Universal predictive systems

Vladimir Vovk

May 7, 2021

Abstract

This paper describes probability forecasting systems that are *universal*, or *universally consistent*, in the sense of being consistent under any data-generating distribution, assuming that the observations are produced independently in the IID fashion. The notion of universal consistency is asymptotic and does not imply any small-sample guarantees of validity. On the other hand, the method of conformal prediction has been recently adapted to producing predictive distributions that satisfy a natural property of small-sample validity, namely they are automatically probabilistically calibrated. The main result of the paper is the existence of universal conformal predictive systems, which output predictive distributions that are both probabilistically calibrated and universally consistent.

The version of this paper at http://alrw.net (Working Paper 18, first posted on 17 April 2017) is updated most often. The conference version is published in the Proceedings of COPA 2019 (Proceedings of Machine Learning Research 105:105–122).

1 Introduction

Predictive distributions are probability distributions for future labels satisfying a natural property of validity. They were introduced independently by Schweder and Hjort [19, Chapter 12] and Shen et al. [21], who also gave several examples of predictive distributions in parametric statistics. Earlier, related notions had been studied extensively by Tilmann Gneiting with co-authors and their predecessors (see, e.g., the review [9]). First nonparametric predictive distributions were constructed in the conference version of [32] based on the method of conformal prediction (see, e.g., [28, 29, 15, 16]). The nonparametric statistical model used in [32] is the one that is standard in machine learning: the observations are produced independently from the same probability measure; we will refer to it as the *IID model* in this paper. To make the notion of predictive distributions applicable in the nonparametric context, [32] slightly generalizes it allowing randomization; unless the amount of training data is very small, randomization affects the predictive distribution very little, but it simplifies definitions.

This paper follows [32, 30] in studying randomized predictive distributions under the IID model. Namely, we construct randomized predictive distributions

that, in addition to the small-sample property of validity that is satisfied automatically, satisfy an asymptotic property of universal consistency; informally, the true conditional distribution of the label and the randomized predictive distribution for it computed from the corresponding object and training data of size n approach each other as $n \to \infty$. (The procedures studied in [32, 30] were based on the Least Squares method and its modifications, and thus far from universally consistent; cf. Example 14 below.)

Our approach is in the spirit of Gneiting et al.'s [8] paradigm (which they trace back to Murphy and Winkler [18]) of *maximizing the sharpness of the predictive distributions subject to calibration*. We, however, refer to calibration as validity, sharpness as efficiency, and include a validity requirement in the definition of predictive distributions (following Shen et al. [21]). Martin and Liu [17, Section 3.3] state a similar "Efficiency Principle": *Subject to the validity constraint, probabilistic inference should be made as efficient as possible.*

We are mostly interested in results about the existence (and in explicit constructions) of randomized predictive distributions that satisfy two appealing properties: the small-sample property of validity and the asymptotic property of universal consistency. However, if we do not insist on the former, randomization becomes superfluous (Theorem 26).

As in [32, 30], our main technical tool will be conformal prediction. Before those papers, conformal prediction was typically applied for computing prediction sets. Conformal predictors are guaranteed to satisfy a property of validity, namely the correct coverage probability, and a remaining desideratum is their efficiency, namely the smallness of their prediction sets. Asymptotically efficient conformal predictors were constructed by Lei et al. [15] in the unsupervised setting and Lei and Wasserman [16] in the supervised setting (namely, for regression). This paper can be considered another step in this direction, where the notion of efficiency is formalized as universal consistency.

For convenience, in this paper we will refer to procedures producing randomized predictive distributions as randomized predictive systems; in particular, conformal predictive systems are procedures producing conformal predictive distributions, i.e., randomized predictive systems obtained by applying the method of conformal prediction.

The main result of this paper (Theorem 31) is that there exists a universally consistent, or universal, conformal predictive system, in the sense that it produces predictive distributions that are consistent under any probability distribution for one observation. The notion of consistency is used in an unusual situation here, and our formalization is based on Belyaev's [3, 4, 24] notion of weakly approaching sequences of distributions. The construction of a universal conformal predictive system adapts standard arguments for universal consistency in classification and regression [25, 7, 11].

The importance of universal consistency is demonstrated in [27, Section 5]; namely, applying the expected utility maximization principle to the predictive distributions produced by a universal predictive system leads, under natural conditions, to asymptotically optimal decisions.

The main part of this paper starts, in Section 2, from definitions of several

basic classes of predictive systems. The most important of those classes is that of conformal predictive systems. The main result of the paper, Theorem 31 stated in Section 4, requires a slight generalization of conformal predictive systems (for which we retain the same name). Two simple versions of Theorem 31 are given in Section 3. The first version (Theorem 24) states the existence of universal predictive systems in a class of predictive systems that is wider than that of conformal predictive systems but still satisfies a small-sample property of validity (albeit a weaker one); we refer to them as Mondrian predictive systems. The other version of Theorem 31 given in Section 3 is even simpler and states the existence of a universal probability forecasting system, which is deterministic and not required to satisfy any small-sample properties of validity (Theorem 26). In conclusion, Section 5 summarizes the paper and lists some natural directions of further research.

Three appendices clarify main results of this paper and provide further information. Appendix A further explores the notion of universal consistency making it more tangible. In particular, Theorem 31 implies that the Lévy distance between the predictive distribution output by a universal conformal predictive system and the true conditional distribution of the label of the test object converges to zero. Appendix B explores another popular notion of validity for probability forecasting systems, marginal calibration, in relation to conformal predictive systems. Finally, Appendix C briefly reviews another conformal counterpart of probability forecasting systems, Venn predictors, which has been widely used in the case of classification. Venn predictors satisfy an interesting additional property of validity as compared with conformal predictive distributions.

The conference version of this paper was published as [26].

Remark 1. There is a widely studied sister notion to predictive distributions with a similar small-sample guarantee of validity, namely confidence distributions: see, e.g., [33]. Both confidence and predictive distributions go back to Fisher's fiducial inference. Whereas, under the nonparametric IID model of this paper, there are no confidence distributions, [32], [30], and this paper argue that there is a meaningful theory of predictive distributions even under the IID model.

2 Predictive distributions

This section defines several basic classes of predictive systems. We start in Subsection 2.1 from defining randomized predictive systems, which are required to satisfy the small-sample property of validity under the IID model. Next, in Subsection 2.2 we define conformal predictive systems, which are a subclass of randomized predictive systems. Subsection 2.3 introduces another subclass of randomized predictive systems, which is wider than the subclass of conformal predictive systems of Subsection 2.2; the elements of this wider subclass are called Mondrian predictive systems. One advantage of Mondrian predictive sys-

tems is that a universal Mondrian predictive system is much easier to construct than a universal conformal predictive system.

2.1 Randomized predictive distributions

In this subsection we give some basic definitions partly following [21] and [32]. Let $\mathbf{X}$ be a measurable space, which we will call the *object space*. The *observation space* is defined to be $\mathbf{Z} := \mathbf{X} \times \mathbb{R}$; its element $z = (x, y)$, where $x \in \mathbf{X}$ and $y \in \mathbb{R}$, is interpreted as an *observation* consisting of an *object* $x \in \mathbf{X}$ and its *label* $y \in \mathbb{R}$. A typical example is where x is a description of a house, and y is its price. The label is assumed to be a real number, and so we are dealing with a problem of regression. Our task is, given *training data* consisting of observations $z_i = (z_i, y_i)$, $i = 1, \dots, n$, and a new (test) object $x_{n+1} \in \mathbf{X}$, to predict the corresponding label y_{n+1}; the pair (x_{n+1}, y_{n+1}) will be referred to as the test observation. We will be interested in procedures whose output is independent of the ordering of the training data $(z_1, \dots, z_n)$; therefore, the training data can also be interpreted as a multiset rather than a sequence.

Let U be the uniform probability measure on the interval $[0, 1]$.

Definition 2. Let $Q : \cup_{n=1}^{\infty} (\mathbf{Z}^{n+1} \times [0, 1]) \to [0, 1]$ be a measurable function. We call it a *randomized predictive system* if it satisfies the following requirements:

R1 i For each n, each training data sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x_{n+1} \in \mathbf{X}$, the function $Q(z_1, \dots, z_n, (x_{n+1}, y), \tau)$ of y and τ is monotonically increasing in both y and τ (i.e., it is monotonically increasing in y for each τ, and it is monotonically increasing in τ for each y).

ii For each n, each training data sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x_{n+1} \in \mathbf{X}$, we have

$$\lim_{y \to -\infty} Q(z_1, \dots, z_n, (x_{n+1}, y), 0) = 0, \tag{1}$$
$$\lim_{y \to \infty} Q(z_1, \dots, z_n, (x_{n+1}, y), 1) = 1.$$

R2 For each n, the distribution of Q, as function of random training observations $z_1 \sim P, \dots, z_n \sim P$, a random test observation $z_{n+1} \sim P$, and a random number $\tau \sim U$, all assumed independent, is uniform, i.e.:

$$\forall \alpha \in [0, 1] : \mathbb{P}\left(Q(z_1, \dots, z_n, z_{n+1}, \tau) \le \alpha\right) = \alpha. \tag{2}$$

The function $Q(z_1, \dots, z_n, (x_{n+1}, \cdot), \tau)$ is the *predictive distribution (function)* output by Q for given training data $z_1, \dots, z_n$, test object x_{n+1}, and $\tau \in [0, 1]$.

Requirement R1 says, essentially, that, as a function of y, Q is a distribution function, apart from a slack caused by the dependence on the random number τ. The size of the slack is

$$Q(z_1, \dots, z_n, (x_{n+1}, y), 1) - Q(z_1, \dots, z_n, (x_{n+1}, y), 0) \tag{3}$$

(remember that Q is monotonically increasing in $\tau \in [0,1]$, according to requirement R1(i)). In typical applications the slack will be small unless there is little training data; see Remark 15 for details.

Requirement R2 says, informally, that the predictive distributions agree with the data-generating mechanism. It has a long history in the theory and practice of forecasting. The review by Gneiting and Katzfuss [9] refers to it as probabilistic calibration and describes it as critical in forecasting; [9, Section 2.2.3] reviews the relevant literature.

Remark 3. Requirements R1 and R2 are the analogues (introduced in [19, Chapter 12] and [21]) of similar requirements in the theory of confidence distributions: see, e.g., [33, Definition 1] or [19, Chapter 3].

Definition 4. Let us say that a randomized predictive system Q is *consistent* for a probability measure P on $\mathbf{Z}$ if, for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid x_{n+1}) \to 0 \qquad (n \to \infty) \tag{4}$$

in probability, where:

- Q_n is the predictive distribution $Q_n : y \mapsto Q(z_1, \dots, z_n, (x_{n+1}, y), \tau)$ output by Q as its forecast for the label y_{n+1} corresponding to the test object x_{n+1} based on the training data $(z_1, \dots, z_n)$, where $z_i = (x_i, y_i)$ for all i;
- $\mathbb{E}_P(f \mid x_{n+1})$ is the conditional expectation of $f(y)$ given $x = x_{n+1}$ under $(x, y) \sim P$;
- $z_i = (x_i, y_i) \sim P$, $i = 1, \dots, n+1$, and $\tau \sim U$, are assumed all independent.

