Conformal e-prediction

Vladimir Vovk

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

On-line Compression Modelling Project (New Series)

Working Paper #26

First posted January 16, 2020. Last revised April 17, 2025.

Project web site:
http://alrw.net

Abstract

This paper discusses a counterpart of conformal prediction for e-values, *conformal e-prediction*. Conformal e-prediction is conceptually simpler and had been developed in the 1990s as a precursor of conformal prediction. When conformal prediction emerged as result of replacing e-values by p-values, it seemed to have important advantages over conformal e-prediction without obvious disadvantages. This paper re-examines relations between conformal prediction and conformal e-prediction systematically from a modern perspective. Conformal e-prediction has advantages of its own, such as the ease of designing conditional conformal e-predictors and the guaranteed validity of cross-conformal e-predictors (whereas for cross-conformal predictors validity is only an empirical fact and can be broken with excessive randomization). Even where conformal prediction has clear advantages, conformal e-prediction can often emulate those advantages, more or less successfully.

Contents

1 Introduction

Conformal prediction is based on the notion of a p-value. At this time p-values are widely discussed and sometimes criticized (see, e.g., [31]), and several alternatives to p-values have been proposed. Perhaps the most popular alternatives are Bayes factors and their non-Bayesian variation, e-values. The terminology of e-values was introduced in [29], and the literature on e-values has been growing quickly; see, e.g., [17, 16, 8].

In fact, e-values were used (under different names) when discussing a precursor of conformal prediction in the 1990s; in this paper we will refer to this precursor as *conformal e-prediction*. One early description of conformal e-prediction is [6]. The paper [25] that first introduced conformal prediction also discusses conformal e-prediction. In this paper, we will occasionally refer to conformal prediction as *conformal p-prediction* in order to emphasize it being based on p-values.

Soon after the publication of [6, 25], conformal e-prediction seems to have disappeared. Perhaps the main reason why it was superseded by conformal prediction was that conformal predictions can be packaged as prediction sets [26, Sect. 2.2], and in this case their property of validity is very easy to state: we just say that the probability of error is at most ϵ at a prespecified significance level ϵ [26, Proposition 2.3]. This was clearly stated only in 2001 [13, Theorem 1], although this statement was implicit in the standard requirement of validity for p-values stated in [25]. The standard requirement of validity for e-values, also stated in [25], does not admit such a simple restatement in terms of probability of error without weakening it drastically; see Appendix B for further details. (While the prediction sets derived from conformal e-prediction can be used to define the property of validity in its strong form, validity becomes a property of the whole family of prediction sets for different significance levels.)

Another reason for conformal e-prediction losing its popularity may have been the finding in 2002 [22, Theorem 1] that, in the on-line mode of prediction, smoothed conformal predictors make errors independently. An important corollary of this stronger property of validity is that small probabilities of errors manifest themselves, with high probability, as a low frequency of errors [26, Corollary 2.5].

The last advantage of conformal prediction that we discuss in this section was found only in 2017 [27], and so it did not contribute to the eclipse of conformal e-prediction in the early 2000s. It was the discovery of conformal predictive distributions, motivated by [18]: in the case of regression, smoothed conformal prediction may produce "conformal predictive distributions", which are automatically well-calibrated.

In this paper we will look systematically at these advantages of conformal prediction except for the last one, which will only be briefly discussed in the concluding section. On one hand, the favourable properties of conformal prediction are often partially satisfied by conformal e-prediction. And on the other hand, conformal e-prediction has several advantages of its own.

For simplicity, in this paper we consider IID (or at least exchangeable) data and concentrate on the problem of pattern recognition (also known as classification). We start in Sect. 2 from the definition of conformal e-prediction and continue in Sect. 3 with discussing its validity. The simplest property of validity (Proposition 2 in Sect. 3) consists in conformal e-predictors producing at each step a valid e-value for the true label: namely, its expectation is at most 1 at each step; the general notion of e-value is introduced right after Proposition 2. Since the expectation is the average over the sample space, we can say that conformal e-predictors are valid in the space domain, or space-wise. A complementary notion of efficiency is efficiency in the time domain, which we consider next.

Proposition 2 does not say anything about the relation between the e-values for the true labels produced at different steps. Is it possible that for some streams of data the average of the e-values produced at different steps tends to 2 while for others it tends to 0.5 with non-zero probability? In Sect. 3 we show that this is impossible in the online prediction protocol, stating both the strong law of large numbers and the law of the iterated logarithm for the e-values produced for the true labels at different steps. This does not fully replace the strong property of independence of errors for conformal prediction, but it can be regarded as a partial replacement. We can see that the e-values produced at different steps are not misleading in the time domain; not only is their expectation at most 1 at each step, their average is at most 1 time-wise in the long run.

The properties of validity established in Sect. 3 are marginal, in the sense that the expectations, or time averages, in them are not conditional on any properties of the observations. For example, if our predictions are for people, in principle we can get very different averages for men and women. In the case of conformal prediction, a simple way of achieving conditional validity is using Mondrian conformal prediction [26, Sect. 4.6]. Mondrian conformal prediction requires a hard partition of observations, such as the partition of people into men and women. Interestingly, conformal e-prediction is much more flexible when trying to achieve conditional validity. Its conditional version does not have to be based on a partition (Sect. 4); e.g., we may require separate validity for men, women, and Europeans. It has been shown recently that this type of conditionality can also be achieved for conformal prediction (see, e.g., [7]), but it is much easier to achieve and appears to be more natural in the case of conformal e-prediction. This can be regarded as an advantage of conformal e-prediction.

What we discuss in Sections 2–4 is "full" conformal e-prediction, and it is computationally inefficient when built on top of many standard prediction algorithms (such as neural networks). Section 5 introduces split conformal e-prediction, which is a simple way to make conformal e-prediction computationally efficient. Similarly to split-conformal prediction (introduced in [15, 14]), split conformal e-prediction can lose in predictive efficiency as compared with full conformal e-prediction.

To prevent loss in computational efficiency without sacrificing predictive efficiency, cross-conformal predictors were introduced in [24]. Cross-conformal predictors are not provably valid [24, Appendix], and this sometimes even shows in experimental results [12]. The limits of violations of validity are given by Rüschendorf's result (see, e.g., [28, Proposition 2]): when merging p-values coming from different folds by taking arithmetic mean (this is essentially what cross-conformal predictors do), the resulting arithmetic mean has to be multiplied by 2 in order to guarantee validity. In the more recent method of jackknife+, introduced in [3] and closely related to cross-conformal prediction, there is a similar factor of 2 [3, Theorem 1], which cannot be removed in general [3, Theorem 2].

In Sect. 6, we introduce a version of cross-conformal prediction based on e-values, which we call *cross-conformal e-prediction*. The situation with cross-conformal e-prediction is very different from cross-conformal prediction, as the arithmetic mean of e-values is always an e-value. This is an obvious fact, and it is shown in [29, Sect. 3] that arithmetic mean is the only useful merging rule. Therefore, cross-conformal e-prediction is always valid. This is a second advantage of conformal e-prediction.

The emphasis of Sections 3–6 is on the validity of conformal e-prediction and its computational efficiency, while Sect. 7 moves on to its predictive efficiency. What are suitable criteria of predictive efficiency? We propose two such criteria in the case of pattern recognition, the "observed log criterion" and the "prior log criterion".

Section 8 concludes and lists some advantages and disadvantages of conformal e-prediction as compared with conformal prediction.

2 Conformal e-predictors

Suppose we are given a training set $z_1, \dots, z_n$ consisting of labelled objects $z_i = (x_i, y_i)$ and our goal is to predict the label of a new object x. In this paper we consider predictors of the following type: for each potential label y for x we would like to have a number $f(z_1, \dots, z_n, x, y)$ reflecting the plausibility of y being the true label of x. An example is conformal transducers [26, Sect. 2.7], which, in the terminology of this paper, may be called *conformal p-predictors*. The output

$$y \mapsto f(z_1, \dots, z_n, x, y)$$

of a conformal p-predictor is the full conformal prediction for the label of x; e.g., it determines the prediction set at each significance level. We will sometimes write $f(z_1, \dots, z_n, z)$, where $z := (x, y)$, instead of $f(z_1, \dots, z_n, x, y)$.

We will use the notation $\mathbf{X}$ for the object space and $\mathbf{Y}$ for the label space (both assumed non-empty). These are measurable spaces from which the objects and labels, respectively, are drawn. Full observations $z = (x, y)$ are drawn from the observation space $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$. For any non-empty set X, X^+ will be the set $\cup_{n=1}^{\infty} X^n$ of all non-empty finite sequences of elements of X.

A *nonconformity e-measure* is a measurable function $A : \mathbf{Z}^+ \to [0, \infty)^+$ that maps any finite sequence $(z_1, \dots, z_m)$, $m \in \{1, 2, \dots\}$, to a finite sequence

$(\alpha_1, \dots, \alpha_m)$ of the same length consisting of nonnegative numbers (*nonconformity scores*) with average at most 1,

$$\frac{1}{m}\sum_{i=1}^{m} \alpha_i \le 1, \tag{1}$$

that satisfies the following property of equivariance: for any $m \in \{2, 3, \dots\}$, any permutation π of $\{1, \dots, m\}$, any $(z_1, \dots, z_m) \in \mathbf{Z}^m$, and any $(\alpha_1, \dots, \alpha_m) \in [0, \infty)^m$,

$$(\alpha_1, \dots, \alpha_m) = A(z_1, \dots, z_m) \Longrightarrow (\alpha_{\pi(1)}, \dots, \alpha_{\pi(m)}) = A(z_{\pi(1)}, \dots, z_{\pi(m)}).$$

Very roughly, this property means that each nonconformity score α_i (supposed to measure the strangeness of z_i as compared with the other observations in the sequence) does not depend on the position of z_i, or of the other observations, in the sequence. The *conformal e-predictor* f corresponding to such A is defined by

$$f(z_1, \dots, z_n, x, y) := \alpha_{n+1}, \text{ where } (\alpha_1, \dots, \alpha_n, \alpha_{n+1}) := A(z_1, \dots, z_n, (x, y)), \tag{2}$$

so that $f : \mathbf{Z}^+ \to [0, \infty)$. A *conformal e-predictor* is a function that can be obtained from a nonconformity e-measure in this way.

