Universally consistent conformal predictive distributions

Vladimir Vovk V.VOVK@RHUL.AC.UK
Royal Holloway, University of London, Egham, Surrey, UK

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Evgueni Smirnov

Abstract

This paper describes conformal predictive systems that are universally consistent in the sense of being consistent under any data-generating distribution, assuming that the observations are produced independently in the IID fashion. Being conformal, these predictive systems satisfy a natural property of small-sample validity, namely they are automatically calibrated in probability.

Keywords: Calibration in probability, conformal prediction, predictive distribution, universal consistency.

1. Introduction

Predictive distributions are probability distributions for future labels satisfying a natural property of validity. They were introduced independently by Schweder and Hjort (2016, Chapter 12), and Shen et al. (2018), 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 Gneiting and Katzfuss 2014). First nonparametric predictive distributions were constructed in the conference version of Vovk et al. (2019) based on the method of conformal prediction (see, e.g., Vovk et al. 2005, 2009; Lei et al. 2013; Lei and Wasserman 2014). The nonparametric statistical model used in Vovk et al. (2019) 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, Vovk et al. (2019) slightly generalize 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 Vovk et al. (2019, 2018) 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 Vovk et al. 2019, 2018 were based on the Least Squares method and its modifications, and thus far from universally consistent)

Our approach is in the spirit of Gneiting et al.'s (2007) paradigm (which they trace back to Murphy and Winkler 1987) 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. 2018).

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 23).

As in Vovk et al. (2019, 2018), 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. (2013) in the unsupervised setting and Lei and Wasserman (2014) 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 28) is that there exists a universally consistent 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 (Belyaev, 1995; Belyaev and Sjöstedt–de Luna, 2000; Sjöstedt–de Luna, 2005) notion of weakly approaching sequences of distributions. The construction of a universally consistent conformal predictive system adapts standard arguments for universal consistency in classification and regression (Stone, 1977; Devroye et al., 1996; Györfi et al., 2002).

The importance of universal consistency is demonstrated in Vovk and Bendtsen (2018, Section 5); namely, applying the expected utility maximization principle to the predictive distributions produced by a universally consistent predictive system leads, under natural conditions, to asymptotically optimal decisions.

We start in Section 2 from defining randomized predictive systems, which are required to satisfy the small-sample property of validity under the IID model. The next section, Section 3, defines conformal predictive systems, which are a subclass of randomized predictive systems. The main result of the paper, Theorem 28 stated in Section 9, requires a slight generalization of conformal predictive systems (for which we retain the same name). Section 4 introduces another subclass of randomized predictive systems, which is wider than the subclass of conformal predictive systems of Section 3; the elements of this wider subclass are called Mondrian predictive systems. A simple version of Theorem 28 given in Section 5 (Theorem 21) states the existence of Mondrian predictive systems that are universally consistent. An example of a universally consistent Mondrian predictive system is given in Section 6, and Section 7 is devoted to a short proof that this predictive system is indeed universally consistent. Section 8 gives an even shorter proof of the existence of a universally consistent probability forecasting system (Theorem 23), which is deterministic and not required to satisfy any small-sample properties of validity. Theorem 28 stated in

Section 9 asserts the existence of universally consistent conformal predictive systems. An example of such a conformal predictive system is given in Section 10, and it is shown in Section 11 to be universally consistent. One advantage of Theorem 28 over the result of Section 5 (Theorem 21) is that, as compared with Mondrian predictive systems, conformal predictive systems enjoy a stronger small-sample property of validity (see Remarks 9 and 20). In conclusion, Section 12 lists some natural directions of further research.

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., Xie and Singh (2013). 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, Vovk et al. (2019, 2018) and this paper argue that there is a meaningful theory of predictive distributions even under the IID model.

2. Randomized predictive distributions

In this section we give some basic definitions partly following Shen et al. (2018) and Vovk et al. (2019). 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* $y \in \mathbb{R}$. 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 A measurable function $Q : \cup_{n=1}^{\infty}(\mathbf{Z}^{n+1} \times [0, 1]) \to [0, 1]$ is called 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 $Q(z_1, \dots, z_n, (x_{n+1}, y), \tau)$ is monotonically increasing in both y and τ (i.e., monotonically increasing in y for each τ and 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 $x_{n+1} \in \mathbf{X}$,

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

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:

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

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 very little training data; see Remark 12 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 (2014) refers to it as probabilistic calibration and describes it as critical in forecasting; Gneiting and Katzfuss (2014, Section 2.2.3) review the relevant literature.

