Universally consistent predictive distributions

Vladimir Vovk

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

On-line Compression Modelling Project (New Series)

Working Paper #18

First posted April 17, 2017. Last revised April 24, 2017.

Project web site:
http://alrw.net

Abstract

This paper describes simple universally consistent procedures of probability forecasting that satisfy a natural property of small-sample validity, under the assumption that the observations are produced independently in the IID fashion.

Contents

1 Introduction

This paper continues the study of conformal predictive distributions started in [13]. Conformal predictive distributions always satisfy a small-sample property of validity under the assumption that the observations are generated from the same probability distribution in the IID fashion. For convenience we will refer to procedures producing predictive distributions as predictive systems; in particular, conformal predictive systems are procedures producing conformal predictive distributions.

The main result of this paper 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 [2, 3, 10] 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 [11, 4, 5].

We start in Section 2 from defining randomized predictive systems, which are required to satisfy the small-sample property of validity under the IID assumption. The next section defines conformal predictive systems, which are a subclass of randomized predictive systems. The main result of the paper, Theorem 3, is proved in Section 4. Section 5 introduces another subclass of randomized predictive systems, which is wider than the subclass of conformal predictive systems; the elements of this wider subclass are called Mondrian predictive systems. A special case of Theorem 3 given in Section 6 states the existence of Mondrian predictive systems that are universally consistent. An example of a universally consistent Mondrian predictive system is given in Section 7, and Section 8 is devoted to a short proof that this predictive system is indeed universally consistent. After that, a slightly more complicated example of a conformal predictive system is given in Section 9, and it is shown in Section 10 to be universally consistent. In conclusion, Section 11 gives two natural directions of further research.

There is a widely studied sister notion to predictive distributions with a similar small-sample guarantee of validity, namely confidence distributions: see, e.g., [14]. Both confidence and predictive distributions go back to Fisher's fiducial inference. Whereas, under the nonparametric IID assumption of this paper, there are no confidence distributions, [13] and this paper argue that there is a meaningful theory of predictive distributions.

2 Randomized predictive distributions

In this section we give some basic definitions mainly following [13]. 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 a *training sequence* of observations $z_i = (z_i, y_i)$, $i = 1, \ldots, n$, and a new test object $x_{n+1} \in \mathbf{X}$, to predict the label y_{n+1} of the $(n+1)$th observation.

Let U be the uniform probability measure on the interval $[0,1]$. A measurable function $Q : \cup_{n=1}^{\infty}(\mathbf{Z}^{n+1} \times [0,1]) \to [0,1]$, $n = 1, 2, \ldots$, is called a *randomized predictive system* if it satisfies the following requirements:

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

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

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

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

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

The function $Q(z_1, \ldots, z_n, (x_{n+1}, \cdot), \cdot)$ is the *randomized predictive distribution (function)* output by Q for a given training sequence $z_1, \ldots, z_n$ and test object x_{n+1}.

Remark 1. Requirements R1 and R2 are the analogues (introduced in [8]) of similar requirements in the theory of confidence distributions: see, e.g., [14, Definition 1].

3 Conformal predictive distributions

A way of producing randomized predictive distributions has been proposed in [13]. 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 $(z_1, \ldots, z_n) \in \mathbf{Z}^n$, any $z_{n+1} \in \mathbf{Z}$, and any permutation π of $\{1, \ldots, n\}$,

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

The *conformal transducer* determined by a conformity measure A is defined as

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

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

$$\begin{aligned} \alpha_i^y &:= A(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, (x_{n+1}, y), z_i), \qquad i = 1, \ldots, n, \\ \alpha_{n+1}^y &:= A(z_1, \ldots, z_n, (x_{n+1}, y)). \end{aligned} \tag{2}$$

A function is a *conformal transducer* if it is the conformal transducer determined by some conformity measure. A *conformal predictive system* is a function which is both a conformal transducer and a randomized predictive system. If Q is a conformal predictive system, $Q(z_1, \ldots, z_n, (x_{n+1}, \cdot), \cdot)$ are the corresponding *conformal predictive distributions* (or, more fully, conformal predictive distribution functions).