It is clear that the notion of consistency given in Definition 4 does not depend on the choice of the version of the conditional expectation $\mathbb{E}_P(f \mid \cdot)$ in (4). The integral in (4) is not quite standard since we did not require Q_n to be exactly a distribution function, so we understand $\int f \mathrm{d}Q_n$ as $\int f \mathrm{d}\bar{Q}_n$ with the measure $\bar{Q}_n$ on $\mathbb{R}$ defined by $\bar{Q}_n((u, v]) := Q_n(v+) - Q_n(u+)$ for any interval $(u, v]$ of this form (nonempty, open on the left, and closed on the right) in $\mathbb{R}$.

Definition 5. A randomized predictive system Q is *universal*, or *universally consistent*, if it is consistent for any probability measure P on $\mathbf{Z}$.

As already mentioned in Section 1, Definition 5 is based on Belyaev's (see, e.g., [4]). Our goal is construction of universal randomized predictive systems.

2.2 Conformal predictive distributions

A way of producing randomized predictive distributions under the IID model has been proposed in [32]. This subsection reviews a basic version, and Subsection 4.1 introduces a simple extension.

Definition 6. A *conformity measure* is a measurable function $A : \cup_{n=1}^{\infty} \mathbf{Z}^{n+1} \to \mathbb{R}$ that is invariant with respect to permutations of the training observations, i.e.: for any n, any sequence $(z_1, \ldots, z_n) \in \mathbf{Z}^n$, any $z_{n+1} \in \mathbf{Z}$, and any permutation π of the set $\{1, \ldots, n\}$,

$$A(z_1, \ldots, z_n, z_{n+1}) = A\left(z_{\pi(1)}, \ldots, z_{\pi(n)}, z_{n+1}\right).$$

The standard interpretation of a conformity measure A is that the value $A(z_1, \ldots, z_n, z_{n+1})$ measures how well the new observation z_{n+1} conforms to the comparison data $(z_1, \ldots, z_n)$. In the context of this paper, and conformal predictive distributions in general, $A(z_1, \ldots, z_n, z_{n+1})$, where $z_{n+1} = (x_{n+1}, y_{n+1})$, measures how large the label y_{n+1} is, in view of the corresponding object x_{n+1} and comparison data $z_1, \ldots, z_n$.

Definition 7. Given a conformity measure A, we define the corresponding *conformal transducer* as

$$\begin{aligned} Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) := \frac{1}{n+1} \left|\left\{i = 1, \ldots, n+1 \mid \alpha_i^y < \alpha_{n+1}^y\right\}\right| \\ + \frac{\tau}{n+1} \left|\left\{i = 1, \ldots, n+1 \mid \alpha_i^y = \alpha_{n+1}^y\right\}\right|, \end{aligned} \quad (5)$$

where $n \in \{1, 2, \ldots\}$, $(z_1, \ldots, z_n) \in \mathbf{Z}^n$ is training data, $x_{n+1} \in \mathbf{X}$ is a test object, and for each $y \in \mathbb{R}$ the corresponding *conformity scores* α_i^y are defined by

$$\begin{aligned} \alpha_i^y &:= A(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, (x_{n+1}, y), z_i), \qquad i = 1, \ldots, n, \\ \alpha_{n+1}^y &:= A(z_1, \ldots, z_n, (x_{n+1}, y)). \end{aligned} \quad (6)$$

A function is a *conformal transducer* if it is the conformal transducer corresponding to some conformity measure.

The usual interpretation of (5) is as a randomized p-value obtained when testing the IID model for the training data extended by adding the test object x_{n+1} combined with a postulated label y (cf. Remark 16 at the end of this subsection). It is important in this definition that, for each postulated label y, all elements of the *augmented training data* sequence $(z_1, \ldots, z_n, (x_{n+1}, y))$ are treated symmetrically.

Definition 8. A *conformal predictive system* is a function that is both a conformal transducer and a randomized predictive system. If Q is a conformal predictive system, $Q(z_1, \ldots, z_n, (x_{n+1}, \cdot), \tau)$ are the corresponding *conformal predictive distributions* (or, more fully, conformal predictive distribution functions).

Example 9. The simplest non-trivial conformal predictive system is a version of the classical Dempster–Hill procedure (to use the terminology of [32]; Dempster [6] referred to it as direct probabilities and Hill [12, 13] as Bayesian nonparametric predictive inference, which was abbreviated to nonparametric predictive inference by Coolen [1]). The conformity measure is

$$A(z_1, \ldots, z_n, (x_{n+1}, y_{n+1})) := y_{n+1}, \quad (7)$$

so that it ignores the objects. Since the objects are ignored, we will write y_i in place of $z_i = (x_i, y_i) \in \mathbf{Z}$, omitting the objects from our notation. Now suppose we are given training data $y_1, \dots, y_n$ and are interested in the conformal predictive distribution for the next label y_{n+1}; for simplicity, we will assume that $y_1, \dots, y_n$ are all different. The conformity scores (6) are $\alpha_i^y = y_i$ and $\alpha_{n+1}^y = y$, and so the conformal predictive distribution is

$$Q(y_1, \dots, y_n, y, \tau) = \begin{cases} (i + \tau)/(n+1) & \text{if } y \in (y_{(i)}, y_{(i+1)}),\ i = 0, \dots, n \\ (i - 1 + 2\tau)/(n+1) & \text{if } y = y_{(i)},\ i = 1, \dots, n, \end{cases}$$

where $y_{(1)}, \dots, y_{(n)}$ is the sequence $y_1, \dots, y_n$ sorted in the increasing order, $y_{(0)} := -\infty$, and $y_{(n+1)} := \infty$. A more intuitive (and equally informative) representation can be given in terms of the intervals

$$Q(y_1, \dots, y_n, y) := [Q(y_1, \dots, y_n, y, 0), Q(y_1, \dots, y_n, y, 1)]\,;$$

namely,

$$Q(y_1, \dots, y_n, y) = \begin{cases} [i/(n+1), (i+1)/(n+1)] & \text{if } y \in (y_{(i)}, y_{(i+1)}),\ i = 0, \dots, n \\ [(i-1)/(n+1), (i+1)/(n+1)] & \text{if } y = y_{(i)},\ i = 1, \dots, n. \end{cases} \tag{8}$$

For a further discussion of the Dempster–Hill procedure in the context of conformal prediction, see [32]. For another example of a conformal predictive system (depending on the objects in a simple but non-trivial way), see Example 13.

Remark 10. Requirement R2 in the previous subsection is sometimes referred to as the frequentist validity of predictive or confidence distributions (see, e.g., [33] and [21]). It can be argued that there is no need to appeal to frequencies in these and similar cases (see, e.g., [20]). However, the property of validity enjoyed by conformal predictive systems is truly frequentist: for them R2 (see (2)) can be strengthened to say that the random numbers $Q(z_1, \dots, z_n, z_{n+1}, \tau_n)$, $n = 1, 2, \dots$, are distributed uniformly in $[0, 1]$ and independently, provided $z_n \sim P$ and $\tau_n \sim U$, $n = 1, 2, \dots$, are all independent [28, Theorem 8.1]. In combination with the law of large numbers this implies, e.g., that for $\epsilon \in (0, 1)$ the frequency of the event

$$Q(z_1, \dots, z_n, z_{n+1}, \tau_n) \in \left[\frac{\epsilon}{2}, 1 - \frac{\epsilon}{2}\right]$$

(i.e., the frequency of the central $(1 - \epsilon)$-prediction interval covering the true label) converges to $1 - \epsilon$ as $n \to \infty$. Notice that this frequentist conclusion depends on the independence of $Q(z_1, \dots, z_n, z_{n+1}, \tau_n)$ for different n; R2 alone is not sufficient.

For a natural class of conformity measures the corresponding conformal transducers are automatically conformal predictive systems.

Definition 11. A conformity measure A is *monotonic* if $A(z_1, \dots, z_{n+1})$ is:

- monotonically increasing in y_{n+1}, namely

  $$y_{n+1} \le y'_{n+1} \implies A(z_1, \dots, z_n, (x_{n+1}, y_{n+1})) \le A(z_1, \dots, z_n, (x_{n+1}, y'_{n+1}));$$

- monotonically decreasing in y_1, namely

  $$y_1 \le y'_1 \implies A((x_1, y_1), z_2, \dots, z_n, z_{n+1}) \ge A((x_1, y'_1), z_2, \dots, z_n, z_{n+1})$$

  (which is equivalent to being monotonically decreasing in y_i for any $i = 2, \dots, n$).

Let A_n be the restriction of A to $\mathbf{Z}^{n+1}$.

Lemma 12. *Suppose a monotonic conformity measure A satisfies the following three conditions:*

- *for all n, all training data sequences $(z_1, \dots, z_n)$, and all test objects x_{n+1},*

  $$\inf_y A(z_1, \dots, z_n, (x_{n+1}, y)) = \inf A_n, \tag{9}$$

  $$\sup_y A(z_1, \dots, z_n, (x_{n+1}, y)) = \sup A_n; \tag{10}$$

- *for each n, the $\inf_y$ in (9) is either attained for all $(z_1, \dots, z_n)$ and x_{n+1} or not attained for all $(z_1, \dots, z_n)$ and x_{n+1};*

- *for each n, the $\sup_y$ in (10) is either attained for all $(z_1, \dots, z_n)$ and x_{n+1} or not attained for all $(z_1, \dots, z_n)$ and x_{n+1}.*

Then the conformal transducer corresponding to A is a randomized predictive system.

As usual, the two inf in (9) are allowed to take value $-\infty$, and the two sup in (10) are allowed to take value ∞. The conditions of Lemma 12 will be satisfied if (9) and (10) hold with $\inf A_n$ and $\sup A_n$ replaced by $-\infty$ and ∞, respectively; we will usually use this simplified version of the lemma (except for the proof of our main result, where we will need a $[0, 1]$-valued conformity measure). Before proving Lemma 12, we will give less trivial examples of conformal predictive systems (cf. Example 9).