Remark 3 Requirements R1 and R2 are the analogues (introduced in Schweder and Hjort 2016, Chapter 12, and Shen et al. 2018) of similar requirements in the theory of confidence distributions: see, e.g., Xie and Singh (2013, Definition 1), or Schweder and Hjort (2016, 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 $z_i=(x_i,y_i)$;
- $\mathbb{E}_P(f\mid x_{n+1})$ 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 in $\mathbb{R}$.

Definition 5 A randomized predictive system Q is *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., Belyaev and Sjöstedt–de Luna 2000). Our goal is construction of universally consistent randomized predictive systems.

3. Conformal predictive distributions

A way of producing randomized predictive distributions under the IID model has been proposed in Vovk et al. (2019). This section reviews a basic version, and Section 9 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: for any n, any sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, any $z_{n+1} \in \mathbf{Z}$, and any permutation π of $\{1, \dots, n\}$,

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

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

Definition 7 The *conformal transducer* corresponding to a conformity measure A is defined as

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

where $n \in \{1, 2, \dots\}$, $(z_1, \dots, z_n) \in \mathbf{Z}^n$ 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

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

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 13 at the end of this section).

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, $Q(z_1, \dots, z_n, (x_{n+1}, \cdot), \tau)$ are the corresponding *conformal predictive distributions* (or, more fully, conformal predictive distribution functions).

Remark 9 Requirement R2 in the previous section is sometimes referred to as the frequentist validity of predictive or confidence distributions (see, e.g., Xie and Singh 2013 and Shen et al. 2018). It can be argued that there is no need to appeal to frequencies in these and similar cases (see, e.g., Shafer 2017). 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 (Vovk et al., 2005, 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 10 A conformity measure A is *monotonic* if $A(z_1,\dots,z_{n+1})$ is:

- monotonically increasing in y_{n+1},

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

- monotonically decreasing in y_1,

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

  (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 11 *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},*

$$\inf_y A(z_1,\dots,z_n,(x_{n+1},y))=\inf A_n, \tag{7}$$

$$\sup_y A(z_1,\dots,z_n,(x_{n+1},y))=\sup A_n; \tag{8}$$

- *for each n, the $\inf_y$ in (7) 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 (8) 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 (7) are allowed to take value $-\infty$, and the two sup in (8) are allowed to take value ∞. The conditions of Lemma 11 will be satisfied if (7) and (8) 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).

Remark 12 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, in the most interesting cases considered in Vovk et al. (2019, 2018) the difference (3) is $1/(n+1)$ except for at most n values of y. A randomized predictive system can be universally consistent only if the difference (3) is small with high probability.

Proof of Lemma 11 We need to check requirements R1 and R2. R2 is the standard property of validity for conformal transducers (see, e.g., Vovk et al. 2005, Theorem 8.1). The intuition behind the proof of this property is given in Remark 13 at the end of this section.

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

$$Q(z_1, \dots, z_n, (x_{n+1}, y), \tau) = \frac{1}{n+1} \sum_{i=1}^{n+1} \left(1_{\{\alpha_i^y < \alpha_{n+1}^y\}} + \tau 1_{\{\alpha_i^y = \alpha_{n+1}^y\}} \right), \tag{9}$$

where $1_{\{E\}}$ stands for the indicator function of a property E, and it suffices to prove that each addend in (9) 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 (7) 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 (7) 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 13 The proof of Lemma 11 refers to Vovk et al. (2005) 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.

4. 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 14 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 15. The idea behind a taxonomy is to determine the comparison class for computing the p-value (5); instead of using all available data we will 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 15 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

$$(i \sim j \mid z_1, \dots, z_{n+1}) \implies (\pi(i) \sim \pi(j) \mid z_{\pi(1)}, \dots, z_{\pi(n+1)}). \tag{10}$$

Remark 16 The notion of taxonomy used in this paper is introduced in Vovk and Petej (2014) under the name of Venn taxonomies and subsumes Mondrian taxonomies as defined in Vovk et al. (2005, Section 4.5), Venn taxonomies as defined in Vovk et al. (2005, Section 6.3), and n-taxonomies as defined in Balasubramanian et al. (2014, Section 2.2). A narrower notion of taxonomy requires that (10) hold for all permutations π of $\{1, \dots, n+1\}$; the taxonomy of Section 6 belongs to this narrower class.

Definition 17 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

$$\begin{aligned} &Q(z_1, \dots, z_n, (x_{n+1}, y), \tau) \\ &\qquad := \frac{\left|\left\{i \in \kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y)) \mid \alpha_i^y < \alpha_{n+1}^y\right\}\right|}{\left|\kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y))\right|} \\ &\qquad\quad + \tau \frac{\left|\left\{i \in \kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y)) \mid \alpha_i^y = \alpha_{n+1}^y\right\}\right|}{\left|\kappa(n+1 \mid z_1, \dots, z_n, (x_{n+1}, y))\right|}, \end{aligned} \tag{11}$$

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 Section 2.