Remark 2. Requirement R2 in the previous section is sometimes referred to as the frequentist validity of predictive or confidence distributions (see, e.g., [14] and [8]). It can be argued that there is no need to appeal to frequencies in these and similar cases (see, e.g., [7, Section 1.4]). However, the property of validity enjoyed by conformal predictive systems is truly frequentist: for them R2 (see (1)) can be strengthened to say that the random numbers $Q(z_1, \ldots, z_n, z_{n+1}, \tau_n)$, $n = 1, 2, \ldots$, are distributed uniformly in $[0, 1]$ and independently, provided $z_n \sim P$ and $\tau_n \sim U$, $n = 1, 2, \ldots$, are independent. In combination with the law of large numbers this implies, e.g., that the frequency of the event

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

(i.e., the frequency of the central $(1 - \epsilon)$-confidence interval failing to cover the true label) converges to ϵ. Notice that this frequentist conclusion depends on the independence of $Q(z_1, \ldots, z_n, z_{n+1}, \tau_n)$; R2 alone is not sufficient.

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

- monotonically increasing in y_{n+1},

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

- monotonically decreasing in y_1,

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

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

If, additionally, a monotonic conformity measure A satisfies, for all n, all training sequences $z_1, \ldots, z_{n+1}$, and all test objects x_{n+1},

$$\begin{aligned} \inf_y A(z_1, \ldots, z_n, (x_{n+1}, y)) &= \inf A, \\ \sup_y A(z_1, \ldots, z_n, (x_{n+1}, y)) &= \sup A, \end{aligned} \tag{3}$$

where both inf are attained and both sup are attained, the corresponding conformal transducer will be a randomized predictive system. (This statement will remain true after replacing either or both of "attained" by "not attained".)

4 Universally consistent conformal predictive distributions

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 $Q_n : y \mapsto Q(z_1, \dots, z_n, (x_{n+1}, y), \tau_n)$ (for a given τ_n) output by Q as its forecast for the label y_{n+1} of x_{n+1} based on the training set $(z_1, \dots, z_n)$, 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 $(x, y) \sim P$;
- $z_n \sim P$ and $\tau_n \sim U$, $n = 1, 2, \dots$, are assumed independent.

It is clear that this notion of consistency 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 $\int f \mathrm{d}Q'$ with the measure Q' on $\mathbb{R}$ defined by $Q'((u, v]) := Q_n(v+) - Q_n(u+)$ for any interval $(u, v]$ of this form in $\mathbb{R}$. Remember that the observations $z_1, z_2, \dots$ are assumed to be generated from P in the IID fashion, and the internal randomization $\tau_1, \tau_2, \dots$ in Q is assumed to be independent of them. The 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, this definition is based on Belyaev's (see, e.g., [3]).

Theorem 3. *Suppose the measurable space $\mathbf{X}$ is standard Borel (see, e.g., [6, Definition 12.5]). There exists a universally consistent conformal predictive system.*

In Section 9 we will construct a conformal predictive system that will be shown in Section 10 to be universally consistent.

5 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.

A *Mondrian taxonomy* κ is a 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 $\{1, \dots, n+1\}$ which is equivariant in the sense to be defined shortly. The notation $(i \sim j \mid z_1, \dots, z_{n+1})$ means that i is equivalent to j under the relation assigned by κ to $(z_1, \dots, z_{n+1})$. 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. 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 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

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

(Our Mondrian taxonomies subsume both Mondrian taxonomies as defined in [12, Section 4.5] and Venn taxonomies as defined in [12, Section 6.3] and, especially, [1, Section 2.2].) 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* associated with a Mondrian taxonomy κ and a conformity measure A is

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

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

Lemma 4. *If a Mondrian taxonomy does not depend on the labels and a conformity measure is monotonic and satisfies* (3)*, the corresponding Mondrian transducer will be a randomized (and, therefore, Mondrian) predictive system.*

Proof. Under the conditions of the lemma, the conformity scores (defined by (2)) α_i^y are monotonically increasing in y when $i = n+1$ and monotonically decreasing in y when $i = 1,\dots,n$. Since the equivalence class of $n+1$ in (5) does not depend on y, the value of (5) is monotonically increasing in y. This proves R1(i). R1(ii) follows immediately from (3). The proof of R2 is standard (see, e.g., [12, Section 8.7]). □

These properties, apart from (3), will be satisfied by the conformity measure and Mondrian taxonomy defined in Section 7 to prove a special case of the main result of this paper. It will be obvious that the corresponding Mondrian transducer satisfies R1(i) (and so is a Mondrian predictive system) despite (3) being violated.

Remark 5. One advantage of conformal predictive systems over Mondrian predictive systems is that the former satisfy a stronger version of R2, as explained in Remark 2.

6 Universally consistent Mondrian predictive distributions and probability forecasting systems

Since every conformal predictive system is a Mondrian predictive system, we have the following corollary of Theorem 3.