Example 13. In this example we will modify the conformity measure (7) of the Dempster–Hill procedure by making it dependent, in a very simple way, on the objects; it will satisfy all conditions of Lemma 12 (with $-\infty$ and ∞ on the right-hand sides of (9) and (10), respectively). Namely, we set

$$A\big((x_1, y_1), \dots, (x_n, y_n), (x_{n+1}, y_{n+1})\big) := y_{n+1} - \hat{y}_{n+1}, \tag{11}$$

where $\hat{y}_{n+1}$ is the label y_i corresponding to the *nearest neighbour* x_i of x_{n+1}: $i \in \arg\min_{k \in \{1, \dots, n\}} \rho(x_k, x_{n+1})$, ρ being a measurable metric on the object

space $\mathbf{X}$. In this example we only consider the case where the pairwise distances $\rho(x_i, x_j)$, $i, j \in \{1, \dots, n+1\}$, are all different; the definition will be completed in Example 32. For each $i \in \{1, \dots, n\}$, let $\hat{y}_i$ be the label y_j corresponding to the nearest neighbour x_j to x_i among $x_1, \dots, x_n$: $j \in \arg\min_{k \in \{1,\dots,n\} \setminus \{i\}} \rho(x_k, x_i)$. Let $I \subseteq \{1, \dots, n\}$ be the set of $i \in \{1, \dots, n\}$ such that x_i is closer to x_{n+1} than to any of x_j, $j \in \{1, \dots, n\} \setminus \{i\}$. The conformity scores (6) are $\alpha^y_{n+1} = y - \hat{y}_{n+1}$, where $\hat{y}_{n+1}$ is defined as in (11), $\alpha^y_i = y_i - y$ if $i \in I$, and $\alpha^y_i = y_i - \hat{y}_i$ if $i \in \{1, \dots, n\} \setminus I$. Solving the equation $\alpha^y_i = \alpha^y_{n+1}$ (cf. (5)) gives $y = C_i := (\hat{y}_{n+1} + y_i)/2$ if $i \in I$ and

$$y = C_i := \hat{y}_{n+1} + (y_i - \hat{y}_i) \tag{12}$$

if $i \in \{1, \dots, n\} \setminus I$. Assuming, for simplicity, that $C_1, \dots, C_n$ are all different, we obtain the conformal predictive distribution

$$Q(y_1, \dots, y_n, y) = \begin{cases} [i/(n+1), (i+1)/(n+1)] & \text{if } y \in (C_{(i)}, C_{(i+1)}),\ i = 0, \dots, n \\ [(i-1)/(n+1), (i+1)/(n+1)] & \text{if } y = C_{(i)},\ i = 1, \dots, n \end{cases} \tag{13}$$

(cf. (8)), where $C_{(1)}, \dots, C_{(n)}$ is the sequence $C_1, \dots, C_n$ sorted in the increasing order, $C_{(0)} := -\infty$, and $C_{(n+1)} := \infty$.

The naive nearest-neighbour modification of the Dempster–Hill predictive distribution (8) would be (13) with all C_i defined by (12). The conformal predictive distribution is different only for $i \in I$, and I is typically a small set (its expected size is 1). For such i the conformal predictive distribution modifies the residual $y_i - \hat{y}_i$ in (12) by replacing it by $(y_i - \hat{y}_{n+1})/2$. Intuitively, the nearest neighbour to x_i in the augmented set $\{x_1, \dots, x_{n+1}\}$ is x_{n+1}, so we would like to use y_{n+1} instead of $\hat{y}_i$; but since we do not know y_{n+1} as yet, we have to settle for its estimate $\hat{y}_{n+1}$, and the resulting loss of accuracy is counterbalanced by halving the new residual. This seemingly minor further modification ensures the small-sample property of validity R2.

Example 14. Another natural conformity measure is (11) with $\hat{y}_{n+1}$ being the Least Squares prediction of y_{n+1} computed for the object x_{n+1} given $z_1, \dots, z_n$ as training data; this makes $y_{n+1} - \hat{y}_{n+1}$ the deleted residual for y_{n+1}. Alternative definitions use ordinary residuals (where (x_{n+1}, y_{n+1}) is added to the training data) and studentized residuals (which are half-way between deleted and ordinary residuals, in a certain sense). These conformity measures give rise to what is called Least Squares Prediction Machines in [32]. Only the studentized version is a randomized predictive system; the other two versions satisfy property R1(i) only under the assumption of the absence of high-leverage points. See [32] for an in-depth study of properties of Least Squares Prediction Machines, especially of their asymptotic efficiency under a standard Gaussian linear model. Kernelized versions are studied in [30].

Remark 15. The degree to which a randomized predictive system is affected by randomness, for given training data $(z_1, \dots, z_n)$, test object x_{n+1}, and postulated label y, is (3). As already mentioned, in interesting cases this difference will

be small. For example, for the Dempster–Hill predictive system (Example 9), the nearest neighbour predictive system (Example 13), and Least Squares Prediction Machines (Example 14), the difference (3) is $1/(n+1)$ except for at most n values of y, apart from pathological cases (see, e.g., (8) and (13)). A randomized predictive system can be universal only if the difference (3) is small with high probability.

Proof of Lemma 12. We need to check requirements R1 and R2. R2 is the standard property of validity for conformal transducers (see, e.g., [28, Theorem 8.1]). The intuition behind the proof of this property is given in Remark 16 at the end of this subsection.

The second statement of R1(i) is that (5) is monotonically increasing in τ; this follows from (5) being a linear function of τ with a nonnegative slope (the slope is in fact always positive as $i = n+1$ is allowed).

The first statement of R1(i) is that (5) is monotonically increasing in y. We can rewrite (5) as

$$Q(z_1,\dots,z_n,(x_{n+1},y),\tau) = \frac{1}{n+1}\sum_{i=1}^{n+1}\left(1_{\{\alpha_i^y<\alpha_{n+1}^y\}} + \tau 1_{\{\alpha_i^y=\alpha_{n+1}^y\}}\right), \quad (14)$$

where $1_{\{E\}}$ stands for the indicator function of a property E, and it suffices to prove that each addend in (14) is monotonically increasing in y; we will assume $i \le n$ (the case $i = n+1$ is trivial). This follows from α_i^y being monotonically decreasing in y and α_{n+1}^y being monotonically increasing in y, and therefore,

$$1_{\{\alpha_i^y<\alpha_{n+1}^y\}} + \tau 1_{\{\alpha_i^y=\alpha_{n+1}^y\}}$$

taking all or some of the values 0, τ, 1 in this order as y increases.

For concreteness, we will prove only the first statement of R1(ii), (1). Fix an n. First let us assume that the $\inf_y$ in (9) is attained for all $(z_1,\dots,z_n)$ and x_{n+1}. We will have $\alpha_{n+1}^y = \inf A_n$ for sufficiently small y, and plugging $\tau := 0$ into (5) will give 0, as required. It remains to consider the case where the $\inf_y$ in (9) is not attained for any $(z_1,\dots,z_n)$ and x_{n+1}. Since $\min_{i=1,\dots,n}\alpha_i^0 > \inf A$, we will have, for sufficiently small y,

$$\alpha_{n+1}^y < \min_{i=1,\dots,n}\alpha_i^0 \le \min_{i=1,\dots,n}\alpha_i^y,$$

and so plugging $\tau := 0$ into (5) will again give 0. □

Remark 16. The proof of Lemma 12 refers to [28] for a complete proof of R2. However, the intuition behind the proof is easy to explain. Setting $\tau := 1$ and assuming that there are no ties among the conformity scores, the right-hand side of (5) evaluated at $y := y_{n+1}$ is the rank of the last observation (x_{n+1},y_{n+1}) in the augmented training data $(z_1,\dots,z_n,(x_{n+1},y_{n+1}))$. Under the IID model (and the weaker assumption of the exchangeability of all the $n+1$ observations), the rank is uniformly distributed in the set $\{1,\dots,n+1\}$. Dividing by $n+1$ and making $\tau \sim U$ leads to (5) (evaluated at $y := y_{n+1}$) being uniformly distributed in $[0,1]$ (even if some conformity scores are tied). This makes (5) a *bona fide* randomized p-value for testing the IID model.

2.3 Mondrian predictive distributions

First we simplify our task by allowing Mondrian predictive distributions, which are more general than conformal predictive distributions but enjoy the same property of validity R2.

Definition 17. A *taxonomy* κ is an equivariant measurable function that assigns to each sequence $(z_1, \dots, z_n, z_{n+1}) \in \mathbf{Z}^{n+1}$, for each $n \in \{1, 2, \dots\}$, an equivalence relation $\sim$ on $\{1, \dots, n+1\}$.

The requirement that κ be equivariant will be spelled out in Definition 18. The idea behind a taxonomy is to determine the comparison class for computing the p-value (5); instead of using all available data we only use the observations that are equivalent to the test observation (intuitively, similar to it in some respect, with the aim of making the p-value more relevant).

The notation $(i \sim j \mid z_1, \dots, z_{n+1})$, where $i, j \in \{1, \dots, n+1\}$, means that i is equivalent to j under the equivalence relation assigned by κ to $(z_1, \dots, z_{n+1})$ (where κ is always clear from the context and not reflected in our notation). The measurability of κ means that, for all n, i, and j, the set $\{(z_1, \dots, z_{n+1}) \mid (i \sim j \mid z_1, \dots, z_{n+1})\}$ is measurable.

Definition 18. A permutation π of $\{1, \dots, n+1\}$ *respects* an equivalence relation $\sim$ if $\pi(i) \sim i$ for all $i = 1, \dots, n+1$. The requirement that a Mondrian taxonomy κ be *equivariant* means that, for each n, each $(z_1, \dots, z_{n+1}) \in \mathbf{Z}^{n+1}$, and each permutation π of $\{1, \dots, n+1\}$ respecting the equivalence relation assigned by κ to $(z_1, \dots, z_{n+1})$, we have

$$(i \sim j \mid z_1, \dots, z_{n+1}) \Longrightarrow (\pi(i) \sim \pi(j) \mid z_{\pi(1)}, \dots, z_{\pi(n+1)}). \tag{15}$$

Remark 19. The notion of taxonomy used in this paper is introduced in [31] under the name of Venn taxonomies and subsumes Mondrian taxonomies as defined in [28, Section 4.5], Venn taxonomies as defined in [28, Section 6.3], and n-taxonomies as defined in [2, Section 2.2]. A narrower notion of taxonomy requires that (15) hold for all permutations π of $\{1, \dots, n+1\}$; the taxonomy of Subsection 3.2 belongs to this narrower class.

Definition 20. Define

$$\kappa(j \mid z_1, \dots, z_{n+1}) := \{i \in \{1, \dots, n+1\} \mid (i \sim j \mid z_1, \dots, z_{n+1})\}$$

to be the equivalence class of j. The *Mondrian transducer* corresponding to a taxonomy κ and a conformity measure A is

$$\begin{aligned} &Q(z_1, \dots, z_n, (x_{n+1}, y), \tau) \\ &\quad := \frac{\left|\left\{i \in \kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y)) \mid \alpha_i^y < \alpha_{n+1}^y\right\}\right|}{\left|\kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y))\right|} \\ &\qquad + \tau \frac{\left|\left\{i \in \kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y)) \mid \alpha_i^y = \alpha_{n+1}^y\right\}\right|}{\left|\kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y))\right|}, \end{aligned} \tag{16}$$

where $n \in \{1, 2, \ldots\}$, $(z_1, \ldots, z_n) \in \mathbf{Z}^n$ is training data, $x_{n+1} \in \mathbf{X}$ is a test object, and for each $y \in \mathbf{Y}$ the corresponding conformity scores α_i^y and α_{n+1}^y are still defined by (6). A function is a *Mondrian transducer* if it is the Mondrian transducer corresponding to some taxonomy and conformity measure. A *Mondrian predictive system* is a function that is both a Mondrian transducer and a randomized predictive system, as defined in Subsection 2.1.

Notice that the denominator in (16) is always positive. The Mondrian p-value (16) differs from the original p-value (5) in that it uses only the equivalence class of the test observation (with a postulated label) as comparison class. See [28, Fig. 4.3], for the origin of the attribute "Mondrian".