Notice that the denominator in (11) is always positive. The Mondrian p-value (11) 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 Vovk et al. (2005, Fig. 4.3), for the origin of the attribute "Mondrian".

Lemma 18 *If a taxonomy does not depend on the labels and a conformity measure is monotonic and satisfies the three conditions of Lemma 11, the corresponding Mondrian transducer will be a randomized (and, therefore, Mondrian) predictive system.*

Proof As in Lemma 11, 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 (11) does not depend on y, the value of (11) is monotonically increasing in y: it suffices to replace (9) 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 11. In combination with the obvious monotonicity in τ, this proves R1(i). R1(ii) is demonstrated as in Lemma 11. The proof of R2 is standard and valid for any taxonomy (see, e.g., Vovk et al. 2005, Section 8.7); the intuition behind it is given in Remark 19 below. ■

The properties listed in Lemma 18 will be satisfied by the conformity measure and taxonomy defined in Section 6 to prove Theorem 21, a weaker form of the main result of this paper.

Remark 19 Remark 13 can be easily adapted to Mondrian predictive systems. For $\tau := 1$ and assuming no ties among the conformity scores, the right-hand side of (11) 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 (11) being uniformly distributed in $[0,1]$, provided $y := y_{n+1}$ and $\tau \sim U$.

Remark 20 One advantage of conformal predictive systems over Mondrian predictive systems is that the former satisfy a stronger version of R2, as explained in Remark 9.

5. Universally consistent Mondrian predictive systems and probability forecasting systems

Our results (Theorems 21, 23, and 28) will assume that the object space $\mathbf{X}$ is standard Borel (see, e.g., Kechris 1995, Definition 12.5); the class of standard Borel spaces is very wide and contains, e.g., all Euclidean spaces $\mathbb{R}^d$. In this section we start from an easy result (Theorem 21) and its adaptation to deterministic forecasting (Theorem 23).

Theorem 21 *If the object space $\mathbf{X}$ is standard Borel, there exists a universally consistent Mondrian predictive system.*

In Section 6 we will construct a Mondrian predictive system that will be shown in Section 7 to be universally consistent.

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 21 covering this case (Theorem 23 below).

Definition 22 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 *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 23 *If the object space $\mathbf{X}$ is standard Borel, there exists a universally consistent probability forecasting system.*

Theorem 23 will be proved in Section 8.

6. 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 Kechris 1995), $\mathbf{X}$ 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 24 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 25 *The trivial conformity measure is monotonic and satisfies all other conditions of Lemma 11. 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 (7) is always $-\infty$ and never attained, and the supremum on the left-hand side of (8) is always ∞ and never attained. By definition, the histogram taxonomy does not depend on the labels. It remains to apply Lemma 18. ∎

Definition 26 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 section, where we will see that it is universally consistent.

7. Proof of Theorem 21

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:

$$\mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) - \mathbb{E}_P(f \mid x_{n+1}) \to 0, \tag{12}$$

$$\int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \to 0, \tag{13}$$

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 (12) follows by Paul Lévy's martingale convergence theorem (Shiryaev, 2019, 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 (12) (in P^∞-probability).

It remains to prove (13). Let $\epsilon > 0$; we will show that

$$\left| \int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \right| \le \epsilon \tag{14}$$

with high probability for large enough n. By Devroye et al. (1996, 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 (14) holds with high conditional probability given $N > K$ for large enough K. Moreover, it suffices to prove that, for large enough K, (14) 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

$$\left| \frac{1}{N} \sum_{i \in I} f(y_i) - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \right| \le \epsilon/2 \tag{15}$$

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

$$\int f \mathrm{d}Q_n = \frac{1}{N+1} \sum_{i \in I} f(y_i); \tag{16}$$

this follows from $\bar{Q}_n$ (in the notation of Section 2) 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 (16) does not depend on the random number τ.

8. Proof of Theorem 23

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 section. 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 section.) If $I = \emptyset$, let $Q_n(\cdot)$ be the distribution function of the probability measure concentrated at 0. We still have (15) with high probability, and we have (16) with N in place of $N+1$.

9. Universally consistent conformal predictive systems

In this section 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 9).

First we extend the notion of a conformity measure.

Definition 27 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 Section 3, 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:

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

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 11. Conformal predictive systems are defined literally as before.

Theorem 28 *Suppose the object space* $\mathbf{X}$ *is standard Borel. There exists a universally consistent conformal predictive system.*

In Section 10 we will construct a conformal predictive system that will be shown in Section 11 to be universally consistent. The corresponding randomized conformity measure will be monotonic and satisfy all the conditions of Lemma 11 (with objects replaced by extended objects).