Corollary 6. *If the measurable space* $\mathbf{X}$ *is standard Borel, there exists a universally consistent Mondrian predictive system.*

In fact, the proof of Corollary 6 is easier than the proof of Theorem 3, and we will use the former as a gentle introduction to the letter. Namely, in Section 7 we will construct a Mondrian predictive system that will be shown in Section 8 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, the following corollary of the proof of Corollary 6 covers this case. 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 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 sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x_{n+1} \in \mathbf{X}$,

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

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.

Corollary 7. *If the measurable space* $\mathbf{X}$ *is standard Borel, there exists a universally consistent probability forecasting system.*

7 Histogram Mondrian predictive distributions

Since $\mathbf{X}$ is assumed standard Borel, we can set, without loss of generality, $\mathbf{X} := \mathbb{R}$. 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 of $\mathcal{P}_n$ that includes $x \in \mathbf{X}$. The conformity measure A is defined by $A(z_1, \dots, z_n, z_{n+1}) := y_{n+1}$. This conformity measure will be called the *trivial conformity measure*. The Mondrian 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 8. *The trivial conformity measure is monotonic but does not always satisfy* (3)*; still, the Mondrian transducer corresponding to it and the histogram taxonomy is a randomized predictive system.*

Proof. The first statement is obvious, and the second follows from the proof of Lemma 4. □

The Mondrian predictive distributions corresponding to the trivial conformity measure and histogram taxonomy will be called the *histogram Mondrian predictive distributions* and denoted Q_n in the next section, where we will see that they are universally consistent.

8 Proof of Corollary 6

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 this section. We can split (4) into two tasks:

$$\mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) - \mathbb{E}_P(f \mid x) \to 0, \tag{6}$$

$$\int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \to 0, \tag{7}$$

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 (6) (not only in probability but also almost surely) follows by Paul Lévy's martingale convergence theorem [9, Theorem VII.4.3]; Paul Lévy's theorem is applicable since, by our assumption, the partitions $\mathcal{P}_n$ are nested.

It remains to prove (7). Let $\epsilon > 0$; we will show that

$$\left| \int f \mathrm{d}Q_n - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \right| \le \epsilon \tag{8}$$

with high probability for large enough n. By [4, the proof of Theorem 6.2], the number N of observations z_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 (8) holds with high conditional probability given $N > K$ for large enough K. Moreover, it suffices to prove that, for large enough K, (8) 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

$$\left| \frac{1}{N} \sum_{i \in I} f(y_i) - \mathbb{E}_P(f \mid \mathcal{P}_n(x_{n+1})) \right| \le \epsilon/2 \tag{9}$$

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

$$\int f \mathrm{d}Q_n = \frac{1}{N+1} \sum_{i \in I} f(y_i). \tag{10}$$

9 Histogram conformal predictive distributions

The *histogram conformity measure* $A(z_1, \dots, z_n, z_{n+1})$ is 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 of those objects; formally,

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

If $N = 0$, set, e.g.,

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

Since the histogram conformity measure is monotonic and satisfies (3), the corresponding conformal transducer is a conformal predictive system. In the next section we will show that it is universally consistent.

10 Proof of Theorem 3

The proof in this section is an elaboration of the proof of Corollary 6 in Section 8. The difference is that now we have a different definition of Q_n. It suffices to show that (8) holds with probability at least $1 - \epsilon$ for large enough n, where $\epsilon > 0$ is a given (arbitrarily small) positive constant. We will be using the notation introduced in Section 8, such as N and I.

Set $C := \sup |f|$. Remember that $\epsilon > 0$ is a given positive constant. Let B be so large that $y \in [-B, B]$ with probability at least $1 - 0.01\epsilon/C$ when $(x, y) \sim P$. Since f is uniformly continuous over $[-B, B]$, there is a partition

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

of the interval $[-B, B]$ such that $\left| f(y_{i+1}^*) - f(y_i^*) \right| \le 0.01\epsilon$ for $i = 1, \dots, m$. Let us assume, without loss of generality, that $m > 10$.

