shorter version (not including the end of Appendix B referred to in this paper) has been published as arXiv technical report 2001.05989 [cs.LG], November 2024.

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

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