10. Histogram conformal predictive systems

In this section we will use the same partitions $\mathcal{P}_n$ of $\mathbf{X} = \mathbb{R}$ as in Section 6.

Definition 29 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 11 (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 section we will show that it is universally consistent.

11. Proof of Theorem 28

The proof in this section is an elaboration of the proof of Theorem 21 in Section 7. The difference is that now we have a different definition of Q_n. It suffices to show that (14) holds with probability at least $1 - \epsilon$ for large enough n, where $\epsilon > 0$ is a given (arbitrarily small) positive constant. In view of (15), it suffices to prove that

$$\left| \int f \mathrm{d}Q_n - \frac{1}{N} \sum_{i \in I} f(y_i) \right| \le \epsilon/2 \tag{17}$$

holds with probability at least $1 - \epsilon/2$ for large enough n. In this section we are using the notation introduced in Section 7, 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 30 *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 $\mathbb{P}(E) \ge 1 - \delta_1\delta_2$,

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

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 30 that we will need:

Lemma 31 *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 30 we have

$$\mathbb{P}\left(\mathbb{P}(y \in [-B, B] \mid x \in \mathcal{P}_n(x')) > 1 - 0.01\epsilon/C\right) \geq 1 - 0.1\epsilon, \tag{18}$$

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 (18) 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

$$\max_{y \in [y_j^*, y_{j+1}^*]} f(y) - \min_{y \in [y_j^*, y_{j+1}^*]} f(y) \leq 0.01\epsilon \tag{19}$$

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 Section 8, 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 32 *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 30, 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 (17) as follows (as explained later):

$$\left|\int f\mathrm{d}Q_n-\frac{1}{N}\sum_{i\in I}f(y_i)\right|=\left|\int f\mathrm{d}Q_n-\int f\mathrm{d}Q_n^*\right| \tag{20}$$

$$\le\left|\int_{(-B,B]}f\mathrm{d}Q_n-\int_{(-B,B]}f\mathrm{d}Q_n^*\right| + C\left(Q_n^*(-B)+1-Q_n^*(B)+Q_n(-B)+1-Q_n(B)\right) \tag{21}$$

$$\le\left|\sum_{i=0}^{m}f(y_i^*)\left(Q_n(y_{i+1}^*)-Q_n(y_i^*)\right)-\sum_{i=0}^{m}f(y_i^*)\left(Q_n^*(y_{i+1}^*)-Q_n^*(y_i^*)\right)\right| + 0.02\epsilon+C\left(0.08\frac{\epsilon}{C}+\frac{2}{K}+\frac{0.22\epsilon}{C(m+1)}+\frac{2}{n}\right) \tag{22}$$

$$\le\sum_{i=0}^{m}|f(y_i^*)|\left|Q_n(y_{i+1}^*)-Q_n^*(y_{i+1}^*)-Q_n(y_i^*)+Q_n^*(y_i^*)\right|+0.2\epsilon \tag{23}$$

$$\le\sum_{i=0}^{m}|f(y_i^*)|\left(\frac{2}{K}+\frac{0.22\epsilon}{C(m+1)}+\frac{2}{n}\right)+0.2\epsilon \tag{24}$$

$$\le\frac{2C(m+1)}{K}+0.42\epsilon+\frac{2C(m+1)}{n}\le 0.5\epsilon. \tag{25}$$

Inequality (21) holds always. Inequality (22) 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 (19) and Lemmas 31 and 32: the addend 0.02ϵ arises by (19) 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 31 (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 32 to reduce bounding $Q_n(-B)$ and $1-Q_n(-B)$ to bounding $Q_n^*(-B)$ and $1-Q_n^*(-B)$, respectively. Inequality (23) holds for sufficiently large K and n. Inequality (24) holds with probability at least $1-0.11\epsilon$ by Lemma 32, but this probability has already been accounted for. And finally, the second inequality in (25) holds for sufficiently large K and n. Therefore, the whole chain (20)–(25) holds with probability at least $1-0.22\epsilon\ge 1-\epsilon/2$. This proves (17), 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.

12. Conclusion

This paper constructs a universally consistent Mondrian predictive system and, which is somewhat more involved, a universally consistent conformal predictive system. There are many interesting directions of further research. These are the most obvious ones:

- Investigate the best rate at which conformal predictive distributions and the true conditional distributions can approach each other.
- 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 9).
- Construct more natural, and perhaps even practically useful, universally consistent randomized predictive systems.

Acknowledgments

Many thanks to the COPA 2019 reviewers for their thoughtful comments and to Astra Zeneca and Stena Line for their support.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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.

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

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

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.

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.

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