When given a training set $z_1, \dots, z_n$ and a test object x, the full prediction for x according to a conformal e-predictor f is the family of *conformal e-values*

$$(f(z_1, \dots, z_n, x, y) \mid y \in \mathbf{Y}) \,. \tag{3}$$

We can regard the family (3) of e-values for each potential label y as a soft set predictor. By thresholding f at some level, we can get a (hard) set predictor (as in (41) below); we include in the prediction set for the label of x the potential labels y for which the conformal e-value is less than a chosen level. But there is no need to choose the level in advance, and we interpret the conformal e-value $f(z_1, \dots, z_n, x, y)$ of y as the degree to which y is excluded from the soft prediction set. We want our conformal prediction to be valid and efficient, where validity means that we do not want the true label to be excluded (i.e., to have a large e-value) while efficiency means that we want to exclude all other labels.

The full prediction (3) for the label of x can be summarized as, e.g., the *point prediction*

$$\hat{y} \in \arg\min_y f(z_1, \dots, z_n, x, y)$$

(assuming the min is attained at a single label), the *e-confidence*

$$\min_{y \ne \hat{y}} f(z_1, \dots, z_n, x, y),$$

and the *e-credibility* $f(z_1, \dots, z_n, x, \hat{y})$. (See [26, Sect. 3.5.1] for their p-counterparts.) We can make a confident point prediction when the e-confidence is large while the e-credibility is not.

Let us say that a nonconformity e-measure and the corresponding conformal e-predictor are *admissible* if we always have "=" in place of "$\le$" in the definition (1). If a conformal e-predictor is not admissible, we can make its predictions more confident without sacrificing their validity. Therefore, we will usually concentrate on admissible nonconformity e-measures and admissible conformal e-predictors.

A *nonnegative nonconformity measure* $A : \mathbf{Z}^+ \to [0,\infty)^+$ is defined as a nonconformity e-measure except that the condition (1) is omitted. Given a nonnegative nonconformity measure A, we can always define the corresponding admissible nonconformity e-measure A' by normalizing A:

$$A'(z_1,\dots,z_m) := \frac{m}{\sum_{i=1}^m \alpha_i}(\alpha_1,\dots,\alpha_m), \tag{4}$$

where $(\alpha_1,\dots,\alpha_m) := A(z_1,\dots,z_m)$; if $A(z_1,\dots,z_m) = (0,\dots,0)$, we set $A'(z_1,\dots,z_m) := (1,\dots,1)$ in order to ensure that A' is admissible. We will say that the corresponding conformal e-predictor *is based on* A.

A further generalization of nonnegative nonconformity measures, *nonconformity measures*, is obtained by dropping the condition of nonnegativity; these are functions $A : \mathbf{Z}^+ \to \mathbb{R}^+$. They are used, explicitly or implicitly, in conformal p-prediction. See, e.g., [1, Sect. 1.3] (and Remark 1 below).

The conformal e-predictor proposed in [6] for binary pattern recognition problems (with $\mathbf{Y} = \{-1, 1\}$) is based on support vector machines (SVM); let us fix all relevant parameters, such as the kernel. It is defined as

$$f(z_1,\dots,z_{n+1}) := \begin{cases} (n+1)/\left|\mathrm{SV}\right| & \text{if } n+1 \in \mathrm{SV} \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where SV is the set of indices of support vectors: $i \in \mathrm{SV}$ if and only if z_i, $i \in \{1,\dots,n+1\}$, is a support vector for the SVM constructed from $z_1,\dots,z_{n+1}$ as training set. It is based on the indicator function of being a support vector; indeed, the right-hand side of (4) then becomes

$$\left(\frac{(n+1)1_{\{i\in\mathrm{SV}\}}}{\sum_{i=1}^{n+1} 1_{\{i\in\mathrm{SV}\}}}\right)_{i=1}^{n+1},$$

which agrees with (5). When given a training set $z_1,\dots,z_n$ and a new object x, this conformal e-predictor goes through all potential labels y for x and for each constructs an SVM and outputs $f(z_1,\dots,z_n,x,y)$. It makes it computationally inefficient.

Remark 1. In conformal prediction, several natural ways of defining nonconformity measures have been discussed and widely used in literature. The definition at the beginning of this section parallels the definition of nonconformity measures in [1, Sect. 1.3]; under these definitions, nonconformity measures map sequences of observations to sequences of nonconformity scores. Under other definitions, nonconformity measures may map, e.g., a bag of observations and another observation z to one nonconformity score (that for z), and there are several varieties of such definitions; see, e.g., [26, Sect. 2.9.3]. Each of these varieties could have been adapted to conformal e-prediction.

3 Validity of conformal e-predictors

The following obvious proposition asserts the validity of conformal e-predictors. Let us write $Z_1, Z_2, \dots$ for the random elements whose realizations are the observed data $z_1, z_2, \dots$; more generally, (X, Y) or Z are random elements with values in the observation space $\mathbf{Z}$. As usual, a finite sequence of random elements is *exchangeable* if its joint distribution does not change if it is permuted (and an infinite sequence is exchangeable if its joint distribution does not change if its first n elements are permuted, for any n). The difference between exchangeability and being IID (independent and identically distributed) can be substantial for finite data sequences (but for infinite data sequences the difference between the two assumptions disappears, provided $\mathbf{Z}$ is a standard Borel space, according to de Finetti's theorem [26, Sect. A.5.1]).

Proposition 2. *For any conformal e-predictor f and any n, if $Z_1, \dots, Z_n, (X, Y)$ are IID (or exchangeable),*

$$\mathbb{E} f(Z_1, \dots, Z_n, X, Y) \le 1 \tag{6}$$

(with "$=$" in place of "$\le$" if f is admissible).

Proof. It follows from the definition of conformal e-predictors that

$$\mathbb{E}\left(f(Z_1, \dots, Z_n, X, Y) \mid \wr Z_1, \dots, Z_n, (X, Y) \int\right) \le 1,$$

and it remains to average over the multisets $\wr Z_1, \dots, Z_n, (X, Y) \int$. □

The property of validity given in Proposition 2 says that conformal e-predictors output valid e-values for the true labels. Formally, an *e-variable* is a random variable E satisfying $\mathbb{E}(E) \le 1$ under the data-generating distribution (and the values it takes are *e-values*). We do not expect the e-values for the true labels to be large because, by Markov's inequality, $\mathbb{P}(E \ge C) \le 1/C$ for any constant $C > 1$.

Proposition 2 involves a "space average", i.e., an average over the sample space. The analogous property of validity for conformal prediction can be stated in terms of error probabilities for set predictions (see, e.g., [26, Proposition 2.1]), which is particularly intuitive. This is the first advantage of conformal prediction pointed out in Sect. 1. It disappears when we move to conformal e-prediction: validity has to be defined in a more complicated way in order to avoid making this property much weaker (however, it has been argued [17] that e-values are more intuitive than p-values).

Another advantage of conformal prediction is that, in the online mode of prediction (to be defined shortly), conformal predictors (in their smoothed version) make errors at different steps independently (see, e.g., [26, Theorem 11.1]).

Without independence, it is possible, even when the probability of error at each step is ϵ, for the long-term relative frequency of errors over consecutive steps to be either 0 or 1 (1 with probability ϵ). The independence of errors forces the long-term frequency of errors to be ϵ almost surely (which is also asserted in [26, Proposition 2.1]). We can say that independence implies ergodicity, i.e., the almost sure coincidence of time and space averages.

The independence of conformal p-values in the online mode of prediction is the strongest property of validity in conformal prediction as presented in [26]. In conformal e-prediction, independence is lost, but ergodicity still holds, at least for bounded conformal e-predictors.

To give an example demonstrating the loss of independence in conformal e-prediction, we first define the online prediction protocol. In the online protocol for conformal e-prediction, we observe an object x_1, apply the conformal e-predictor to compute the conformal e-values $f(x_1, y)$ for all possible labels $y \in \mathbf{Y}$, observe the true label y_1, record the e-value $e_1 := f(x_1, y_1)$ for it, observe another object x_2, apply the conformal e-predictor to compute the e-values $f(x_1, y_1, x_2, y)$ for all possible labels $y \in \mathbf{Y}$, observe the true label y_2, record the e-value $e_2 := f(x_1, y_1, x_2, y_2)$ for it, etc. In the case of conformal prediction, we can instead record conformal p-values p_n (or record whether a mistake is made at a given significance level). While the conformal p-values are independent (for smoothed conformal predictors), the following example shows that the conformal e-values are not independent already in a toy binary situation (perhaps this is the simplest non-trivial example of conformal e-prediction, albeit it is not particularly interesting per se). We write E_n for the e-value e_n regarded as a random variable (an e-variable).

Example 3. Let the random observations $Z_1, Z_2, \dots$ correspond to tossing a fair coin: $Z_n \in \{0, 1\}$ and $Z_n = 1$ with probability $1/2$ independently. Let A be the identity nonconformity measure,

$$A(z_1, \dots, z_m) := (z_1, \dots, z_m),$$

and consider the conformal e-predictor based on A (via (4)). Then the e-variables that it outputs satisfy

$$E_{n+1} = \begin{cases} 0 & \text{with probability } 1/2 \\ \frac{n+1}{k+1} & \text{with probability } 1/2 \end{cases} \qquad (7)$$

given $Z_1, \dots, Z_n$, where $k := \sum_{i=1}^n Z_i$. This assumes $(Z_1, \dots, Z_n) \neq 0$; if $(Z_1, \dots, Z_n) = (0, \dots, 0)$, the 0 in (7) should be replaced by 1. Since $E_1, \dots, E_n$ uniquely determine $Z_1, \dots, Z_n$ unless $(E_1, \dots, E_n) = (1, \dots, 1)$, we still have (7) given $E_1, \dots, E_n$ such that $(E_1, \dots, E_n) \neq (1, \dots, 1)$. Therefore,

$$\mathbb{E}(E_{n+1} \mid E_1, \dots, E_n) = \frac{n+1}{2(k+1)}$$

provided $(E_1, \dots, E_n) \neq (1, \dots, 1)$, and the expression on the right-hand side may well be different from 1 (albeit will typically be close to 1). This shows that the e-values are not independent: their conditional expectations may be different from their marginal expectations.