Lemma 21. *If a taxonomy does not depend on the labels and a conformity measure is monotonic and satisfies the three conditions of Lemma 12, the corresponding Mondrian transducer will be a randomized (and, therefore, Mondrian) predictive system.*

Proof. As in Lemma 12, the conformity scores (defined by (6)) α_i^y are monotonically increasing in y when $i = n + 1$ and monotonically decreasing in y when $i = 1, \ldots, n$. Since the equivalence class of $n + 1$ in (16) does not depend on y, the value of (16) is monotonically increasing in y: it suffices to replace (14) by

$$Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) = \frac{1}{|\kappa(n+1 \mid z_1, \ldots, z_n, (x_{n+1}, y))|} \sum_{i \in \kappa(n+1 \mid z_1, \ldots, z_n, (x_{n+1}, y))} \left(1_{\{\alpha_i^y < \alpha_{n+1}^y\}} + \tau 1_{\{\alpha_i^y = \alpha_{n+1}^y\}} \right)$$

in the argument of Lemma 12. In combination with the obvious monotonicity in τ, this proves R1(i). R1(ii) is demonstrated as in Lemma 12. The proof of R2 is standard and valid for any taxonomy (see, e.g., [28, Section 8.7]); the intuition behind it is given in Remark 22 below. □

The properties listed in Lemma 21 will be satisfied by the conformity measure and taxonomy defined in Subsection 3.2 to prove Theorem 24, a weaker form of the main result of this paper.

Remark 22. Remark 16 can be easily adapted to Mondrian predictive systems. For $\tau := 1$ and assuming no ties among the conformity scores, the right-hand side of (16) at $y := y_{n+1}$ is the rank of the last observation (x_{n+1}, y_{n+1}) in its equivalence class divided by the size of the equivalence class. Let us introduce another notion of equivalence: sequences $(z_1, \ldots, z_{n+1})$ and $(z'_1, \ldots, z'_{n+1})$ in $\mathbf{Z}^{n+1}$ are *equivalent* if

$$(z'_1, \ldots, z'_{n+1}) = \left(z_{\pi(1)}, \ldots, z_{\pi(n+1)} \right)$$

for some permutation π of $\{1, \ldots, n + 1\}$ that respects the equivalence relation assigned by κ to $(z_1, \ldots, z_{n+1})$; this is indeed an equivalence relation since κ is equivariant. The stochastic mechanism generating the augmented training data

(the IID model) can be represented as generating an equivalence class (which is always finite) and then generating the actual sequence of observations in $\mathbf{Z}^{n+1}$ from the uniform probability distribution on the equivalence class. Already the second step ensures that the rank is distributed uniformly in the set of its possible values, which leads to (16) being uniformly distributed in $[0,1]$, provided $y := y_{n+1}$ and $\tau \sim U$.

Remark 23. One advantage of conformal predictive systems over Mondrian predictive systems is that the former satisfy a stronger version of R2, as explained in Remark 10.

3 Basic results on universal consistency

A simplified version of the main result of this paper is given in Subsection 3.1 as Theorem 24, and it states the existence of Mondrian predictive systems that are universal. An example of a universal Mondrian predictive system is given in Subsection 3.2, and Subsection 3.3 is devoted to a short proof that this predictive system is indeed universal. Subsection 3.4 gives an even shorter proof of the existence of a universal probability forecasting system (Theorem 26).

3.1 Universal Mondrian predictive systems and probability forecasting systems

Our results (Theorems 24, 26, and 31) will assume that the object space $\mathbf{X}$ is standard Borel (see, e.g., [14, Definition 12.5]); the class of standard Borel spaces is very wide and contains, e.g., all Euclidean spaces $\mathbb{R}^d$. In this subsection we start from an easy result (Theorem 24) and its adaptation to deterministic forecasting (Theorem 26).

Theorem 24. *If the object space $\mathbf{X}$ is standard Borel, there exists a universal Mondrian predictive system.*

In Subsection 3.2 we will construct a Mondrian predictive system that will be shown in Subsection 3.3 to be universal.

Belyaev's generalization of weak convergence can also be applied in the situation where we do not insist on small-sample validity; for completeness, we will state a simple corollary of the proof of Theorem 24 covering this case (Theorem 26 below).

Definition 25. A *probability forecasting system* is a measurable function $Q : \cup_{n=1}^{\infty} \mathbf{Z}^{n+1} \to [0,1]$ such that:

- for each n, each training data sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x_{n+1} \in \mathbf{X}$, $Q(z_1, \dots, z_n, (x_{n+1}, y))$ is monotonically increasing in y;
- for each n, each training data sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x_{n+1} \in \mathbf{X}$, we have

$$\lim_{y \to -\infty} Q(z_1, \dots, z_n, (x_{n+1}, y)) = 0,$$

$$\lim_{y\to\infty} Q(z_1,\dots,z_n,(x_{n+1},y)) = 1;$$

- for each n, each training data sequence $(z_1,\dots,z_n) \in \mathbf{Z}^n$, and each test object $x_{n+1} \in \mathbf{X}$, the function $Q(z_1,\dots,z_n,(x_{n+1},\cdot))$ is right-continuous (and therefore, a *bona fide* distribution function).

A probability forecasting system Q is *universal*, or *universally consistent*, if, for any probability measure P on $\mathbf{Z}$ and any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$, (4) holds in probability, where $Q_n : y \mapsto Q(z_1,\dots,z_n,(x_{n+1},y))$, assuming $z_n \sim P$ are independent.

Theorem 26. *If the object space* $\mathbf{X}$ *is standard Borel, there exists a universal probability forecasting system.*

Theorem 26 will be proved in Subsection 3.4.

3.2 Histogram Mondrian predictive systems

Remember that the measurable space $\mathbf{X}$ is assumed to be standard Borel. Since every standard Borel space is isomorphic to $\mathbb{R}$ or a countable set with discrete σ-algebra (combine Theorems 13.6 and 15.6 in [14]), $\mathbf{X}$ is isomorphic to a Borel subset of $\mathbb{R}$. Therefore, we can set, without loss of generality, $\mathbf{X} := \mathbb{R}$, which we will do.

Definition 27. Fix a monotonically decreasing sequence h_n of powers of 2 such that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Let $\mathcal{P}_n$ be the partition of $\mathbf{X}$ into the intervals $[kh_n,(k+1)h_n)$, where k are integers. We will use the notation $\mathcal{P}_n(x)$ for the interval (cell) of $\mathcal{P}_n$ that includes $x \in \mathbf{X}$. Let A be the conformity measure defined by $A(z_1,\dots,z_n,z_{n+1}) := y_{n+1}$, where y_{n+1} is the label in z_{n+1}. This conformity measure will be called the *trivial conformity measure*. The taxonomy under which $(i \sim j \mid z_1,\dots,z_{n+1})$ is defined to mean $x_j \in \mathcal{P}_n(x_i)$ is called the *histogram taxonomy*.

Lemma 28. *The trivial conformity measure is monotonic and satisfies all other conditions of Lemma 12. Therefore, the Mondrian transducer corresponding to it and the histogram taxonomy is a randomized predictive system.*

Proof. The infimum on the left-hand side of (9) is always $-\infty$ and never attained, and the supremum on the left-hand side of (10) is always ∞ and never attained. By definition, the histogram taxonomy does not depend on the labels. It remains to apply Lemma 21. □

Definition 29. The Mondrian predictive system corresponding to the trivial conformity measure and histogram taxonomy is called the *histogram Mondrian predictive system*.

The histogram Mondrian predictive system will be denoted Q in the next subsection, where we will see that it is universal.

3.3 Proof of Theorem 24

Let us fix a probability measure P on $\mathbf{Z}$; our goal is to prove the convergence (4) in probability. We fix a version of the conditional expectation $\mathbb{E}_P(f \mid x)$, $x \in \mathbf{X}$, and use it throughout the rest of this paper. We can split (4) into two tasks:

$$\mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) - \mathbb{E}_P(f \mid x_{n+1}) \to 0, \tag{17}$$

$$\int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \to 0, \tag{18}$$

where $\mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1}))$ is the conditional expectation of $f(y)$ given $x \in \mathcal{P}_n(x_{n+1})$ under $(x, y) \sim P$.

The convergence (17) follows by Paul Lévy's martingale convergence theorem [23, Theorem 7.4.3]. Paul Lévy's theorem is applicable since, by our assumption, the partitions $\mathcal{P}_n$ are nested (as h_n are powers of 2) and, therefore, the random variables $\mathbb{E}_P(f \mid \mathcal{F}_n)$ form a martingale, where $\mathcal{F}_n$ is the σ-algebra on $\mathbf{X} \times \mathbb{R}$ generated by $\mathcal{P}_n$. This theorem implies $\mathbb{E}_P(f \mid \mathcal{P}_n(x)) - \mathbb{E}_P(f \mid x) \to 0$ P-almost surely and, therefore, in probability when $(x, y) \sim P$. The last convergence is clearly equivalent to (17) (in P^∞-probability).

It remains to prove (18). Let $\epsilon > 0$; we will show that

$$\left| \int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \right| \le \epsilon \tag{19}$$

with high probability for large enough n. By [7, the proof of Theorem 6.2], the number N of observations $z_i = (x_i, y_i)$ among $z_1, \dots, z_n$ such that $x_i \in \mathcal{P}_n(x_{n+1})$ tends to infinity in probability. Therefore, it suffices to prove that (19) holds with high conditional probability given $N > K$ for large enough K. Moreover, it suffices to prove that, for large enough K, (19) holds with high conditional probability given $x_1, \dots, x_{n+1}$ such that at least K of objects x_i among $x_1, \dots, x_n$ belong to $\mathcal{P}_n(x_{n+1})$. (The remaining randomness is in the labels.) Let $I \subseteq \{1, \dots, n\}$ be the indices of those objects; remember that our notation for $|I|$ is N. By the law of large numbers, the probability (over the random labels) of

$$\left| \frac{1}{N} \sum_{i \in I} f(y_i) - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \right| \le \epsilon/2 \tag{20}$$

can be made arbitrarily high by increasing K. It remains to notice that

$$\int f \mathrm{d}Q_n = \frac{1}{N+1} \sum_{i \in I} f(y_i); \tag{21}$$

this follows from $\bar{Q}_n$ (in the notation of Subsection 2.1) being concentrated at the points y_i, $i \in I$, and assigning weight $a_i/(N+1)$ to each such y_i, where a_i is its multiplicity in the multiset $\{y_i \mid i \in I\}$ (our use of the same notation for sets and multisets is always counterbalanced by using unambiguous descriptors). Interestingly, $\int f \mathrm{d}Q_n$ in (21) does not depend on the random number τ.

