Inductive randomness predictors

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 #44

First posted March 4, 2025. Last revised March 22, 2025.

Project web site:
http://alrw.net

Abstract

This paper introduces inductive randomness predictors, which form a superset of inductive conformal predictors. It turns out that a typical inductive conformal predictor is strictly dominated by an inductive randomness predictor, although the improvement is not great, at most a factor of $e \approx 2.72$. The dominating inductive conformal predictors are more complicated and more difficult to compute; besides, an improvement by a factor of e is rare.

Contents

1 Introduction

Randomness predictors were introduced and studied in [15]. Their definition is trivial (it is a straightforward application of the definition of p-values), but they include conformal predictors as their proper subclass, and conformal predictors have been widely implemented (see, e.g., [2,3]), used (see, e.g., [13]), and studied (see, e.g., [1]). Both [15] and the follow-up paper [16] concentrate on negative results about randomness predictors, showing that the difference in predictive efficiency between conformal and randomness prediction is not great. (While [15] covers worst-case difference, [16] also treats difference on average.) This paper concentrates, instead, on positive results, giving examples of situations were randomness predictors have an advantage over conformal predictors.

Both conformal and randomness predictors are valid (ensure the desired coverage probability) under the assumption of randomness, which is standard in machine learning. The main advantage of randomness prediction, if real, may lie in its *efficiency*, which is defined, informally, as the smallness of the p-values that it produces for false labels. A major limitation of conformal predictors, discussed in detail in [18], is that the p-values that they output can never drop below $\frac{1}{n+1}$, where n is the length of the training sequence. An advantage of randomness predictors is that the lower bound improves to $\frac{1}{\mathrm{e}(n+1)}$. The factor of e (the base of natural logarithms, $\mathrm{e} \approx 2.72$) in the denominator is negligible by the usual standards of the algorithmic theory of randomness, but substantial from the point of view of standard machine learning and statistics.

The most popular kind of conformal predictors is inductive conformal predictors. Their main advantage is that they can be used on top of generic point predictors without prohibitive computational costs, whereas full conformal prediction is computationally efficient only on top of a relatively narrow class of point predictors. The smallest p-value that can be achieved by an inductive conformal predictor is $\frac{1}{m+1}$, where m is the number of "calibration examples" (to be defined later). This paper introduces and studies inductive randomness predictors, which are also computationally efficient (at least in the part that depends on the actual data rather than merely on m).

We will start in Sect. 2 from the main definitions, including that of inductive randomness predictors. Section 3 is devoted to computing binary inductive randomness predictors, whose use is illustrated on two examples. The topic of Sect. 4 is the inadmissibility of inductive conformal predictors as inductive randomness predictors. Section 5 introduces a more interesting class of inductive randomness predictors, which we call "sequential". While both binary and sequential inductive randomness predictors sometimes achieve p-values of $\frac{1}{\mathrm{e}(m+1)}$, only sequential inductive randomness predictors dominate inductive conformal predictors. The short Sect. 6 concludes.

Let $\mathbb{N}_0 := \{0, 1, \dots\}$ and $\mathbb{N}_1 := \{1, 2, \dots\}$ be the two standard sets of natural numbers.

2 Definitions

The prediction problem considered in this paper is the same as in [15,16]. We are given a training sequence $z_1, \dots, z_n$, where $z_i = (x_i, y_i)$ (an *example*) consists of an *object* $x_i \in \mathbf{X}$ and a *label* $y_i \in \mathbf{Y}$, and a test object $x_{n+1} \in \mathbf{X}$. Our task is to predict the label y_{n+1} of x_{n+1}. The object space $\mathbf{X}$ and the label space $\mathbf{Y}$ are non-empty measurable spaces, and the length n of the training sequence is fixed. To exclude trivialities, let us assume that $n \geq 2$ and that the σ-algebra on $\mathbf{Y}$ is different from $\{\emptyset, \mathbf{Y}\}$ (i.e., that $\mathbf{Y}$ contains at least two essentially distinct elements).

In the definition of an inductive conformal predictor we will follow [17, Sect. 4.2.2]. The training sequence $z_1, \dots, z_n$ is split into two parts: the *proper training sequence* $z_1, \dots, z_l$ of size l and the *calibration sequence* $z_{l+1}, \dots, z_n$ of size $m := n - l$; we will assume $l \in \mathbb{N}_1$ and $m \in \mathbb{N}_1$. An *inductive nonconformity measure* is a measurable function $A : \mathbf{Z}^{l+1} \to \mathbb{R}$, where $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is the *example space*. The *inductive conformal predictor* (ICP) based on A outputs the prediction p-function

$$f(y) := \frac{|\{j = l+1, \dots, n+1 \mid \alpha_j \geq \alpha_{n+1}\}|}{m+1} \in \left[\frac{1}{m+1}, 1\right], \tag{1}$$