Nevertheless, various properties of ergodicity still hold. For example, the following proposition asserts a simple property of ergodicity for the conformal e-values $e_n := f(x_1, y_1, \ldots, x_n, y_n)$, namely their asymptotic online validity.

Proposition 4. *Suppose the observations (X_n, Y_n), $n = 1, 2, \ldots$, are IID and e-variables E_n are produced by a bounded conformal e-predictor f. Then, in the online prediction protocol,*

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E_n \le 1 \quad a.s. \tag{8}$$

We can replace the "$\le$" by "$=$" if f is admissible.

Proposition 4 shows that the long-term time average of the e-values for the true labels is bounded above by 1 almost surely. In this sense they are time-wise e-values.

Proof of Proposition 4. We follow the proof of [23, Lemma 14] (given as Lemma 3.15 in the first edition of [26]) and use the terminology of [20, Chap. 7].

Let $\mathcal{F}_n$ be the σ-algebra generated by the multiset $\wr Z_1, \ldots, Z_{n-1} \int$ and the observations $Z_n, Z_{n+1}, \ldots$. (These σ-algebras form the *exchangeable filtration* [16, Sect. 5.6].) For each time horizon $N \in \{2, 3, \ldots\}$, the stochastic sequence $(E_n - 1, \mathcal{F}_n)$, $n = N, \ldots, 1$, is a bounded supermartingale difference. By Hoeffding's inequality (see, e.g., [26, Sect. A.6.3]), for any $N \in \{2, 3, \ldots\}$,

$$\mathbb{P}\left\{\frac{1}{N}\sum_{n=1}^{N}(E_n - 1) \ge \epsilon\right\} \le e^{-2\epsilon^2 N/C^2}, \tag{9}$$

where C is an upper bound on the conformal e-predictor f. By the Borel–Cantelli lemma [19, Sect. 2.10, part (a) of the lemma], the internal inequality in (9) holds only for finitely many N for a fixed $\epsilon > 0$. This implies (8).

If f is admissible, we can also apply the same argument to $1 - E_n$ in place of $E_n - 1$. □

The equation (9) in the proof of Proposition 4 can also be interpreted directly, and its advantage is that it is non-asymptotic. It implies that, for any time horizon $N \ge 2$,

$$\mathbb{P}\left\{\frac{1}{N}\sum_{n=1}^{N} E_n \ge 1 + \epsilon\right\} \le e^{-2\epsilon^2 N/C^2}, \tag{10}$$

which is another expression for the e_n being time-wise e-values: their average is approximately bounded above by 1 with high probability (for small ϵ and large N).

Even if we do not assume that the conformal e-predictor is bounded, we can still claim, e.g., that

$$\mathbb{P}\left\{\frac{1}{N}\sum_{n=1}^{N}(E_n \wedge N^{1/3}) \ge 1+\epsilon\right\} \le e^{-2\epsilon^2 N^{1/3}}. \tag{11}$$

The last inequality says that $E_n \wedge N^{1/3}$ are approximate time-wise e-values (assuming that ϵ is small and $N \gg \epsilon^{-6}$), and so E_n are approximate time-wise e-values if we regard an e-value of $N^{1/3}$ as large enough for the difference between $N^{1/3}$ and larger e-values to be considered unimportant (such as an e-value of 100 on Jeffreys's scale [10, Appendix B]). (To derive (11) from (10), just set $C := N^{1/3}$.)

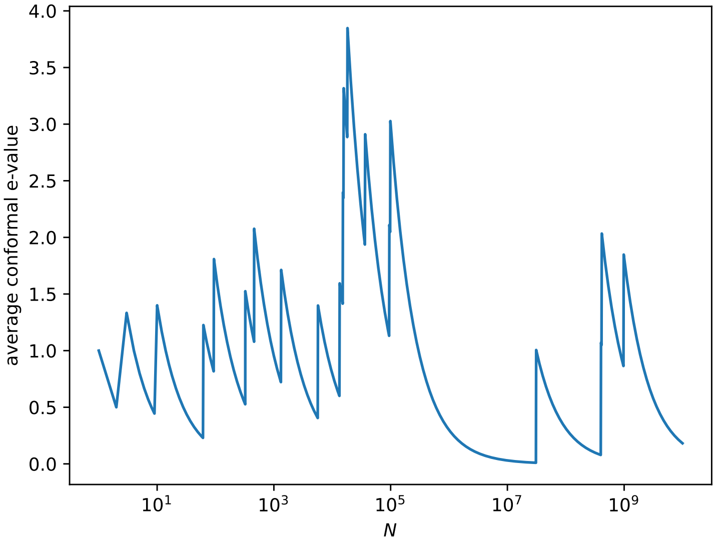

Figure 1: Illustration for Remark 5

Remark 5. Let us check that we cannot simply drop the requirement that E be bounded in Proposition 4. Suppose that, for each n,

$$E_n := \begin{cases} n & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } \frac{n-1}{n} \end{cases}$$

independently of $E_{n+1}, E_{n+2}, \dots$. (For example, $|\mathbf{X}| = 1$, $\mathbf{Y} = \mathbb{R}$, each observation is generated from the same continuous probability measure independently, and the nonconformity e-measure is n at the largest y_n, assuming it is unique, and 0 elsewhere.) Then the distribution of $E_1, \dots, E_N$ is the product distribution

$$\prod_{n=1}^{N} \left(\frac{n-1}{n} \delta_0 + \frac{1}{n} \delta_n \right) \tag{12}$$

(where δ_a is the probability measure concentrated on $\{a\}$), since $E_N, \dots, E_1$ are generated independently. This gives us a product distribution for $E_1, E_2, \dots$, namely (12) with N replaced by ∞. For any $b \in \{1, 2, \dots\}$, by Le Cam's version of Poisson's theorem [19, Sect. 3.12], the probability of the event that b of the E_n for $n \in [2^k, 2^{k+1})$ will satisfy $E_n = n$ tends to $e^{-\lambda}\lambda^b/b!$ as $k \to \infty$, where $\lambda := \ln 2$; therefore, this event will happen infinitely often (for any b). Therefore, (8) will be violated almost surely; namely, the limsup in it will be ∞ almost surely. Figure 1 illustrates this by plotting the average conformal e-value $\frac{1}{N}\sum_{n=1}^{N} e_n$ vs N for a wide range of N, $N \in \{1, \dots, 10^{10}\}$ (with N given on a logarithmic scale).

In Appendix A we will check that Proposition 4 can be strengthened to the following statement of the iterated-logarithm type.

Proposition 6. *Suppose the observations (X_n, Y_n), $n = 1, 2, \dots$, are IID. Then, in the online prediction protocol,*

$$\limsup_{N \to \infty} \sqrt{\frac{N}{\ln \ln N}} \left(\frac{1}{N} \sum_{n=1}^{N} E_n - 1 \right) \le 2^{-1/2} C \quad \textit{a.s.},$$

where C is an upper bound on the conformal e-predictor producing $E_1, E_2, \dots$.

In conclusion of this section, let me state what can be considered to be the main property of validity for conformal e-prediction (it implies the other properties listed here, although by itself it is not particularly intuitive): $\mathbb{E}(E_n \mid \mathcal{F}_{n+1}) \le 1$, where $\mathcal{F}$ is the exchangeable filtration and E_n is a conformal e-variable. It was introduced in the proof of Proposition 4 but is worth stating separately.

4 Conditional conformal e-predictors

Conformal prediction and conformal e-prediction satisfy the marginal property of validity that we have discussed so far. In this section we will discuss stronger properties of validity, including what we called *object-conditional* and *label-conditional* validity in [26, Sect. 4.6]. Suppose, e.g., that we have an algorithm for diagnosing Covid. In object-conditional validity, we divide the objects into separate categories (such as men and women) and require validity inside each of the categories, not just on average. In label-conditional validity, we require validity for each possible label, in this case separate validity for people who have Covid and those who do not. More generally, we can talk about *observation-conditional validity*, where the categories for which we require separate validity are defined in terms of both objects and labels.

The most standard approach to conditional conformal prediction is Mondrian conformal prediction (see, e.g., [26, Sections 4.6.7–4.6.8]), where the observation space $\mathbf{Z}$ is split into a family, often finite, of disjoint categories, and conformal p-values are computed for each category separately. For a relatively narrow group of conformal predictors, object-conditional validity has been established for overlapping categories by Jung et al. [11]. This has been generalized by Gibbs et al. [7] to what may be called "fuzzy categories": a category becomes a nonnegative function on the object space $\mathbf{X}$, with categories in the sense of subsets of $\mathbf{X}$ corresponding to the indicator functions on $\mathbf{X}$. In this section we will see that the steps of moving from disjoint to overlapping and then to "fuzzy" categories become very simple and natural in conformal e-prediction.

Formally, we have a finite set, called *taxonomy*, $\mathcal{K}$ of functions $\kappa : \mathbf{Z} \to [0, \infty)$, and we are only interested in nonconformity e-measures A such that, for all $m \in \{1, 2, \dots\}$ and all $z_1, \dots, z_m \in \mathbf{Z}$,

$$\frac{\sum_{i=1}^{m} \alpha_i \kappa(z_i)}{\sum_{i=1}^{m} \kappa(z_i)} \le 1, \quad \forall \kappa \in \mathcal{K}, \tag{13}$$

where $(\alpha_1, \dots, \alpha_m) := A(z_1, \dots, z_m)$, and $0/0$ is interpreted as 1 when it occurs on the left-hand side. We will say that such an A and the corresponding conformal e-predictor are *$\mathcal{K}$-conditional*. An important special case, a *groupwise taxonomy*, is where all $\kappa \in \mathcal{K}$ take values in $\{0, 1\}$. In this case we regard $\{z \in \mathbf{Z} : \kappa(z) = 1\}$ as the categories. In general, we can still regard κ as fuzzy categories.