3.4 Proof of Theorem 26

Define a probability forecasting system Q by the requirement that

$$Q_n(\cdot) := Q(z_1, \dots, z_n, (x_{n+1}, \cdot))$$

be the distribution function of the empirical probability measure of the multiset $\{y_i \mid i \in I\}$, in the notation of the previous subsection. In other words, the probability measure corresponding to Q_n is concentrated on the set $\{y_i \mid i \in I\}$ and assigns the weight a_i/N to each element y_i of this set, where a_i is its multiplicity in the multiset $\{y_i \mid i \in I\}$. (This is very similar to $\bar{Q}_n$ at the end of the previous subsection.) If $I = \emptyset$, let $Q_n(\cdot)$ be the distribution function of the probability measure concentrated at 0. We still have (20) with high probability, and we have (21) with N in place of $N + 1$.

4 Main result

The main result of this paper, Theorem 31, is stated in Subsection 4.1. It asserts the existence of universal conformal predictive systems. An example of such a conformal predictive system is given in Subsection 4.2, and it is shown in Subsection 4.3 to be universal. One advantage of Theorem 31 over the result of Subsection 3.1 (Theorem 24) is that, as compared with Mondrian predictive systems, conformal predictive systems enjoy a stronger small-sample property of validity (see Remarks 10 and 23).

4.1 Universal conformal predictive systems

In this subsection we will introduce a clearly innocuous extension of conformal predictive systems allowing further randomization. In particular, the extension will not affect the small-sample property of validity, R2 (or its stronger version given in Remark 10).

First we extend the notion of a conformity measure.

Definition 30. A *randomized conformity measure* is a measurable function $A : \cup_{n=1}^{\infty}(\mathbf{Z} \times [0,1])^{n+1} \to \mathbb{R}$ that is invariant with respect to permutations of extended training observations: for any n, any sequence $(z_1, \dots, z_{n+1}) \in \mathbf{Z}^{n+1}$, any sequence $(\theta_1, \dots, \theta_{n+1}) \in [0,1]^{n+1}$, and any permutation π of $\{1, \dots, n\}$,

$$A\big((z_1, \theta_1), \dots, (z_n, \theta_n), (z_{n+1}, \theta_{n+1})\big) = A\left((z_{\pi(1)}, \theta_{\pi(1)}), \dots, (z_{\pi(n)}, \theta_{\pi(n)}), (z_{n+1}, \theta_{n+1})\right).$$

This is essentially Definition 6 of Subsection 2.2, except that each observation is extended by adding a number (later it will be generated randomly from U) that can be used for tie-breaking. We can still use the same definition, given by the right-hand side of (5), of the conformal transducer corresponding to a

randomized conformity measure A, except for replacing each observation in (6) by an extended observation:

$$\begin{aligned}\alpha_i^y &:= A\big((z_1,\theta_1),\dots,(z_{i-1},\theta_{i-1}),(z_{i+1},\theta_{i+1}),\dots,(z_n,\theta_n),(x_{n+1},y,\theta_{n+1}),\\ &\qquad (z_i,\theta_i)\big), \qquad i=1,\dots,n,\\ \alpha_{n+1}^y &:= A\big((z_1,\theta_1),\dots,(z_n,\theta_n),(x_{n+1},y,\theta_{n+1})\big).\end{aligned}$$

Notice that our new definition of conformal transducers is a special case of the old definition, in which the original observation space $\mathbf{Z}$ is replaced by the extended observation space $\mathbf{Z}\times[0,1]$. An extended observation $(z,\theta)=(x,y,\theta)$ will be interpreted to consist of an extended object (x,θ) and a label y. The main difference from the old framework is that now we are only interested in the probability measures on $\mathbf{Z}\times[0,1]$ that are the product of a probability measure P on $\mathbf{Z}$ and the uniform probability measure U on $[0,1]$.

The definitions of randomized predictive systems and monotonic conformity measures generalize by replacing objects x_j by extended objects (x_j,θ_j). We still have Lemma 12. Conformal predictive systems are defined literally as before.

Theorem 31. *Suppose the object space* $\mathbf{X}$ *is standard Borel. There exists a universal conformal predictive system.*

In Subsection 4.2 we will construct a conformal predictive system that will be shown in Subsection 4.3 to be universal. The corresponding randomized conformity measure will be monotonic and satisfy all the conditions of Lemma 12 (with objects replaced by extended objects).

Example 32. We can use the notion of a randomized conformity measure to complete the definition in Example 13. Now we drop the assumption that the pairwise distances among $x_1,\dots,x_{n+1}$ are all different. We can use the same conformity measure (11), except that now the index j of the nearest neighbour x_j of x_i, $i\in\{1,\dots,n+1\}$, is chosen randomly from the uniform probability measure on the set $\arg\min_{k\in\{1,\dots,n\}\setminus\{i\}}\rho(x_k,x_i)$.

4.2 Histogram conformal predictive systems

In this subsection we will use the same partitions $\mathcal{P}_n$ of $\mathbf{X}=\mathbb{R}$ as in Subsection 3.2.

Definition 33. The *histogram conformity measure* is defined to be the randomized conformity measure A with $A((z_1,\theta_1),\dots,(z_n,\theta_n),(z_{n+1},\theta_{n+1}))$ defined as a/N, where N is the number of objects among $x_1,\dots,x_n$ that belong to $\mathcal{P}_n(x_{n+1})$ and a is essentially the rank of y_{n+1} among the labels corresponding to those objects; formally,

$$a := \left|\{i=1,\dots,n \mid x_i\in\mathcal{P}_n(x_{n+1}),(y_i,\theta_i)\le(y_{n+1},\theta_{n+1})\}\right|,$$

where $\le$ refers to the lexicographic order (so that $(y_i, \theta_i) \le (y_{n+1}, \theta_{n+1})$ means that either $y_i < y_{n+1}$ or both $y_i = y_{n+1}$ and $\theta_i \le \theta_{n+1}$). If $N = 0$, set, e.g.,

$$A\big((z_1, \theta_1), \dots, (z_n, \theta_n), (z_{n+1}, \theta_{n+1})\big) := \begin{cases} 1 & \text{if } y_{n+1} \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

Since the histogram conformity measure is monotonic and satisfies all other conditions of Lemma 12 (where now both inf and sup are always attained as 0 and 1, respectively), the corresponding conformal transducer is a conformal predictive system. In the next subsection we will show that it is universal.

4.3 Proof of Theorem 31

The proof in this subsection is an elaboration of the proof of Theorem 24 in Subsection 3.3. The difference is that now we have a different definition of Q_n. It suffices to show that (19) holds with probability at least $1-\epsilon$ for large enough n, where $\epsilon > 0$ is a given (arbitrarily small) positive constant. In view of (20), it suffices to prove that

$$\left| \int f \mathrm{d}Q_n - \frac{1}{N} \sum_{i \in I} f(y_i) \right| \le \epsilon/2 \tag{22}$$

holds with probability at least $1 - \epsilon/2$ for large enough n. In this subsection we are using the notation introduced in Subsection 3.3, such as N and I.

On two occasions we will use the following version of Markov's inequality applicable to any probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Lemma 34. *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ and $E \in \mathcal{F}$ be an event. For any positive constants δ_1 and δ_2, if $\mathbb{P}(E) \ge 1 - \delta_1\delta_2$, then $\mathbb{P}(E \mid \mathcal{G}) > 1 - \delta_1$ with probability at least $1 - \delta_2$.*

Proof. Assuming $\mathbb{P}(E) \ge 1 - \delta_1\delta_2$,

$$\begin{aligned} \mathbb{P}\Big(\mathbb{P}(E \mid \mathcal{G}) \le 1 - \delta_1\Big) &= \mathbb{P}\Big(\mathbb{P}(E^c \mid \mathcal{G}) \ge \delta_1\Big) \\ &\le \frac{\mathbb{E}\left(\mathbb{P}(E^c \mid \mathcal{G})\right)}{\delta_1} = \frac{\mathbb{P}(E^c)}{\delta_1} \le \frac{\delta_1\delta_2}{\delta_1} = \delta_2, \end{aligned}$$

where E^c is the complement of E and the first inequality in the chain is a special case of Markov's. □

Set $C := \sup |f| \vee 10$. Remember that $\epsilon > 0$ is a given positive constant. Let B be so large that $y \in [-B, B]$ with probability at least $1 - 0.001\epsilon^2/C$ when $(x, y) \sim P$. This is the first corollary of Lemma 34 that we will need:

Lemma 35. *For a large enough n, the probability (over the choice of $z_1, \dots, z_n, x_{n+1}$) of the fraction of y_i, $i \in I$, satisfying $y_i \in [-B, B]$ to be more than $1 - 0.02\epsilon/C$ is at least $1 - 0.11\epsilon$.*

Proof. By Lemma 34 we have

$$\mathbb{P}\left(\mathbb{P}(y \in [-B, B] \mid x \in \mathcal{P}_n(x')) > 1 - 0.01\epsilon/C\right) \geq 1 - 0.1\epsilon, \tag{23}$$

where the inner $\mathbb{P}$ is over $(x, y) \sim P$ and the outer $\mathbb{P}$ is over $x' \sim P_{\mathbf{X}}$, $P_{\mathbf{X}}$ being the marginal distribution of P on the object space $\mathbf{X}$. To obtain the statement of the lemma it suffices to combine (23) with the law of large numbers. □

Since f is uniformly continuous over $[-B, B]$, there is a partition

$$-B = y_0^* < y_1^* < \cdots < y_m^* < y_{m+1}^* = B$$

of the interval $[-B, B]$ such that

$$\max_{y \in [y_j^*, y_{j+1}^*]} f(y) - \min_{y \in [y_j^*, y_{j+1}^*]} f(y) \leq 0.01\epsilon \tag{24}$$

for $j = 0, 1, \ldots, m$. Without loss of generality we will assume that $y \in \{y_0^*, \ldots, y_{m+1}^*\}$ with probability zero when $(x, y) \sim P$. We will also assume, without loss of generality, that $m > 10$.

Along with the conformal predictive distribution Q_n we will consider the empirical distribution function Q_n^* of the multiset $\{y_i \mid i \in I\}$ (as defined in Subsection 3.4, where it was denoted Q_n); it exists only when $N > 0$. The next lemma will show that Q_n is typically close to Q_n^*. Let K be an arbitrarily large positive integer.

Lemma 36. *For sufficiently large n, $Q_n(y_j^*)$ and $Q_n^*(y_j^*)$ (both exist and) differ from each other by at most $1/K + 0.11\epsilon/C(m+1) + 1/n$ for all $j = 0, 1, \ldots, m+1$ with probability (over the choice of $z_1, \ldots, z_n, x_{n+1}$ and random numbers $\tau, \theta_1, \ldots, \theta_{n+1}$) at least $1 - 0.11\epsilon$.*

Proof. We can choose n so large that $N \geq K$ with probability at least $1 - 0.01\epsilon^2/C(m+1)(m+2)$. By Lemma 34, for such n the conditional probability that $N \geq K$ given $x_1, \ldots, x_n$ is at least $1 - 0.1\epsilon/C(m+1)$ with probability (over the choice of $x_1, \ldots, x_n$) at least $1 - 0.1\epsilon/(m+2)$. Moreover, we can choose n so large that the fraction of x_i, $i = 1, \ldots, n$, which have at least $K - 1$ other x_i, $i = 1, \ldots, n$, in the same cell of $\mathcal{P}_n$ is at least $1 - 0.11\epsilon/C(m+1)$ with probability at least $1 - 0.11\epsilon/(m+2)$ (indeed, we can choose n satisfying the condition in the previous sentence and generate sufficiently many new observations).