Let K be a large positive integer. We can choose n so large that $N \ge K$ with probability at least $1 - 0.01\epsilon^2/C(m+1)$. 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 $1 - 0.1\epsilon$; this follows from Markov's inequality:

$$\mathbb{P}\Big(\mathbb{P}\left(N < K \mid x_1, \ldots, x_n\right) \geq 0.1\epsilon/C(m+1)\Big) \leq \frac{\mathbb{E}\left(\mathbb{P}\left(N < K \mid x_1, \ldots, x_n\right)\right)}{0.1\epsilon/C(m+1)}$$
$$= \frac{\mathbb{P}(N < K)}{0.1\epsilon/C(m+1)} \leq \frac{0.01\epsilon^2/C(m+1)}{0.1\epsilon/C(m+1)} = 0.1\epsilon.$$

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 $1 - 0.1\epsilon$ (indeed, we can choose n satisfying the condition in the previous sentence and generate sufficiently many new observations). For such n:

- Q_n is still concentrated at the points y_i, $i \in I$; namely, Q_n is constant between adjacent points y_i, $i \in I$;
- let I' be the subset of I consisting of those $i \in I$ for which there is no $j \in I$ such that $y_i = y_j$ and $j < i$, and, for each $i \in I'$, let $c_i \in \{1, 2, \ldots\}$ be the number of times y_i occurs among y_j, $j \in I$ (this is one way of dealing with duplicates among y_i, $i \in I$);
- let Q_n^* be the distribution function of the probability measure on $\mathbb{R}$ that gives weight c_i/N to each y_i, $i \in I'$;
- $Q_n(y_i)$ and $Q_n^*(y_i)$ differ from each other by at most $1/K + 0.11\epsilon/C(m+1)$ for all $i \in I'$ with probability, over the choice of $x_1, \ldots, x_n$, at least $1 - 0.1\epsilon$.

We will also assume n to be large enough to ensure that the probability of the fraction of y_i, $i = 1, \ldots, n$, satisfying $y_i \in [-B, B]$ to be more than $1 - 0.02\epsilon/C$ is at least $1 - 0.1\epsilon$.

We still have (9), but (10) now becomes:

$$\left|\int f \mathrm{d}Q_n - \int f \mathrm{d}Q_n^*\right| \tag{11}$$
$$\leq \left|\int_{-B}^{B} f \mathrm{d}Q_n - \int_{-B}^{B} f \mathrm{d}Q_n^*\right| + C\left(Q_n^*(-B) + 1 - Q_n^*(B) + Q_n(-B) + 1 - Q_n(B)\right) \tag{12}$$
$$\leq \left|\sum_{i=0}^{m} f(y_i^*)\left(Q_n(y_{i+1}^*) - Q_n(y_i^*)\right) - \sum_{i=0}^{m} f(y_i^*)\left(Q_n^*(y_{i+1}^*) - Q_n^*(y_i^*)\right)\right| + 0.02\epsilon + C\Big(0.02\epsilon/C + 0.02\epsilon/C + \frac{1}{K} + 0.11\epsilon/C(m+1)\Big) \tag{13}$$
$$\leq \sum_{i=0}^{m} |f(y_i^*)| \left|Q_n(y_{i+1}^*) - Q_n^*(y_{i+1}^*) - Q_n(y_i^*) + Q_n^*(y_i^*)\right| + 0.1\epsilon \tag{14}$$

$$\le \sum_{i=0}^{m} |f(y_i^*)| \left(\frac{2}{K} + 0.22\epsilon/C(m+1) \right) + 0.1\epsilon \tag{15}$$

$$\le \frac{2C(m+1)}{K} + 0.32\epsilon \le 0.5\epsilon. \tag{16}$$

Inequality (12) holds always, inequality (13) holds with probability at least $1 - 0.1\epsilon - 0.1\epsilon = 1 - 0.2\epsilon$, inequality (14) holds for sufficiently large K, inequality (15) holds with probability at least $1 - 0.1\epsilon$, and finally the last inequality in (16) holds for sufficiently large K; therefore, the whole chain holds with probability at least $1 - 0.3\epsilon \ge 1 - \epsilon/2$. Combining the chain (11)–(16) and (9) gives (8) and, therefore, completes the proof.

To avoid any ambiguity, let me describe once again the role of ϵ, K, and n in this proof. First we fix a positive constant $\epsilon > 0$ (which, however, can be arbitrarily small). We then choose K, which should be sufficiently large for the given ϵ. Finally, we choose n, which should be sufficiently large for the given ϵ and K.

11 Conclusion

These are two obvious directions of further research:

- Replace universal consistency by strong universal consistency (i.e., convergence in probability by convergence almost surely).
- Construct more natural, and perhaps even practically useful, universally consistent randomized predictive distributions.

References

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

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

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

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

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

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

[7] Glenn Shafer. Game-theoretic significance testing. The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper 49, April 2017.

[8] Jieli Shen, Regina Liu, and Min-ge Xie. Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 2017. To appear.

[9] Albert N. Shiryaev. *Вероятность (Probability)*. МЦНМО, Moscow, third edition, 2004.

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

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

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

[13] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 17, March 2017.

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