The requirement (13) formalizes the conditional validity of the conformal e-predictor. But we do not have efficiency inside a category κ if (13) holds as "$<$" rather than "$=$": if we are interested in this category only, we can improve confidence of predictions without violating validity for this category (but perhaps violating validity for other categories). Let us say that a conformal e-predictor is *$\mathcal{K}$-exact* if it is based on a nonconformity e-measure (also called *$\mathcal{K}$-exact*) satisfying (13) with "$\le$" replaced by "$=$" for all m and $z_1, \dots, z_m$. Being exact is stronger than what is usually called admissible in decision theory, but exact conformal e-predictors are ideal in the sense of achieving validity without being strictly dominated inside each category.

The allowed nonconformity vectors $(\alpha_1, \dots, \alpha_m)$ for a $\mathcal{K}$-exact nonconformity e-measure form the intersection of an affine space and the simplex

$$\left\{(\alpha_1, \dots, \alpha_m) \in [0, 1]^m : \alpha_1 + \dots + \alpha_m = m\right\},$$

and its dimension is at least $m - 1 - |\mathcal{K}|$. This affine space can be defined as the shift by the vector $(1, \dots, 1) \in \mathbb{R}^m$ of the orthogonal complement of the vectors

$$\left(\kappa(z_i)\right)_{i=1}^{m}, \quad \kappa \in \mathcal{K}. \tag{14}$$

(Equivalently, we can define the allowed nonconformity vectors $(\alpha_1, \dots, \alpha_m)$ as the intersection of the nonnegative orthant in $\mathbb{R}^m$ and the orthogonal complement of the vectors (14) extended by adding $(1, \dots, 1) \in \mathbb{R}^m$ and shifted by $(1, \dots, 1) \in \mathbb{R}^m$.)

The geometric picture given in the previous paragraph makes the design of $\mathcal{K}$-exact conditional predictors very easy; it is just a matter of picking a point with nonnegative coordinates in a simply described affine space. Even without the requirement of being $\mathcal{K}$-exact, it is a matter of picking a point in a simply described polytope.

Validity results are easy to state for group-wise taxonomies. (Remember that group-wise taxonomies allow overlapping categories, and so cover many more applications as compared with Mondrian conformal prediction.) First let us state the result in the space domain generalizing Proposition 2.

Proposition 7. *Let $\mathcal{K}$ be a group-wise taxonomy and f be a $\mathcal{K}$-exact conformal e-predictor. For any n and any $\kappa \in \mathcal{K}$, if $Z_1, \dots, Z_n, Z$ are IID (or exchangeable),*

$$\mathbb{E}\left(f(Z_1, \dots, Z_n, Z) \mid \kappa(Z) = 1\right) = 1 \quad \textit{a.s.} \tag{15}$$

Proof. Let us fix $\kappa \in \mathcal{K}$ and check that (15) is true even conditionally on the σ-algebra $\mathcal{F}$ generated by $\kappa(Z_1), \dots, \kappa(Z_n), \kappa(Z)$, by all Z_i, $i = 1, \dots, n$, with $\kappa(Z_i) = 0$, and by the multiset B consisting of Z and all Z_i, $i = 1, \dots, n$, with $\kappa(Z_i) = 1$. Since, conditionally on $\mathcal{F}$ and inside the event $\kappa(Z) = 1$, all orderings of B are equiprobable (almost surely), the conditional expectation of $f(Z_1, \dots, Z_n, Z)$ is 1 given $\mathcal{F}$ and $\kappa(Z) = 1$, which implies (15). □

Now we state a validity result in the time domain.

Proposition 8. *Let $\mathcal{K}$ be a group-wise taxonomy. Suppose the observations Z_n, $n = 1, 2, \dots$, are IID. Then, in the online prediction protocol,*

$$\sum_{n=1}^{\infty} 1_{\{\kappa(Z_n)=1\}} = \infty \Longrightarrow \lim_{N \to \infty} \frac{\sum_{n=1}^{N} E_n 1_{\{\kappa(Z_n)=1\}}}{\sum_{n=1}^{N} 1_{\{\kappa(Z_n)=1\}}} = 1 \quad \textit{a.s.} \tag{16}$$

for a bounded $\mathcal{K}$-exact conformal e-predictor producing conformal e-variables E_n.

The event $A \Rightarrow B$ in (16) is defined, as usual, as the union of B and the complement of A; therefore, $A \Rightarrow B$ holds almost surely if the event $A \setminus B$ is null.

Proof of Proposition 8. By the strong law of large numbers, the antecedent of (16) holds with probability 0 or 1, depending on whether $\mathbb{P}(\kappa(Z) = 1) = 0$ or $\mathbb{P}(\kappa(Z) = 1) > 0$, where Z is any of the Z_n. In the case $\mathbb{P}(\kappa(Z) = 1) = 0$, (16) holds vacuously, so let us assume $\mathbb{P}(\kappa(Z) = 1) > 0$. Our goal is to prove that the consequent of (16) holds almost surely.

By the strong law of large numbers,

$$\sum_{n=1}^{N} 1_{\{\kappa(Z_n)=1\}} \sim \mathbb{P}(\kappa(Z) = 1)N \quad \text{a.s.}, \tag{17}$$

as $N \to \infty$. As in the proof of Proposition 4, we can obtain

$$\sum_{n=1}^{N} E_n 1_{\{\kappa(Z_n)=1\}} \sim \mathbb{P}(\kappa(Z)=1)N \quad \text{a.s.} \tag{18}$$

Combining (17) and (18) gives the consequent of (16) holding almost surely. □

Remark 9. The proofs of Propositions 7 and 8 show that their assumptions (exchangeability or being IID) can be weakened, similarly to the case of Mondrian conformal prediction [26, Sect. 11.3.6]. On the other hand, the conclusion of Proposition 8 can be strengthened to give an iterated-logarithm result along the lines of Proposition 6.

5 Split conformal e-predictors

The versions of conformal e-predictors discussed so far are computationally feasible for a large training set only for a narrow class of nonconformity e-measures. The ideas of split conformal e-predictors, discussed in this section, and cross-conformal e-predictors, discussed in the next one, make it possible to extend greatly the practical applicability of conformal e-prediction.

Let us fix a measurable space Σ (a *summary space*). A Σ-*valued split nonconformity measure* is a measurable function $A : \mathbf{Z}^+ \to \Sigma$. Intuitively, $A(z_1, \dots, z_m, z)$ encodes how well z conforms to $z_1, \dots, z_m$. A *normalizing transformation* $N : \Sigma^+ \to [0, \infty)^+$ is an equivariant measurable function that maps every non-empty finite sequence $(\sigma_1, \dots, \sigma_m)$ of elements of Σ to a finite sequence $(\alpha_1, \dots, \alpha_m)$ of the same length of nonnegative numbers whose average is at most 1 (i.e., satisfying (1)). It is *admissible* if (1) holds with "=".

To apply split conformal e-prediction to a training set $z_1, \dots, z_n$, we split it into two parts, the *training set proper* $z_1, \dots, z_{n-c}$ and the *calibration set* $z_{n-c+1}, \dots, z_n$. For a new object x and a potential label y for it, we set

$$f(z_1, \dots, z_n, x, y) := \alpha^y \tag{19}$$

where α^y is defined using the following steps:

$$\begin{aligned} \sigma_i &:= A(z_1, \dots, z_{n-c}, z_{n-c+i}), \quad i = 1, \dots, c, \\ \sigma^y &:= A(z_1, \dots, z_{n-c}, (x, y)), \\ (\alpha_1^y, \dots, \alpha_c^y, \alpha^y) &:= N(\sigma_1, \dots, \sigma_c, \sigma^y). \end{aligned}$$

For many choices of A and N, the split conformal e-predictor (19) will be computationally efficient; this is the case when:

1. Processing the training set proper only once, we can find an easily computable rule transforming z into $A(z_1, \dots, z_{n-c}, z)$.

2. The normalizing transformation N is easily computable.

To give examples of such easily computable A and N, suppose we have chosen a suitable learning architecture, such as neural networks, and a way of training it. In the case of pattern recognition, a trained neural network implements a function $F : \mathbf{X} \to \mathbf{P}(\mathbf{Y})$, where $\mathbf{P}(\mathbf{Y})$ is the set of all probability measures on $\mathbf{Y}$, assumed finite and equipped with the discrete σ-algebra. Given a test object $x \in \mathbf{X}$, this neural network outputs a probability forecast $F(x) \in \mathbf{P}(\mathbf{Y})$ for its label: the true label of x is y with probability $F(x)(\{y\})$. An example of an easily computable (at the prediction stage) rule A is

$$A(z_1, \dots, z_{n-c}, (x, y)) := \frac{1}{F_{z_1,\dots,z_{n-c}}(x)(\{y\})}, \tag{20}$$

where $F_{z_1,\dots,z_{n-c}} : \mathbf{X} \to \mathbf{P}(\mathbf{Y})$ is the neural network found from $z_1, \dots, z_{n-c}$ as training set. (Training might take a long time, but applying the rule to a new object x is typically quick.) An example of an easily computable normalizing transformation is

$$(\sigma_1, \dots, \sigma_m) \mapsto \frac{m}{\sum_{i=1}^m \sigma_i}(\sigma_1, \dots, \sigma_m)$$

(cf. (4)), where the summary space is supposed to be $\Sigma \subseteq [0, \infty)$.

Proposition 2, our statement of validity, continues to hold for split conformal e-predictors.

Proposition 10. *For any split conformal e-predictor f and any n, if $Z_1, \dots, Z_n, (X, Y)$ are exchangeable, we have* (6) *(with "$=$" if f is admissible).*

Proof. It suffices to notice that (6) holds even conditionally on knowing the observations $Z_1, \dots, Z_{n-c}$ and the multiset $\wr Z_{n-c+1}, \dots, Z_n, (X, Y) \int$ since all orderings of $\wr Z_{n-c+1}, \dots, Z_n, (X, Y) \int$ are equiprobable almost surely. □

6 Cross-conformal e-predictors

Split conformal e-predictors are often computationally efficient, but their predictive efficiency (to be discussed in detail in the next section) may suffer as compared with "full" conformal e-predictors discussed in Sections 2–4, since the latter may be said to use the full training set both as training set proper and as calibration set. The idea behind cross-conformal e-prediction is to combine several split conformal predictors in order to achieve better predictive efficiency.