Let us fix $j \in \{0, 1, \ldots, m+1\}$. We will show that, for sufficiently large n, $Q_n(y_j^*)$ and $Q_n^*(y_j^*)$ differ from each other by at most $1/K + 0.11\epsilon/C(m+1) + 1/n$ with probability at least $1 - 0.11\epsilon/(m+2)$. We will only consider the case $N > 0$; we will be able to do so since the probability that $N = 0$ tends to 0 as $n \to \infty$. The conformity score of the extended test observation $(x_{n+1}, y_j^*, \theta_{n+1})$ with the postulated label y_j^* is, almost surely, a/N, where a is the number of observations among (x_i, y_i), $i \in I$, satisfying $y_i \leq y_j^*$. (We could have written $y_i < y_j^*$ since we assumed earlier that $y = y_j^*$ with probability zero.) If a cell of $\mathcal{P}_n$ contains at least K elements of the multiset $\{x_1, \ldots, x_n\}$, the percentage of elements of this cell with conformity score less than a/N is, almost surely,

between $a/N-1/K$ and $a/N+1/K$; this remains true if "less than" is replaced by "at most". (It is here that we are using the fact that our conformity measure is randomized and, therefore, conformity scores are tied with probability zero.) And at most a fraction of $0.11\epsilon/C(m+1)$ of elements of the multiset $\{x_1,\dots,x_n\}$ are not in such a cell, with probability at least $1-0.11\epsilon/(m+2)$. Therefore, the overall percentage of elements of the multiset $\{x_1,\dots,x_n\}$ with conformity score less than a/N is between $a/N-1/K-0.11\epsilon/C(m+1)$ and $a/N+1/K+0.11\epsilon/C(m+1)$, with probability at least $1-0.11\epsilon/(m+2)$; this remains true if "less than" is replaced by "at most". Comparing this with the definition (5), we can see that $Q_n(y_j^*)$ is between $a/N-1/K-0.11\epsilon/C(m+1)-1/n$ and $a/N+1/K+0.11\epsilon/C(m+1)+1/n$, with probability at least $1-0.11\epsilon/(m+2)$. It remains to notice that $Q_n^*(y_j^*)=a/N$ almost surely. □

Now we are ready to complete the proof of the theorem. For sufficiently large n, we can transform the left-hand side of (22) as follows (as explained later):

$$\left|\int f\mathrm{d}Q_n-\frac{1}{N}\sum_{i\in I}f(y_i)\right|=\left|\int f\mathrm{d}Q_n-\int f\mathrm{d}Q_n^*\right| \tag{25}$$

$$\le\left|\int_{(-B,B]}f\mathrm{d}Q_n-\int_{(-B,B]}f\mathrm{d}Q_n^*\right| \tag{26}$$

$$+C\big(Q_n^*(-B)+1-Q_n^*(B)+Q_n(-B)+1-Q_n(B)\big)$$

$$\le\left|\sum_{i=0}^{m}f(y_i^*)\left(Q_n(y_{i+1}^*)-Q_n(y_i^*)\right)-\sum_{i=0}^{m}f(y_i^*)\left(Q_n^*(y_{i+1}^*)-Q_n^*(y_i^*)\right)\right| \tag{27}$$

$$+0.02\epsilon+C\left(0.08\frac{\epsilon}{C}+\frac{2}{K}+\frac{0.22\epsilon}{C(m+1)}+\frac{2}{n}\right)$$

$$\le\sum_{i=0}^{m}|f(y_i^*)|\left|Q_n(y_{i+1}^*)-Q_n^*(y_{i+1}^*)-Q_n(y_i^*)+Q_n^*(y_i^*)\right|+0.2\epsilon \tag{28}$$

$$\le\sum_{i=0}^{m}|f(y_i^*)|\left(\frac{2}{K}+\frac{0.22\epsilon}{C(m+1)}+\frac{2}{n}\right)+0.2\epsilon \tag{29}$$

$$\le\frac{2C(m+1)}{K}+0.42\epsilon+\frac{2C(m+1)}{n}\le 0.5\epsilon. \tag{30}$$

The inequality in line (26) holds always. The inequality in line (27) holds with probability (over the choice of $z_1,\dots,z_n$, x_{n+1}, and random numbers τ and $\theta_1,\dots,\theta_{n+1}$) at least $1-0.11\epsilon-0.11\epsilon=1-0.22\epsilon$ by (24) and Lemmas 35 and 36: the addend 0.02ϵ arises by (24) from replacing integrals by sums, the addend $0.08\epsilon/C$ is four times the upper bound on $Q_n^*(-B)$, or $1-Q_n^*(-B)$, given by Lemma 35 (the factor of four arises from bounding $Q_n^*(-B)$, $1-Q_n^*(-B)$, $Q_n(-B)$, and $1-Q_n(-B)$), and the expression $2/K+0.22\epsilon/C(m+1)+2/n$ arises from applying Lemma 36 to reduce bounding $Q_n(-B)$ and $1-Q_n(-B)$ to bounding $Q_n^*(-B)$ and $1-Q_n^*(-B)$, respectively. The inequality in line (28)

holds for sufficiently large K and n. The inequality in line (29) holds with probability at least $1 - 0.11\epsilon$ by Lemma 36, but this probability has already been accounted for. And finally, the second inequality in line (30) holds for sufficiently large K and n. Therefore, the whole chain (25)–(30) holds with probability at least $1 - 0.22\epsilon \geq 1 - \epsilon/2$. This proves (22), which completes the overall proof.

To avoid any ambiguity, this paragraph will summarize the roles of ϵ, B, m, K, and n in this proof. First we fix a positive constant $\epsilon > 0$ (which, however, can be arbitrarily small). Next we choose B, sufficiently large for the given ϵ, and after that, a sufficiently fine partition of $[-B, B]$ of size m. We then choose K, which should be sufficiently large for the given ϵ and partition. Finally, we choose n, which should be sufficiently large for the given ϵ, partition, and K.

5 Conclusion

This paper constructs a hierarchy of universally consistent predictive systems, where the notion of universal consistency is based on Belyaev's notion of weakly approaching sequences of distributions. The most basic one is the universal probability forecasting system (Theorem 26). It does not satisfy any small-sample properties of validity, and its construction is standard and very simple. The next level in the hierarchy is occupied by the universal Mondrian predictive system (Theorem 24). Mondrian predictive systems are automatically probabilistically calibrated, but their definition is more complicated than that of conformal predictive systems. The main result of this paper is the construction of a universal conformal predictive system. Perhaps the most important potential applications of universal predictive systems are in decision making, where minimizing the expected loss under predictive distributions output by a universal predictive system provides a natural decision strategy.

There are many interesting directions of further research. These are the most obvious ones:

1. Investigate the best rate at which conformal predictive distributions and the true conditional distributions can approach each other.

2. Replace universal consistency by strong universal consistency (i.e., convergence in probability by convergence almost surely), perhaps in the online prediction protocol (as in Remark 10).

3. Construct more natural, and perhaps even practically useful, universal randomized predictive systems.

Acknowledgments

Many thanks to the COPA 2019 and *Pattern Recognition* reviewers for their thoughtful comments and to Amazon, Astra Zeneca, and Stena Line for their support.

References

[1] Thomas Augustin and Frank P. A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124:251–272, 2004.

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

[3] Yuri Belyaev. Bootstrap, resampling, and Mallows metric. Technical report, Department of Mathematical Statistics, Umeå University, Sweden, 1995.

[4] Yuri Belyaev and Sara Sjöstedt–de Luna. Weakly approaching sequences of random distributions. *Journal of Applied Probability*, 37:807–822, 2000.

[5] Claude Dellacherie and Paul-André Meyer. *Probabilités et potentiel*. Hermann, Paris, 1975. Chapters I–IV.

[6] Arthur P. Dempster. On direct probabilities. *Journal of the Royal Statistical Society B*, 25:100–110, 1963.

[7] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[8] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B*, 69:243–268, 2007.

[9] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

[10] Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013.

[11] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

[12] Bruce M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63:677–691, 1968.

[13] Bruce M. Hill. De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In Dennis V. Lindley, José M. Bernardo, Morris H. DeGroot, and Adrian F. M. Smith, editors, *Bayesian Statistics 3*, pages 211–241. Oxford University Press, Oxford, 1988.

[14] Alexander S. Kechris. *Classical Descriptive Set Theory*. Springer, New York, 1995.

[15] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108:278–287, 2013.

[16] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society B*, 76:71–96, 2014.

[17] Ryan Martin and Chuanhai Liu. *Inferential Models: Reasoning with Uncertainty*. CRC Press, Boca Raton, FL, 2016.

[18] Allan H. Murphy and Robert L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338, 1987.

[19] Tore Schweder and Nils L. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge, 2016.

[20] Glenn Shafer. Bayesian, fiducial, frequentist. The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper 50, December 2017.

[21] 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.

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

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

[24] Sara Sjöstedt–de Luna. Some properties of weakly approaching sequences of distributions. *Statistics and Probability Letters*, 75:119–126, 2005.

[25] Charles J. Stone. Consistent nonparametric regression (with discussion). *Annals of Statistics*, 5:595–645, 1977.

[26] Vladimir Vovk. Universally consistent conformal predictive distributions. *Proceedings of Machine Learning Research*, 105:105–122, 2019. COPA 2019.

[27] Vladimir Vovk and Claus Bendtsen. Conformal predictive decision making. *Proceedings of Machine Learning Research*, 91:52–62, 2018. COPA 2018.

[28] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

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

[30] Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman. Conformal predictive distributions with kernels. In Lev Rozonoer, Boris Mirkin, and Ilya Muchnik, editors, *Braverman's Readings in Machine Learning: Key Ideas from Inception to Current State*, volume 11100, pages 103–121. Springer, Cham, Switzerland, 2018.

[31] Vladimir Vovk and Ivan Petej. Venn–Abers predictors. In Nevin L. Zhang and Jin Tian, editors, *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 829–838, Corvallis, OR, 2014. AUAI Press.

[32] 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.

[33] Minge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81:3–39, 2013.

A An implication of universal consistency

The definition of universal consistency is given in terms of the notion of consistency (4), which means, intuitively, that the predictive distribution and the true conditional distribution of the label of the test object approach each other. The notion of approaching given by (4) is in the spirit of the standard notion of weak convergence but it is not really explicit; in particular, it is not clear whether a natural distance between the corresponding distribution functions tends to zero. In this appendix we will see that this is indeed the case for Lévy distance.