A Σ-valued split nonconformity measure A is a *Σ-valued cross-nonconformity measure* if $A(z_1, \dots, z_m, z)$ does not depend on the order of its first m arguments. Given such an A and a normalizing transformation N, the corresponding *cross-conformal e-predictor* is defined as follows. The training sequence $z_1, \dots, z_n$ is randomly split into K non-empty multisets (*folds*) z_{S_k}, $k = 1, \dots, K$, of equal (or as equal as possible) sizes $|S_k|$, where $K \in \{2, 3, \dots\}$ is a parameter of the algorithm, $(S_1, \dots, S_K)$ is a partition of the index set $\{1, \dots, n\}$, and z_{S_k}

consists of all z_i, $i \in S_k$. For each $k \in \{1,\dots,K\}$ and each potential label $y \in \mathbf{Y}$ of the new object x, find the output α_k of the split conformal e-predictor (based on A and N) on the new object x and its postulated label y with $z_{S_{-k}}$ as training set proper and z_{S_k} as calibration set, where $S_{-k} := \cup_{j \ne k} S_j = \{1,\dots,n\} \setminus S_k$ is the complement to S_k (and so $z_{S_{-k}}$ is the complement to the fold z_{S_k}). The corresponding cross-conformal e-predictor is defined by

$$f(z_1,\dots,z_n,x,y) := \frac{1}{K} \sum_{k=1}^{K} \alpha_k.$$

(A slight modification, still provably valid, of this definition is where the arithmetic mean is replaced by the weighted mean with the weights proportional to the sizes $|S_k|$ of the folds.)

Proposition 2 still holds for cross-conformal e-predictors.

Proposition 11. *For any cross-conformal e-predictor f and any n, if $Z_1,\dots,Z_n,(X,Y)$ are exchangeable, we have* (6) *(with "=" if f is admissible).*

Proof. This follows from Proposition 10 and the arithmetic mean of e-variables being an e-variable (this is obvious and discussed in detail in [29, Sect. 3]). □

Remark 12. To compare informally the outputs of cross-conformal predictors [26, Sect. 4.4] and cross-conformal e-predictors, we can use the rough transformation discussed in [29, Remark 2.3]: a p-value of p roughly corresponds to an e-value of $1/p$. Under this transformation, the arithmetic average of e-values corresponds to the harmonic average of p-values, and the harmonic average is always less than or equal to the arithmetic average [9, Theorem 16]. This suggests that cross-conformal e-prediction produces better results than cross-conformal prediction does. In the opposite direction, the arithmetic average of p-values corresponds to the harmonic average of e-values, which again suggests that cross-conformal e-prediction produces better results.

Remark 13. It is easy to combine the ideas of this section and Sect. 4 to design conditional cross-conformal e-predictors (or to combine Sections 5 and 4), but we stick to the simplest cases.

Remark 14. Proposition 4 continues to hold for cross-conformal e-predictors and, therefore, it gives its time-wise property of validity in the online mode. However, the online mode entails a massive loss of their computational efficiency. Intuitively, using cross-conformal e-prediction in the online mode defeats the purpose of cross-conformal prediction: once we process $x_1,y_1,\dots,x_n,y_n$, we would like to apply the rule that we have found (see item 1 on p. 13) to a large number of new objects rather than just one. There are more complicated settings of "weak teachers" (along the lines of [26, Sect. 3.3]) that combine cross-conformal e-prediction with time-wise validity in useful ways, but we will not discuss them further in this paper.

7 Predictive efficiency of conformal e-predictors

So far we have concentrated on the validity of conformal e-predictors. A valid e-predictor is not allowed to output consistently large e-values for the true labels; namely, the expectation of the e-value for the true label should not exceed 1. On the other hand, for the other labels (we will call them *false labels*), we would like their e-values to be as large as possible, and this (informal) desideratum is known as *efficiency* (or predictive efficiency, if there is a risk of confusion with computational efficiency). The topic of this section is ways of measuring the efficiency for conformal e-prediction.

We developed suitable criteria of efficiency for conformal prediction in [26, Sect. 3.1] (whose notation we will use in this section, except that e will stand for e-values). The idea is that a criterion of efficiency cannot be regarded as suitable if in the limiting case of infinite training and test sets it leads to very unnatural optimal nonconformity measures. An example of such an unsuitable criterion of efficiency for conformal prediction is the use of average confidence and credibility, as defined in the first edition of [26] (analogously to our definition in Sect. 2 above) and analysed in [26, Sections 3.1.6–3.1.7] (where extremely awkward features of this criterion are discussed). This section proposes natural criteria of efficiency for conformal e-prediction that lead to natural nonconformity e-measures.

Suppose we have a training set $z_1, \dots, z_n \in \mathbf{Z}$, and we are given a test set $z_{n+1}, \dots, z_{n+k}$, where $z_i = (x_i, y_i)$ for all i. Let $e_i^y := f(z_1, \dots, z_n, x_i, y)$ be the e-value computed by a given conformal e-predictor f for a postulated label y for a test object x_i, $i \in \{n+1, \dots, n+k\}$. The average sum

$$\frac{1}{k} \sum_{i=n+1}^{n+k} \sum_{y \neq y_i} \ln e_i^y \tag{21}$$

of the log e-values for the false test labels may serve as a measure of the predictive efficiency of f on the test set; we would like it to be as large as possible. Let us call it the *observed log criterion* of efficiency (the modifier "observed" will be discussed in Remark 16). The expression (21) is natural insofar as averaging logarithms of e-values is ubiquitous in numerous contexts: see, e.g., the five reasons given in [8, (7) and Sect. 3.1].

Let us find the optimal nonconformity e-measure for the observed log criterion in the "idealised setting" akin to the one considered in [26, Sect. 3.1.4] in the case of conformal prediction. For that it will be convenient to modify slightly the definition of a nonconformity e-measure.

An equivalent definition of a *nonconformity e-measure* A is as a measurable function mapping a nonempty multiset $\lbag z_1, \dots, z_m \rbag$, for any $m \in \{1, 2, \dots\}$, and its element z_i, $i \in \{1, \dots, m\}$, to a nonnegative number satisfying

$$\sum_{i=1}^{m} A(\lbag z_1, \dots, z_m \rbag, z_i) \leq 1$$

for all $m \in \{1, 2, \dots\}$ and all multisets $\Lbag z_1, \dots, z_m \Rbag$ of size m. The corresponding conformal e-predictor is then defined by

$$f(z_1, \dots, z_n, x, y) := A(\Lbag z_1, \dots, z_n, (x, y) \Rbag, (x, y)). \tag{22}$$

It is clear that this definition is equivalent to our original definition (2).

An *idealised nonconformity e-measure* is a function $A : \mathbf{P}(\mathbf{Z}) \times \mathbf{Z} \to [0, \infty)$ such that $A(Q, z)$ is measurable in $z \in \mathbf{Z}$ and $\int A(Q, z) Q(\mathrm{d}z) \leq 1$ for any $Q \in \mathbf{P}(\mathbf{Z})$. The intuition behind this definition is that Q represents an infinite training set, which becomes, once the order of its elements is forgotten, the probability distribution of the data. The corresponding *idealised conformal e-predictor* is $f(Q, x, y) := A(Q, x, y)$ according to (22) (adding another observation to an infinite training set does not change anything); the difference between a nonconformity e-measure and the corresponding conformal e-predictor disappears in the limit.

As in [26, Sect. 3.1.5], we assume that the object space $\mathbf{X}$ and the label space $\mathbf{Y}$ are finite (the latter simply means that we are interested in pattern recognition). We identify a probability measure Q on a finite set A (such as $\mathbf{Y}$) with a function mapping $a \in A$ to $Q(\{a\})$, thus often omitting the curly braces in expressions such as $Q(\{a\})$. If Q is a probability measure on $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, let $Q_{\mathbf{X}} \in \mathbf{P}(\mathbf{X})$ be its marginal probability measure on $\mathbf{X}$, and let $Q_x \in \mathbf{P}(\mathbf{Y})$, $x \in \mathbf{X}$, be the conditional probability measure on $\mathbf{Y}$ given x:

$$Q_{\mathbf{X}}(x) := Q(\{x\} \times \mathbf{Y}), \quad Q_x(y) := \frac{Q(x, y)}{Q_{\mathbf{X}}(x)}.$$

Let us fix the true data-generating probability measure $P \in \mathbf{P}(\mathbf{Z})$, representing a given infinitely large training set. For simplicity, let us assume that $P(z) > 0$ for all $z \in \mathbf{Z}$; this makes $P_x(y)$ well-defined and positive for all $x \in \mathbf{X}$ and $y \in \mathbf{Y}$. To formalize the test set being infinitely large as well, we replace the problem of maximizing (21) by the idealised optimization problem

$$\int_{\mathbf{Z}} \sum_{y' \in \mathbf{Y} \setminus \{y\}} \ln f(P, x, y') \, P(\mathrm{d}(x, y)) \to \max, \tag{23}$$

$f = A$ ranging over the conformal e-predictors (equivalently, over the nonconformity e-measures). This corresponds to letting $k \to \infty$ in (21).

Proposition 15. *The optimal nonconformity e-measure A under the observed log criterion* (23) *is given by the odds*

$$A(P, x, y) := \frac{1}{|\mathbf{Y}| - 1} \frac{1 - P_x(y)}{P_x(y)} \tag{24}$$

against the true label being y conditional on the object x.

Proof. The rest of this section will be mainly devoted to the proof of Proposition 15. First let us check that (24) is indeed a nonconformity e-measure. It is even true that its average is 1 over each P_x:

$$\int A(P,x,y)\,P_x(\mathrm{d}y) = \frac{1}{|\mathbf{Y}|-1}\sum_{y\in\mathbf{Y}}(1-P_x(y)) = 1.$$

We start from the case where we have no objects, corresponding to $|\mathbf{X}| = 1$ (the only object does not carry any information). This case is not only a gentle introduction to the general case, but also corresponds to the situation where our prediction is fully conditional on the object x (remember that earlier we assumed $|\mathbf{X}| < \infty$, and so the full conditioning is feasible for an infinite training set). Now we can regard P to be a probability measure on the label space $\mathbf{Y}$, and our goal is to find the optimal e-variable $E = Q/P$, where Q is an alternative probability measure on $\mathbf{Y}$.

In the case $|\mathbf{X}| = 1$, the optimization problem (23) is

$$\int \ln\frac{Q}{P}\,\mathrm{d}1 - \int \ln\frac{Q}{P}\,\mathrm{d}P \to \max,$$

where 1 is the counting measure on $\mathbf{Y}$ ($1(y) := 1$ for all $y \in \mathbf{Y}$). This optimization problem is equivalent to

$$\int \ln Q\,\mathrm{d}(1-P) \to \max,$$

which gives $Q(y) \propto 1 - P(y)$. So the optimal e-values are

$$E(y) \propto \frac{1-P(y)}{P(y)};$$

i.e., $E(y)$ is proportional to the odds against observing label y. The full expression is

$$E(y) = \frac{1}{|\mathbf{Y}|-1}\frac{1-P(y)}{P(y)},$$

which agrees with (24).

Now let us get rid of the assumption $|\mathbf{X}| = 1$. We will apply the result of the previous paragraph to each P_x, $x \in \mathbf{X}$. To find an explicit expression for the optimal e-variable $E : \mathbf{Z} \to [0,\infty)$, we split it into $E_x : \mathbf{Y} \to [0,\infty)$, $x \in \mathbf{X}$, defined by $E_x(y) := E(x,y)$. Our optimization problem is

$$\int_{\mathbf{X}}\int_{\mathbf{Y}} \ln E_x\,\mathrm{d}(1-P_x)P_{\mathbf{X}}(\mathrm{d}x) \to \max \tag{25}$$

subject to the constraint

$$\int_{\mathbf{X}}\int_{\mathbf{Y}} E_x\,\mathrm{d}P_x P_{\mathbf{X}}(\mathrm{d}x) \le 1; \tag{26}$$

without loss of generality, we replace "$\le$" by "$=$" in (26). Now the constraint (26) can be rewritten as

$$\int_{\mathbf{Y}} E_x\,\mathrm{d}P_x = 1 + \gamma_x, \quad \int_{\mathbf{X}} \gamma_x P_{\mathbf{X}}(\mathrm{d}x) = 0,$$

where the new variables γ_x take values in $[-1,\infty)$. By the result of the previous paragraph,

$$E_x = (1+\gamma_x)\tilde{E}_x,$$

where

$$\tilde{E}_x(y) = \frac{1}{|\mathbf{Y}|-1}\frac{1-P_x(y)}{P_x(y)}$$

is the normalized version of E_x. Maximizing the overall objective function in (25) can be rewritten as

$$\int_{\mathbf{X}}\int_{\mathbf{Y}} \ln \tilde{E}_x \,\mathrm{d}(1-P_x)P_{\mathbf{X}}(\mathrm{d}x) + (|\mathbf{Y}|-1)\int_{\mathbf{X}} \ln(1+\gamma_x)P_{\mathbf{X}}(\mathrm{d}x) \to \max,$$

and so our optimization problem reduces to

$$\int_{\mathbf{X}} \ln(1+\gamma_x)P_{\mathbf{X}}(\mathrm{d}x) \to \max \tag{27}$$

under the constraint

$$\int_{\mathbf{X}} \gamma_x P_{\mathbf{X}}(\mathrm{d}x) = 0.$$

By Jensen's inequality, the max in (27) is 0, and it is attained for $\gamma_x = 0$. This completes the proof of Proposition 15. □

Remark 16. In the terminology of [26, Sect. 3.1], the criterion of efficiency (21) is "observed" in that the efficiency of the conformal e-predictor on a test observation z_i is measured by an expression, namely $\sum_{y\ne y_i}\ln e_i^y$, that depends on the observed true label y_i. A natural alternative to (21) is the *prior log criterion*

$$\frac{1}{k}\sum_{i=n+1}^{n+k}\sum_{y}\ln e_i^y \tag{28}$$

obtained by replacing $\sum_{y\ne y_i}$ by $\sum_y$. The criterion (28) is called "prior" since for it the dependence on the true label disappears; we can compute the sum $\sum_y$ prior to observing the true label. Its idealised version is

$$\int_{\mathbf{X}} \sum_{y'\in\mathbf{Y}} \ln f(P,x,y')P_{\mathbf{X}}(\mathrm{d}x) \to \max.$$

Adapting the argument given above to this idealised criterion (and slightly simplifying the argument), we can see that the optimal nonconformity e-measure under this criterion is

$$A(P,x,y) := \frac{1}{|\mathbf{Y}|P_x(y)}.$$

(We have already used this nonconformity e-measure with P_x replaced by its estimate: see (20) above.) Notice that while in conformal prediction suitable ("conditionally proper") observed and prior criteria of efficiency lead to the same optimal conformity measures [26, Theorem 3.1], in the case of conformal e-prediction the optimal nonconformity e-measures are different, albeit typically very close for a large label space $\mathbf{Y}$.

8 Conclusion

In this paper we have discussed three strengths of conformal prediction (even if two of them briefly):

1. As set predictors, conformal predictors possess a property of validity that is both reasonably strong and very simple: their probability of error is bounded by a prespecified constant. This advantage is lost in conformal e-prediction (the probability of error being bounded by a prespecified constant becomes a very weak property of validity for conformal e-predictors, and stronger properties of validity, such as the one given at the end of B, are somewhat less intuitive).

2. In the online prediction protocol, conformal predictors as set predictors make errors at different steps independently. Conformal e-predictors satisfy a weakened version of this property, as discussed in Sect. 3.

3. In the case of regression, conformal prediction can be used for producing conformal predictive distributions, and so conformal predictors can also be used as probabilistic predictors. In this role, conformal predictors are automatically well-calibrated. This advantage is lost for conformal e-prediction.

Two of these strengths, 1 and 3, appear to be clear advantages of conformal prediction over conformal e-prediction. For strength 2, this is less obvious, but the picture for conformal prediction still appears simpler and nicer.

Conformal e-prediction has at least two advantages of its own:

- Designing conditional conformal e-predictors is much easier and using e-values adds flexibility, especially as compared with Mondrian conformal predictors; see Sect. 4.

- Cross-conformal e-predictors are provably valid, unlike cross-conformal predictors, as discussed in Sect. 6.

Looking for further advantages of conformal e-prediction is an interesting direction of further research.

Acknowledgments

Many thanks to Aaditya Ramdas for pointing out deficiencies in presentation in an earlier version and to Alex Balinsky and Ilia Nouretdinov for useful discussions. The anonymous reviewers' comments helped greatly in improving the presentation further. This research has been partially supported by Astra Zeneca, Stena Line, and Mitie.

References

[1] Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications.* Elsevier, Amsterdam, 2014.

[2] Alexander A. Balinsky and Alexander D. Balinsky. Enhancing conformal prediction using e-test statistics. *Proceedings of Machine Learning Research*, 230:65–72, 2024. COPA 2024.

[3] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *Annals of Statistics*, 49:486–507, 2021.

[4] Rajendra Bhatia and Chandler Davis. A better bound on the variance. *American Mathematical Monthly*, 107:353–357, 2000.

[5] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press, Cambridge, 2009.

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

[7] Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. Technical Report arXiv:2305.12616 [stat.ME], arXiv.org e-Print archive, June 2024.

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

[9] G. H. Hardy, John E. Littlewood, and George Pólya. *Inequalities.* Cambridge University Press, Cambridge, second edition, 1952.

[10] Harold Jeffreys. *Theory of Probability.* Oxford University Press, Oxford, third edition, 1961.

[11] Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. Technical Report arXiv:2209.15145 [cs.LG], arXiv.org e-Print archive, September 2022. ICLR 2023.

[12] Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, and Tuve Löfström. On the calibration of aggregated conformal predictors. *Proceedings of Machine Learning Research*, 60:154–173, 2017.

[13] Ilia Nouretdinov, Tom Melluish, and Vladimir Vovk. Ridge Regression Confidence Machine. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 385–392, San Francisco, CA, 2001. Morgan Kaufmann.

[14] Harris Papadopoulos, Konstantinos Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive Confidence Machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Proceedings of the Thirteenth European Conference on Machine Learning*, volume 2430 of *Lecture Notes in Computer Science*, pages 345–356, Berlin, 2002. Springer.

[15] Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Qualified predictions for large data sets in the case of pattern recognition. In *Proceedings of the First International Conference on Machine Learning and Applications*, pages 159–163, Las Vegas, NV, 2002. CSREA Press.

[16] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38:576–601, 2023.

[17] 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.

[18] Jieli Shen, Regina Liu, and Minge Xie. Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195:126–140, 2018.

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

[20] Albert N. Shiryaev. *Probability-2*. Springer, New York, third edition, 2019.

[21] William F. Stout. *Almost Sure Convergence*. Academic Press, New York, 1974.

[22] Vladimir Vovk. On-line Confidence Machines are well-calibrated. In *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science*, pages 187–196, Los Alamitos, CA, 2002. IEEE Computer Society.

[23] Vladimir Vovk. A universal well-calibrated algorithm for on-line classification. *Journal of Machine Learning Research*, 5:575–604, 2004.

[24] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

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

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

[27] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Minge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474, 2019. COPA 2017 Special Issue.

[28] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107:791–808, 2020.

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

[30] Abraham Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

[31] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar, editors. Special Issue on p-values. *American Statistician*, 73 (Supplement 1), 2019.

A Law of the iterated logarithm for e-flows

The main goal of this appendix is to prove the law of the iterated logarithm in the form of Proposition 6. As I could not find this version of the law of the iterated logarithm in literature, I will state it in a general form in this appendix.

Let $\mathcal{F}_n$, where $n \in \mathbb{Z}$ and $\mathbb{Z}$ stands for the set of all integer numbers, be an increasing sequence of σ-algebras on a given sample space Ω,

$$\dots \subseteq \mathcal{G}_{-1} \subseteq \mathcal{G}_0 \subseteq \mathcal{G}_1 \subseteq \dots,$$

and ξ_n, $n \in \mathbb{Z}$, be an adapted sequence of random variables, meaning that each ξ_n is $\mathcal{G}_n$-measurable. Such an adapted sequence is an *e-flow* if it is nonnegative and

$$\mathbb{E}(\xi_n \mid \mathcal{G}_{n-1}) \le 1, \quad n \in \mathbb{Z}. \tag{29}$$

Let us say that the e-flow is *exact* if (29) holds with "=" in place of "≤".