It will be convenient to represent the data-generating probability measure P on the observation space $\mathbf{Z} = \mathbf{X} \times \mathbb{R}$ as the combination of the marginal distribution $P_{\mathbf{X}}$ on $\mathbf{X}$, $P_{\mathbf{X}}(E) := P(E \times \mathbb{R})$ for the measurable $E \subseteq \mathbf{X}$, and a regular conditional distribution of $y \in \mathbb{R}$ given $x \in \mathbf{X}$. Let us fix a regular distribution function $F_P(y \mid x)$ of y given x for $(x, y) \sim P$. Its defining property is that, for each $x \in \mathbf{X}$, $F_P(\cdot \mid x)$ is a distribution function on $\mathbb{R}$, and for each $u \in \mathbb{R}$, $F_P(u \mid x)$ is a version of the conditional probability that $y \le u$ given x when $(x, y) \sim P$. (It exists by, e.g., Theorem 2.7.4 in [22], in combination with Doob's theorem [5, I.18].) Intuitively, $(x, y) \sim P$ can be thought of as generated in two steps: first $x \sim P_{\mathbf{X}}$ and then $y \sim F_P(\cdot \mid x)$.

Definition 37. A *DF-type random function* is a function Q on a sample space whose values are increasing functions $y \in \mathbb{R} \mapsto Q(y) \in [0, 1]$ and which is measurable in the sense of $Q(y)$ being measurable for each $y \in \mathbb{R}$.

Definition 37 is a generalization of the standard notion of a CDF-valued random quantity [9, Section 2.1], in which the random functions $y \mapsto Q(y)$ are required, additionally, to be right-continuous and satisfy $\lim_{y \to -\infty} Q(y) = 0$ and $\lim_{y \to \infty} Q(y) = 1$.

Definition 38. A sequence of DF-type random functions Q_n *weakly approaches* (in probability) another sequence of DF-type random functions Q'_n if, for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\left| \int f \mathrm{d}\bar{Q}_n - \int f \mathrm{d}\bar{Q}'_n \right| \to 0 \qquad (n \to \infty) \tag{31}$$

in probability, where $\bar{Q}$, for an increasing function $Q : \mathbb{R} \to [0,1]$, is defined at the end of Subsection 2.1. This relation between Q_n and Q'_n is symmetric (in fact, an equivalence relation) and denoted $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$.

Definition 4 can be restated in terms of Definition 38, namely: a randomized predictive system Q is consistent for P if and only if the sequences of random functions Q_n and $F_P(\cdot \mid x_n)$ weakly approach each other assuming $z_i = (x_i, y_i) \sim P$, $i = 1, \dots, n+1$, and $\tau \sim U$ are all independent.

Remark 39. Definition 38 essentially follows Belyaev [3, 4]. It is a version of Belyaev and Sjöstedt–de Luna's [4] Definition 3. Their Definition 1 assumes that both Q_n and Q'_n are deterministic, and their Definition 2 assumes that one of them is deterministic; all mathematical results in [4] use either Definition 1 or Definition 2.

The notion of tightness is as important in the theory of weakly approaching sequences of distribution functions (or probability measures) as it is in the standard theory of weak convergence (e.g., it is used constantly in [4] and [24]). The following definition is modelled on [4, Definition 4].

Definition 40. A sequence of DF-type random functions Q_n is *tight (in probability)* if, for any $\epsilon > 0$ and $\delta > 0$, there is $C \in (0, \infty)$ such that

$$\liminf_{n \to \infty} \mathbb{P}\Big(\bar{Q}_n([-C, C]) \geq 1 - \epsilon\Big) \geq 1 - \delta. \tag{32}$$

Notice that a tight sequence Q_n always satisfies $\bar{Q}_n(\mathbb{R}) \to 1$ in probability as $n \to \infty$.

The next two lemmas show that the sequences of DF-type random functions considered in this paper are tight.

Lemma 41. *The sequence of random functions $F_P(\cdot \mid x_n)$ is tight, assuming $z_i = (x_i, y_i) \sim P$, $i = 1, 2, \dots$, are all independent.*

Proof. In the current IID case the requirement of being tight, as applied to the sequence of random functions $F_P(\cdot \mid x_n)$, can be rewritten as: for any $\epsilon > 0$ and $\delta > 0$, there is $C \in [0, \infty)$ such that

$$\mathbb{P}\Big(F_P([-C, C] \mid x) \geq 1 - \epsilon\Big) \geq 1 - \delta$$

(cf. (32)), where $\mathbb{P}$ refers to $x \sim P_{\mathbf{X}}$. It remains to use the σ-additivity of the probability measure $P_{\mathbf{X}}$. □

The following lemma generalizes [4, Lemma 5].

Lemma 42. *If $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$ and the sequence Q_n is tight, Q'_n is also tight.*

Proof. Suppose $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$, Q_n is tight, but Q'_n is not tight; we will arrive at a contradiction.

There exist $\epsilon > 0$ and $\delta > 0$ such that, for any $C \in [0, \infty)$, there are infinitely many n such that

$$\mathbb{P}\Big(\bar{Q}'_n([-C, C]) \geq 1 - \epsilon\Big) < 1 - \delta. \tag{33}$$

Fix such ϵ and δ. Fix $C \in (1, \infty)$ such that, from some n on,

$$\mathbb{P}\Big(\bar{Q}_n([-C+1, C-1]) \geq 1 - \epsilon/2\Big) \geq 1 - \delta/2. \tag{34}$$

Let $f : \mathbb{R} \to [0, 1]$ be a continuous function satisfying

$$f(u) = \begin{cases} 0 & \text{if } |u| \leq C - 1 \\ 1 & \text{if } |u| \geq C \end{cases}$$

(there are no restrictions for $|u| \in (C-1, C)$, apart from f being continuous and taking values in $[0, 1]$). Passing to a subsequence, we assume, without loss of generality, that (33) holds for all n. By (33) and (34), we have

$$\mathbb{P}\left(\int f \mathrm{d}\bar{Q}'_n > 0.9\epsilon\right) > 0.9\delta \tag{35}$$

$$\mathbb{P}\left(\int f \mathrm{d}\bar{Q}_n \leq \epsilon/2\right) \geq 1 - \delta/2, \tag{36}$$

respectively, both inequalities holding from some n on; (35) also uses $Q'_n(\mathbb{R}) \to 1$ (in probability), which follows from $Q_n(\mathbb{R}) \to 1$ and $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$ (specialized to $f := 1$). By the inequality $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$, (35) and (36) give

$$\mathbb{P}\left(\int f \mathrm{d}\bar{Q}'_n - \int f \mathrm{d}\bar{Q}_n > 0.4\epsilon\right) > 0.4\delta$$

from some n on, which contradicts $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$. □

The following lemma says that for tight sequences of DF-type random functions the notion of weakly approaching sequences agrees with a standard notion of closeness of distribution functions, Lévy distance; it generalizes the necessity half of [4, Lemma 10]. Let us extend Lévy distance to arbitrary increasing functions $Q, Q' : \mathbb{R} \to [0, 1]$:

$$\rho_L(Q, Q') := \inf\left\{h > 0 \mid \forall u \in \mathbb{R} : Q(u - h) - h \leq Q'(x) \leq Q(u + h) + h\right\}.$$

Lemma 43. *If $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$ and the sequence Q_n is tight, the Lévy distance between Q_n and Q'_n tends to 0 in probability.*

Proof. Suppose $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$, the sequences Q_n and Q'_n are tight, but the convergence $\rho_L(Q_n, Q'_n) \to 0$ in probability fails. Our goal is to arrive at a contradiction (cf. Lemma 42).

Fix $\epsilon > 0$ and $\delta > 0$ such that $\rho_L(Q_n, Q'_n) > \epsilon$ with probability more than δ for infinitely many n. Passing to a subsequence, we assume, without loss of generality, that

$$\mathbb{P}\left(\rho_L(Q_n, Q'_n) > \epsilon\right) > \delta \tag{37}$$

for all n. As both Q_n and Q'_n are tight, there is $C \in (0, \infty)$ such that, from some n on,

$$\mathbb{P}\left(Q_n([-C, C]) \geq 1 - \frac{\epsilon}{2}\right) \geq 1 - \frac{\delta}{4} \tag{38}$$

and

$$\mathbb{P}\left(Q'_n([-C, C]) \geq 1 - \frac{\epsilon}{2}\right) \geq 1 - \frac{\delta}{4}. \tag{39}$$

The conjunction of (37)–(39) implies

$$\mathbb{P}\Big(\exists u \in [-C, C - \epsilon] : Q'_n(u) > Q_n(u + \epsilon) + \epsilon \\ \text{or } Q_n(u) > Q'_n(u + \epsilon) + \epsilon\Big) > \frac{\delta}{2} \tag{40}$$

from some n on (this follows from the inequality $\mathbb{P}(A \cap B \cap C) \geq \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - 2$ combined with the conjunction of the events under the probability sign in (37)–(39) implying the event under the probability sign in (40)). Inequality (40) implies that there exists a positive integer N such that

$$\mathbb{P}\Bigg(\exists i \in \{1, \ldots, N\} : Q'_n\left(-C + 2C\frac{i-1}{N}\right) > Q_n\left(-C + 2C\frac{i}{N}\right) + \frac{2C}{N} \\ \text{or } Q_n\left(-C + 2C\frac{i-1}{N}\right) > Q'_n\left(-C + 2C\frac{i}{N}\right) + \frac{2C}{N}\Bigg) > \frac{\delta}{2} \tag{41}$$

from some n on. Fix such an N.

On the other hand, $Q_n \overset{\text{wa}}{\longleftrightarrow} Q'_n$ implies

$$\limsup_{n\to\infty}\left(\int f \mathrm{d}\bar{Q}_n - \int f \mathrm{d}\bar{Q}'_n\right) \leq 0 \tag{42}$$

(cf. (31)), which, when applied to a continuous function $f : \mathbb{R} \to [0, 1]$ satisfying

$$f(u) = \begin{cases} 1 & \text{if } u \leq -C + 2C\frac{i-1}{N} \\ 0 & \text{if } u \geq -C + 2C\frac{i}{N} \end{cases}$$

for a given $i \in \{1, \ldots, N\}$ and in conjunction with $\bar{Q}_n(\mathbb{R}) \to 1$, implies

$$\limsup_{n\to\infty}\left(\bar{Q}_n\left(-C + 2C\frac{i-1}{N}\right) - \bar{Q}'_n\left(-C + 2C\frac{i}{N}\right)\right) \leq 0, \tag{43}$$

where we continue to use the same notation $\bar{Q}_n$ and $\bar{Q}'_n$ for the distribution functions of the probability measures $\bar{Q}_n$ and $\bar{Q}'_n$. Both (42) and (43) hold in probability, and the latter implies that

$$\mathbb{P}\left(\bar{Q}_n\left(-C+2C\frac{i-1}{N}\right)-\bar{Q}'_n\left(-C+2C\frac{i}{N}\right)\le\frac{2C}{N}\right)\ge 1-\frac{\delta}{4N} \tag{44}$$

from some n on. Combining the inequality (44) for $i=1,\dots,N$ and the same inequalities with $\bar{Q}_n$ and $\bar{Q}'_n$ swapped (therefore, combining $2N$ inequalities in total) gives the negation of (41). This contradiction completes the proof of the lemma. □

Combining Theorem 31 with Lemmas 41 and 43, we can see that the Lévy distance between the predictive distribution output by a universal conformal predictive system and the true conditional distribution of the label of the test object indeed converges to zero.