The main example of an e-flow in the context of this paper is

$$\mathcal{G}_n := \begin{cases} \mathcal{F}_{-n} & \text{if } n < 0 \\ \mathcal{F}_1 & \text{otherwise,} \end{cases}$$

where $\mathcal{F}_n$, $n = 1, 2, \dots$, is the *exchangeable filtration* (introduced in Sect. 3: $\mathcal{F}_n$ is generated by $\wr Z_1, \dots, Z_{n-1} \int, Z_n, Z_{n+1}, \dots)$, and

$$\xi_n := \begin{cases} E_{-n} & \text{if } n < 0 \\ 1 & \text{otherwise,} \end{cases}$$

where E_n is the nth conformal e-variable, $E_n := f(Z_1, \dots, Z_n)$. Let us call this the *conformal e-flow*. The fact that the adapted sequence $(\xi_n, \mathcal{G}_n)$ defined in this way is an e-flow is the main property of validity of conformal e-prediction.

The main part of the conformal e-flow is $(\xi_n, \mathcal{G}_n)$ for $n < 0$; the extension to all integer n is vacuous. Exact e-flows correspond to admissible conformal predictors.

There are two natural laws of the iterated logarithm for e-flows: the *forward law* describes the behaviour, as $N \to \infty$, of the sums $\sum_{n=1}^{N}(\xi_n - 1)$, and the *backward law* describes the behaviour of the sums $\sum_{n=-N}^{-1}(\xi_n - 1)$. The forward law is just a restatement of the standard martingale law of the iterated logarithm, since with each forward sequence ξ_n, $n \geq 1$, we can associate a supermartingale X that carries the same information by setting

$$X_n := \sum_{i=1}^{n}(\xi_i - 1),$$

with $X_0 := 0$; remember that the defining property of being a supermartingale is $\mathbb{E}(X_n \mid \mathcal{F}_{n-1}) \leq X_{n-1}$. Applying the standard law of the iterated logarithm to the supermartingale X_n gives us the following corollary.

Corollary 17. *For any bounded e-flow* $(\xi_n, \mathcal{G}_n)$*,*

$$\limsup_{N\to\infty}(N \ln\ln N)^{-1/2}\sum_{n=1}^{N}(\xi_n - 1) \leq \sqrt{2(C-1)} \quad \textit{a.s.},$$

where $C > 1$ *is an upper bound for* ξ_n*, and there exists an e-flow* $(\xi_n, \mathcal{G}_n)$ *bounded above by* C *(assuming* $C > 1$*) such that*

$$\limsup_{N\to\infty}(N \ln\ln N)^{-1/2}\sum_{n=1}^{N}(\xi_n - 1) = \sqrt{2(C-1)} \quad \textit{a.s.}$$

(Alternatively, the proof of the backward law of the iterated logarithm given later in this appendix will also prove the forward law, after trivial modifications.) The assumption $C > 1$ only excludes a trivial case.

Such a reduction to the supermartingale case is impossible for the backward law; even though there are numerous laws of the iterated logarithm for reverse martingales, they are not applicable in our current context. Luckily, however, the standard proof of the law of the iterated logarithm can be easily adapted to the backward case and allows us to establish the following backward law of the iterated logarithm.

Proposition 18. *For any bounded e-flow* $(\xi_n, \mathcal{G}_n)$*, we have*

$$\limsup_{N\to\infty}(N \ln\ln N)^{-1/2}\sum_{n=-N}^{-1}(\xi_n - 1) \leq \sqrt{2(C-1)} \quad \textit{a.s.}, \tag{30}$$

where $C > 1$ *is an upper bound, and there exists an e-flow* $(\xi_n, \mathcal{G}_n)$ *bounded above by* C *(assuming* $C > 1$*) that satisfies*

$$\limsup_{N\to\infty}(N \ln\ln N)^{-1/2}\sum_{n=-N}^{-1}(\xi_n - 1) = \sqrt{2(C-1)} \quad \textit{a.s.} \tag{31}$$

The proof of Proposition 18 uses the standard basic scheme (see, e.g., [20, Sect. 4.4]). As a first step we fix an e-flow $(\xi_n, \mathcal{G}_n)$ and assume, without loss of generality, that it is exact (an easy way to see that there is no loss of generality is to apply the idea of coupling [5, Sect. 7.4]).

We need an auxiliary exponential supermartingale from Stout's proof of the law of the iterated logarithm for martingales [21, Lemma 5.4.1].

Lemma 19. *Let $\eta_1, \eta_2, \dots$ be a martingale difference bounded above by a constant c, $\eta_n \le c$, w.r.t. to a filtration $\mathcal{F}_0, \mathcal{F}_1, \dots$. Set $S_n := \sum_{i=1}^{n} \eta_i$. Let $\delta \in (0, 1/c]$ be another constant. Then*

$$T_n := \exp\left(\delta S_n - \frac{\delta^2}{2}\left(1 + \frac{\delta c}{2}\right) \sum_{i=1}^{n} \mathbb{E}(\eta_i^2 \mid \mathcal{F}_{i-1})\right) \tag{32}$$

is a supermartingale.

In fact, Stout's lemma only assumes that η_n is a supermartingale difference, but we do not need this generality. Let us derive a corollary of this lemma that will allow us to establish (30). First we notice that $\mathbb{E}(\eta_i^2 \mid \mathcal{F}_{i-1}) \le c$ (the largest $\mathbb{E}(\eta_i^2 \mid \mathcal{F}_{i-1}) = c$ is attained for $\eta_i \in \{-1, c\}$ taking value c with probability $1/(c+1)$); this is spelled out in the following lemma.

Lemma 20. *For any $c > 0$, $\max_\eta \mathbb{E}(\eta^2) = c$, η ranging over the random variables with $\mathbb{E}(\eta) = 0$ and $\eta \in [-1, c]$.*

Proof. This is a special case of the Bhatia–Davis inequality [4]: see Theorem 1 there and the remark after its first proof. □

Since T_n defined by (32) is a test supermartingale, we have, by Ville's inequality,

$$\mathbb{P}\left\{\max_{n=0,\dots,N} T_n \ge \gamma\right\} \le \frac{1}{\gamma}$$

for any $\gamma > 1$ and any upper bound N on n. By Lemma 20 (modified to cover $\mathbb{E}(\eta^2 \mid \mathcal{F})$ in place of $\mathbb{E}(\eta^2)$), this implies

$$\mathbb{P}\left\{\max_{n=0,\dots,N} S_n \ge \frac{\delta}{2}\left(1 + \frac{\delta c}{2}\right) cN + \frac{\ln \gamma}{\delta}\right\} \le \frac{1}{\gamma}. \tag{33}$$

The minimum over δ of the sum

$$\frac{\delta}{2} cN + \frac{\ln \gamma}{\delta}$$

(with the term $\frac{\delta c}{2}$ in (33) ignored for now) is attained at

$$\delta := \sqrt{2 \ln \gamma / (cN)}, \tag{34}$$

and substituting this expression for δ into (33) gives

$$\mathbb{P}\left\{\max_{n=0,\dots,N} S_n \ge \sqrt{2cN \ln \gamma} + \frac{c \ln \gamma}{2}\right\} \le \frac{1}{\gamma} \tag{35}$$

(the condition $\delta \leq 1/c$ will be satisfied when we apply this inequality later in the proof). The inequality (35) is also applicable to $-S_n$ in place of S_n provided $c \geq 1$, and for any $c > 0$ it becomes applicable to $-S_n$ in place of S_n if we replace the second entry of c in it by 1 and assume $\delta \leq 1$.

Proof of Proposition 18. Let $\lambda > 1$ (later we will also let $\lambda \to 1$) and set $n_k := \lceil \lambda^k \rceil$, $k = 1, 2, \ldots$. For a given k, we will use the notation, for $k, n \in \{1, 2, \ldots\}$,

$$\eta_n^k := \xi_{-n_k-1+n} - 1 \text{ and } S_n^k := \sum_{i=1}^{n} \eta_i,$$

so that η_n^k, $n = 1, 2, \ldots$, is a martingale difference taking values in $[-1, C-1]$ and S_n^k, $n = 1, 2, \ldots$, is a martingale. Set $c := C - 1$.

Later we will choose a suitable function $\psi : \{1, 2, \ldots\} \to \mathbb{R}$; roughly, $\psi(n)$ will be an upper bound for $\sum_{i=-n}^{-1}(\xi_i - 1)$. Let A_k be the event that $S_{n_k}^k \geq \psi(n_k)$ and B_k be the event that $-S_n^k \geq \psi(n_k - n_{k-1})$ for some $n \in (0, n_k - n_{k-1}]$. Namely, we will choose ψ in such a way that, for sufficiently large k,

$$\mathbb{P}(A_k) \leq \mathbb{P}(S_n^k \geq \psi(n_k) \text{ for some } n \leq n_k) \leq k^{-\lambda} \tag{36}$$

and

$$\mathbb{P}(B_k) = \mathbb{P}(-S_n^k \geq \psi(n_k - n_{k-1}) \text{ for some } n \leq n_k - n_{k-1}) \leq k^{-\lambda}. \tag{37}$$

By the Borel–Cantelli lemma, as $\sum_k \mathbb{P}(A_k) < \infty$ and $\sum_k \mathbb{P}(B_k) < \infty$, A_k and B_k will hold only for finitely many k.

In order for the inequalities "$\leq k^{-\lambda}$" in (36) and (37) to hold, we can set, according to (35) (with $-S_n$ in place of S_n in the case of (37)),

$$\begin{aligned}
\psi(n_k) &= \sqrt{2cn_k \ln(k^\lambda)} + \frac{c\ln(k^\lambda)}{2} \sim \sqrt{2cn_k\lambda \ln k} \\
\psi(n_k - n_{k-1}) &= \sqrt{2c(n_k - n_{k-1})\ln(k^\lambda)} + \frac{\ln(k^\lambda)}{2} \sim \sqrt{2c(n_k - n_{k-1})\lambda \ln k}.
\end{aligned}$$

Therefore, we can choose $\psi(n) \sim \sqrt{2c\lambda n \ln\ln n}$. The conditions $\delta \leq 1/c$ and $\delta \leq 1$ mentioned earlier indeed hold, from some k on, for (34), $\gamma := k^\lambda$, and $N := n_k$.

According to (36) and (37), we will have, from some k on,

$$\begin{aligned}
&S_{n_k}^k \leq \psi(n_k), \\
&-S_n^k \leq \psi(n_k - n_{k-1}) \text{ for all } n \leq n_k - n_{k-1}.
\end{aligned}$$

For any sufficiently large N, these inequalities imply the inequality in

$$\sum_{n=-N}^{-1}(\xi_n - 1) = \sum_{n=1}^{n_k} \eta_n^k - \sum_{n=1}^{n_k - N} \eta_n^k = S_{n_k}^k - S_{n_k-N}^k \leq \psi(n_k) + \psi(n_k - n_{k-1})$$

$$\sim \sqrt{2c\lambda n_k \ln\ln n_k} + \sqrt{2c\lambda(n_k - n_{k-1})\ln\ln(n_k - n_{k-1})},$$

where k is the value satisfying $N \in [n_{k-1}, n_k)$. Since we can take λ arbitrarily close to 1, this completes the proof of (30).

To prove (31) for some bounded e-flow, suppose

$$\xi_n = \begin{cases} C & \text{with probability } 1/C \\ 0 & \text{with probability } 1 - 1/C \end{cases}$$

independently for $n = -1, -2, \dots$. Then $\operatorname{var}(\xi_n) = C - 1$, and applying the standard law of the iterated logarithm gives (31). □

Our argument given above proves the following more standard law of the iterated algorithm (which we do not need in this paper).

Proposition 21. *Let η_n, $n \in \mathbb{Z}$, be a bounded two-sided sequence of random variables adapted to a filtration $\mathcal{G}_n$, $n \in \mathbb{Z}$. Suppose $\mathbb{E}(\eta_n \mid \mathcal{G}_{n-1}) = 0$ for all $n \in \mathbb{Z}$. Set*

$$S_N := \sum_{n=-N}^{-1} \eta_n \text{ and } A_N := \sum_{n=-N}^{-1} \mathbb{E}(\eta_n^2 \mid \mathcal{G}_{n-1}). \tag{38}$$

Then

$$\limsup_{N\to\infty} \frac{S_N}{\sqrt{2A_N \ln\ln A_N}} = 1 \quad \text{a.s.}$$

provided $A_N \to \infty$ a.s. as $N \to \infty$.

In the proof of Proposition 21 we should use the stopping times $\tau_k := \min\{n : A_n \ge \lambda^k\}$ instead of the constant stopping times n_k. Of course, this proposition will remain true if we replace (38) by

$$S_N := \sum_{n=1}^{N} \eta_n \text{ and } A_N := \sum_{n=1}^{N} \mathbb{E}(\eta_n^2 \mid \mathcal{G}_{n-1}),$$

but then it will become just a special case of the standard martingale law of the iterated logarithm.

B Bounding the error probability of conformal e-predictors via Markov's inequality

The first advantage of conformal prediction mentioned in Sect. 1 is that conformal predictors can be used as set predictors, in which case their property of validity can be expressed as a low probability of error. In principle, this can also be done in the case of conformal e-prediction, and has been done in [2, Sect. 2], but it leads to a predictor that is not admissible in the terminology of statistical decision theory [30, Sect. 1.3]. This is the topic of this appendix. For simplicity we will assume that the object and label spaces $\mathbf{X}$ and $\mathbf{Y}$ are finite.

Let us fix the size n of the training set. A *set predictor* is a function $\Gamma : \mathbf{Z}^n \times \mathbf{X} \to 2^{\mathbf{Y}}$. It is *$\epsilon$-valid*, where $\epsilon > 0$, if $\mathbb{P}(Y \notin \Gamma(Z_1, \dots, Z_n, X)) \le \epsilon$ provided $Z_1, \dots, Z_n, (X, Y)$ are exchangeable; in other words, if the probability of error is bounded by ϵ. An example of an ϵ-valid set predictor is the *conformal ϵ-predictor*

$$\Gamma^\epsilon(z_1, \dots, z_n, x) := \{y : f(z_1, \dots, z_n, x, y) > \epsilon\},$$

where f is a conformal p-predictor [26, Proposition 2.3]. Conformal ϵ-predictors are just conformal predictors packaged as set predictors.

Let $\epsilon \in (0, 1)$ (informally, this is our target probability of error). The *BB-predictor* [2, Sect. 2] associated with an admissible conformal e-predictor f at significance level $1/\epsilon$ is defined as

$$\Gamma(z_1, \dots, z_n, x) := \{y : f(z_1, \dots, z_n, x, y) < 1/\epsilon\}. \tag{39}$$

In combination with Markov's inequality, Proposition 2 implies that the BB-predictor defined in this way is ϵ-valid.

An ϵ-valid set predictor Γ is *inadmissible* if there exists an ϵ-valid set predictor Γ' such that

$$\Gamma'(z_1, \dots, z_n, x) \subseteq \Gamma(z_1, \dots, z_n, x) \tag{40}$$

for all $z_1, \dots, z_n, x$ and the inclusion is strict for some $z_1, \dots, z_n, x$. Otherwise, Γ is *admissible.* We will say that Γ' *dominates* Γ if (40) holds for all $z_1, \dots, z_n, x$ and that the domination is *strict* if the inclusion in (40) is strict for some $z_1, \dots, z_n, x$.

First let us check that each BB-predictor is dominated by a conformal predictor.

Proposition 22. *Let $\epsilon \in (0, 1)$. The BB-predictor associated with an admissible conformal e-predictor f at significance level $1/\epsilon$ is ϵ-valid, but it is dominated by the conformal ϵ-predictor based on f's nonconformity e-measure.*

Proof. The validity was checked earlier. Let us check that the BB-predictor is dominated by the conformal ϵ-predictor Γ^ϵ (based on f's nonconformity e-measure as conformity measure): if

$$\frac{|\{i = 1, \dots, n+1 : \alpha_i \ge \alpha_{n+1}\}|}{n+1} > \epsilon,$$

then we have $\alpha_i \ge \alpha_{n+1}$ for at least $\lfloor (n+1)\epsilon \rfloor + 1$ α_is, which implies

$$\frac{\alpha_{n+1}}{\frac{1}{n+1}\sum_{i=1}^{n+1} \alpha_i} \le \frac{n+1}{\lfloor (n+1)\epsilon \rfloor + 1} < \frac{1}{\epsilon}. \qquad \square$$

In typical cases a BB-predictor will be inadmissible, being strictly dominated by the conformal predictor based on the same nonconformity (e-)measure. This will be formalized in the next proposition, for which we need two definitions.

A nonconformity e-measure A is *generic* if it always outputs $(\alpha_1, \ldots, \alpha_{n+1})$ that are all different (they may be only slightly different, or alternatively we can add slight randomization). In this case we will also say that the corresponding conformal e-predictor is generic.

A nonconformity measure A is *ϵ-categorical* if its nonconformity scores take only two values, 0 and $1/\epsilon$: it only outputs $(\alpha_1, \ldots, \alpha_{n+1})$ with $\alpha_i \in \{0, 1/\epsilon\}$ for all $i \in \{1, \ldots, n+1\}$. Categorical nonconformity e-measures A are a way of encoding set predictors: such an A represents the set predictor

$$\Gamma(z_1, \ldots, z_n, x) := \{y \in \mathbf{Y} : f(z_1, \ldots, z_n, x, y) > 0\},$$

where f is the corresponding conformal e-predictor. The next proposition says that the only way to avoid inadmissibility of the BB-predictor based on a generic nonconformity e-measure A is to make A very close to being ϵ-categorical, where the *deviation from being ϵ-categorical* is measured by

$$d(A) := \max_{(z_1, \ldots, z_{n+1}) \in \mathbf{Z}^{n+1}} \left(\sum_{i:\alpha_i \geq 1/\epsilon} (\alpha_i - 1/\epsilon) + \sum_{i:\alpha_i < 1/\epsilon} \alpha_i \right),$$

$\alpha_1, \ldots, \alpha_{n+1}$ are the nonconformity scores for $z_1, \ldots, z_{n+1}$, and i ranges over $\{1, \ldots, n+1\}$. We will also say that $d(f) := d(A)$ is the deviation from being ϵ-categorical for the conformal e-predictor f associated with A.

Proposition 23. *Let $\epsilon \in (0, 1)$. The BB-predictor associated with an admissible conformal e-predictor f at significance level $1/\epsilon$ is inadmissible if f is generic and $d(f) \geq 1/\epsilon$.*

The lower bound of $1/\epsilon$ on the deviation $d(f)$ in Proposition 23 is very small for a large size n of the training set, as the typical order of magnitude for $d(f)$ is n. Therefore, BB-predictors are typically inadmissible.

Proof of Proposition 23. In order for the BB-predictor to be admissible, for each $(z_1, \ldots, z_{n+1}) \in \mathbf{Z}^{n+1}$, the $\lfloor (n+1)\epsilon \rfloor$ largest α_i in the corresponding $(\alpha_1, \ldots, \alpha_{n+1})$ should all be at least $1/\epsilon$. Therefore, the deviation from being ϵ-categorical should be at most

$$n + 1 - \frac{\lfloor (n+1)\epsilon \rfloor}{\epsilon} < n + 1 - \frac{(n+1)\epsilon - 1}{\epsilon} = 1/\epsilon. \qquad \square$$

The key reason for the inadmissibility of the BB-predictor in typical situations is the strength of the validity property of conformal e-prediction. To see this strength more clearly, let us generalize the definition (39) by allowing the significance level α to be any positive number:

$$\Gamma^\alpha(z_1, \ldots, z_n, x) := \{y : f(z_1, \ldots, z_n, x, y) < \alpha\}. \tag{41}$$

The identity

$$\mathbb{E}(E) = \int_0^\infty \mathbb{P}(E \geq \alpha) \, \mathrm{d}\alpha$$

allows us to state the validity property for f in terms of Γ^{α} as follows: the probability of error for Γ^{α} should integrate to at most 1 over α. This is much stronger than requiring the probability of error for Γ^{α} to be at most $1/\alpha$ for each α. The new requirement is joint rather than being a separate requirement for each α.