B Marginal calibration

The main notion of validity (R2 in Definition 2) used in this paper is, in the terminology of [9, Definition 3(b)], being probabilistically calibrated. This property is generally regarded to be the most important of several calibration properties considered in probability forecasting. The following definition gives another popular calibration property [9, Definition 3(a)] as applied to conformal predictive systems (of course, this property is applicable in a much wider context). The number of training observations will be referred to as the *sample size*.

Definition 44. A conformal predictive system is *marginally calibrated* for a sample size n and a probability measure P on $\mathbf{Z}^{n+1}$ if, for any $y\in\mathbb{R}$,

$$\mathbb{E}\,Q(z_1,\dots,z_n,(x_{n+1},y),\tau)=\mathbb{P}(y_{n+1}\le y), \tag{45}$$

where both $\mathbb{E}$ and $\mathbb{P}$ are over $(z_1,\dots,z_n,(x_{n+1},y_{n+1}),\tau)\sim P\times U$.

In this appendix we will see that conformal predictive systems are not always marginally calibrated under the IID model. But we will start from an easier statement.

The probabilistic calibration property R2 for a given sample size n depends only on the observations $(z_1,\dots,z_{n+1})$ being generated from an exchangeable distribution on $\mathbf{Z}^{n+1}$ (and $\tau\sim U$ independently): see Remark 16 or, e.g., [28, Theorem 8.1]. The following example shows that there are conformal predictive systems that are not marginally calibrated for some sample size n and an exchangeable probability measure on $\mathbf{Z}^{n+1}$, even among conformal predictive systems corresponding to conformity measures satisfying the conditions of Lemma 12.

Example 45. Set $n:=1$, suppose $|\mathbf{X}|>1$, and let the data be generated from the exchangeable probability measure P on $\mathbf{Z}^2$ that assigns equal weights $1/2$

to the sequences $((x^{-1},-1),(x^1,1))$ and $((x^1,1),(x^{-1},-1))$ in $\mathbf{Z}^2$, where x^{-1} and x^1 are two distinct elements of $\mathbf{X}$ (fixed for the rest of this appendix). Let a conformity measure A satisfy

$$A((x_1,y_1),(x_2,y_2)) = \begin{cases} y_2 & \text{if } x_2 = x^1 \\ 3y_2 + 2 & \text{if } x_2 = x^{-1}; \end{cases} \tag{46}$$

it is clear that A can be extended to the whole of $\mathbf{X}$ and to all sample sizes n in such a way that it satisfies all conditions in Lemma 12. For $y = 0$, the right-hand side of (45) is $1/2$, whereas the left-hand side is different, namely $3/4$:

- with probability $1/2$ the training and test data form the sequence $((x^{-1},-1),(x^1,1))$, and so the conformal predictive distribution is

$$Q((x^{-1},-1),(x^1,y)) = \begin{cases} [0,1/2] & \text{if } y < -1 \\ [0,1] & \text{if } y = -1 \\ [1/2,1] & \text{if } y > -1; \end{cases} \tag{47}$$

  the position $y = -1$ of the jump is found from the condition $\alpha_1^y = \alpha_2^y$, i.e., $3 \times (-1) + 2 = y$;

- with probability $1/2$ the training and test data form the sequence $((x^1,1),(x^{-1},-1))$, and so the conformal predictive distribution is

$$Q((x^1,1),(x^{-1},y)) = \begin{cases} [0,1/2] & \text{if } y < -1/3 \\ [0,1] & \text{if } y = -1/3 \\ [1/2,1] & \text{if } y > -1/3; \end{cases} \tag{48}$$

  the position $y = -1/3$ of the jump is found from the same condition $\alpha_1^y = \alpha_2^y$, which becomes $1 = 3y + 2$;

- therefore, the mean value of the conformal predictive distribution at $y = 0$ is $3/4$.

Example 45 can be strengthened by replacing the assumption of exchangeability by the IID model.

Example 46. Now we assume that the two observations are generated independently from the probability measure on $\mathbf{Z}$ assigning equal weights $1/2$ to the observations $(x^{-1},-1)$ and $(x^1,1)$. We consider the same conformity measure as in Example 45: see (46). For $y = 0$, the right-hand side of (45) is still $1/2$, and the left-hand side becomes $5/8$:

- with probability $1/4$ the training and test data form the sequence $((x^{-1},-1),(x^1,1))$, and so the conformal predictive distribution is (47), averaging $3/4$ at $y = 0$;

- with probability 1/4 the training and test data form the sequence $((x^1,1),(x^{-1},-1))$, and so the conformal predictive distribution is (48), averaging 3/4 at $y=0$;
- with probability 1/4 the training and test data form the sequence $((x^{-1},-1),(x^{-1},-1))$, and so the conformal predictive distribution is

$$Q((x^{-1},-1),(x^{-1},y)) = \begin{cases} [0,1/2] & \text{if } y < -1 \\ [0,1] & \text{if } y = -1 \\ [1/2,1] & \text{if } y > -1, \end{cases}$$

  which is 3/4 on average at $y=0$; the position $y=-1$ of the jump is found from the condition $\alpha_1^y = \alpha_2^y$, which now is $3\times(-1)+2 = 3y+2$;
- finally, with probability 1/4 the training and test data form the sequence $((x^1,1),(x^1,1))$, and so the conformal predictive distribution is

$$Q((x^1,1),(x^1,y)) = \begin{cases} [0,1/2] & \text{if } y < 1 \\ [0,1] & \text{if } y = 1 \\ [1/2,1] & \text{if } y > 1, \end{cases}$$

  which is 1/4 on average at $y=0$; the position $y=1$ of the jump is found from the condition $\alpha_1^y = \alpha_2^y$, which now is $1 = y$;
- therefore, the mean value of the conformal predictive distribution at $y=0$ is 5/8.

C Venn predictors

In [32] and this paper, conformal prediction is adapted to probability forecasting. An older method of probability forecasting enjoying properties of validity similar to those of conformal prediction is Venn prediction [28, Chapter 6]. This appendix reviews Venn prediction and its properties of validity. We fix the sample size n.

Definition 47. Let κ be a taxonomy (as defined in Definition 17). The *Venn predictor* corresponding to κ is the family $\{Q_u \mid u \in \mathbb{R}\}$ of distribution functions defined by

$$Q_u(z_1,\dots,z_n,(x_{n+1},y)) := \frac{|\{i \in \kappa(n+1 \mid z_1,\dots,z_n,(x_{n+1},u)) \mid y_i \le y\}|}{|\kappa(n+1 \mid z_1,\dots,z_n,(x_{n+1},u))|} \tag{49}$$

for any training data $(z_1,\dots,z_n)$ and test object x_{n+1}.

The definition (49) is similar to, but simpler than, (16). The intuition is that the Venn predictor contains (for $u := y_{n+1}$ being the actual label in the test observation) the true empirical distribution function of the labels in the observations that are similar, in a suitable sense (determined by κ), to the test observation. The Venn prediction (49) is useful when the distribution functions Q_u are close to each other for different u. Whereas this is a reasonable assumption for a suitable choice of κ in the case of classification (such as binary classification, $y_i \in \{0, 1\}$, in [31]), in the case of regression it might make more sense to restrict attention to

$$\{Q_u \mid u \in \Gamma^\epsilon(z_1, \dots, z_n, x_{n+1})\}$$

for a conformal predictor Γ (see, e.g., [28, Section 2.2]) and a small significance level $\epsilon > 0$.

The following theorem shows that Venn predictors are ideal in the technical sense of [10, Definition 2.2] (and independent work by Tsyplakov).

Theorem 48. *Let $\mathcal{G}$ be the σ-algebra on $\mathbf{Z}^{n+1}$ consisting of the measurable subsets E of $\mathbf{Z}^{n+1}$ that are predictably invariant with respect to the taxonomy κ in the following sense: if a permutation π of $\{1, \dots, n+1\}$ respects the equivalence relation $\sim$ assigned by κ to $(z_1, \dots, z_{n+1})$ (in the sense of Definition 18) and leaves $n+1$ in the same equivalence class, then*

$$(z_1, \dots, z_{n+1}) \in E \Longrightarrow (z_{\pi(1)}, \dots, z_{\pi(n+1)}) \in E.$$

For any $y \in \mathbb{R}$,

$$Q_{y_{n+1}}(z_1, \dots, z_n, (x_{n+1}, y)) = \mathbb{P}(y_{n+1} \le y \mid \mathcal{G}), \tag{50}$$

where $(z_1, \dots, z_n, (x_{n+1}, y_{n+1}))$ are generated from an exchangeable probability distribution on $\mathbf{Z}^{n+1}$.

Equation (50) expresses the condition of being *ideal*, with respect to some information base (namely, the σ-algebra $\mathcal{G}$). According to [10, Theorem 2.8], this means that Venn predictors are both marginally and probabilistically calibrated, in the sense of one of their component distribution functions, namely $Q_{y_{n+1}}(z_1, \dots, z_n, (x_{n+1}, \cdot))$, being such. And according to [10, Theorem 2.11], in the case of the binary label taking values in $\{0, 1\}$, being probabilistically calibrated is equivalent to being *conditionally calibrated*

$$\mathbb{P}(y_{n+1} = 1 \mid p_{n+1}) = p_{n+1}, \tag{51}$$

where $p_{n+1} := 1 - Q_{y_{n+1}}(z_1, \dots, z_n, (x_{n+1}, 0))$ is the predicted probability that $y_{n+1} = 1$. Equation (51) for Venn predictors, in the case of binary classification, is Theorem 1 in [31].

Proof of Theorem 48. Fix $y \in \mathbb{R}$. Let P be the data-generating distribution (an exchangeable probability distribution on $\mathbf{Z}^{n+1}$), Q be the random variable

$Q_{y_{n+1}}(z_1, \dots, z_n, (x_{n+1}, y))$, and $E \in \mathcal{G}$. Notice that Q is $\mathcal{G}$-measurable. Our goal is to prove

$$\int_E 1_{\{y_{n+1} \le y\}} \mathrm{d}P = \int_E Q \mathrm{d}P, \tag{52}$$

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

There are finitely many equivalence relations on the set $\{1, \dots, n+1\}$. For each of them the set of data sequences $(z_1, \dots, z_{n+1})$ that are assigned this equivalence relation by the taxonomy κ is measurable (by the requirement of measurability in the definition of a taxonomy) and, moreover, is an element of $\mathcal{G}$. Therefore, E can be decomposed into a disjoint union of elements of $\mathcal{G}$ all of whose elements are assigned the same equivalence relation by κ. We will assume, without loss of generality, that all elements of E are assigned the same equivalence relation, which is fixed to the end of this proof. Let $\kappa(j)$ stand for the equivalence class of $j \in \{1, \dots, n+1\}$.

Let us say that two data sequences in E are *similar* if, for any equivalence class $C \subseteq \{1, \dots, n+1\}$, they have the same numbers of observations with indices in C and with labels less than or equal to y. Following the same argument as in the previous paragraph, we further assume that all elements of E are similar.

Now we can see that both sides of (52) are equal to

$$P(E) \frac{|\{i \in \kappa(n+1) \mid y_i \le y\}|}{|\kappa(n+1)|}$$

(cf. (49)). □