where the αs are defined by

$$\alpha_j := A(z_1, \dots, z_l, z_j), \quad j = l+1, \dots, n, \tag{2}$$

$$\alpha_{n+1} := A(z_1, \dots, z_l, x_{n+1}, y) \tag{3}$$

To define and discuss inductive randomness predictors, we will need several auxiliary notions. The *upper randomness probability* of a measurable set $E \subseteq \mathbf{Z}^{n+1}$ is defined in [17, Sect. 9.1.1] as

$$\mathbb{P}^{\mathrm{R}}(E) := \sup_{Q \in \mathfrak{P}(\mathbf{Z})} Q^{n+1}(E), \tag{4}$$

where we use the notation $\mathfrak{P}(Z)$ for the set of all probability measures on a measurable set Z. An *inductive nonconformity measure* is a measurable function $A : \mathbf{Z}^{l+1} \to \mathbf{S}$, where $\mathbf{S}$ is a measurable space which we will call the *summary space*; typically, $\mathbf{S} \subseteq \mathbb{R}$, and so our new definition is a very slight modification of the old one. Similarly to (4), we define the upper randomness probability of a measurable set $E \subseteq \mathbf{S}^{m+1}$ as

$$\mathbb{P}^{\mathrm{R}}(E) := \sup_{Q \in \mathfrak{P}(\mathbf{S})} Q^{m+1}(E).$$

(Therefore, the notation $\mathbb{P}^{\mathrm{R}}$ is overloaded, but it should never lead to confusion in this paper.) An *aggregating p-variable* $P : \mathbf{S}^{m+1} \to [0,1]$ is defined to be a randomness p-variable on $\mathbf{S}^{m+1}$; its defining requirement is

$$\forall \epsilon \in (0,1) : \mathbb{P}^{\mathrm{R}}(\{P \leq \epsilon\}) \leq \epsilon. \tag{5}$$

A *randomness predictor*, as defined in [15, 16], is a p-variable $P : \mathbf{Z}^{n+1} \to [0, 1]$, meaning that it is required to satisfy (5).

In inductive randomness prediction, the training sequence $z_1, \dots, z_n$ is still split into the proper training sequence $z_1, \dots, z_l$ and the calibration sequence $z_{l+1}, \dots, z_n$. The *inductive randomness predictor* (IRP) based on (sometimes we will say "corresponding to") an inductive nonconformity measure A and an aggregating p-variable P is defined to be the randomness predictor

$$P_A(z_1, \dots, z_{n+1}) := P(\alpha_{l+1}, \dots, \alpha_{n+1}),$$

where

$$\alpha_j := A(z_1, \dots, z_l, z_j), \quad j = l + 1, \dots, n + 1. \tag{6}$$

Given a training sequence $z_1, \dots, z_n$ and a test object x_{n+1}, the IRP P_A outputs the prediction p-function

$$f(y) = f(y; z_1, \dots, z_n, x_{n+1}) := P_A(z_1, \dots, z_n, x_{n+1}, y). \tag{7}$$

This function itself can be considered to be the IRP's prediction for y_{n+1}. Alternatively, we can choose a *significance level* $\epsilon > 0$ (i.e., our target probability of error) and output the prediction set

$$\Gamma^\epsilon := \{y \in \mathbf{Y} \mid f(y) > \epsilon\} \tag{8}$$

as our prediction for y_{n+1}. By the definition of p-variable, the probability of error (meaning $y_{n+1} \notin \Gamma^\epsilon$) will not exceed ϵ.

IRPs considered in this paper will often output prediction p-functions of an especially simple kind. Let us say that the prediction function (7) is a *hedged prediction set* if it has the form

$$f(y) = \begin{cases} 1 & \text{if } y \in E \\ c & \text{otherwise,} \end{cases}$$

where $E \subseteq \mathbf{Y}$ is the prediction set associated with it and $c \in [0, 1)$ reflects our confidence in this prediction set; the smaller c the greater confidence. We will refer to c as the *incertitude* of the prediction set E. As always, the expression "prediction interval" will be applied to prediction sets that happen to be intervals of the real line, and the corresponding hedged prediction sets will be called hedged prediction intervals.

Remark 1. In our analysis of inductive randomness predictors, we will assume that all $n + 1$ examples under consideration are IID, although it will be obvious that it is sufficient to assume that only the calibration and test examples are IID.

ICPs are a special case of IRPs based on the aggregating p-variable

$$\Pi(\alpha_{l+1}, \dots, \alpha_{n+1}) := \frac{|\{j = l + 1, \dots, n + 1 \mid \alpha_j \geq \alpha_{n+1}\}|}{m + 1},$$

$$(\alpha_{l+1}, \dots, \alpha_{n+1}) \in \mathbf{S}^{m+1}. \quad (9)$$

Therefore, we will use the notation Π_A for the ICP based on an inductive nonconformity measure A.

In statistical hypothesis testing (see, e.g., [4, Sect. 3.2]) it is customary to define p-variables via "test statistics". In this spirit, we can define an *aggregating function* as any measurable function $B : \mathbf{S}^{m+1} \to \mathbb{R}$. It defines the aggregating p-variable

$$P_B(\alpha_{l+1}, \dots, \alpha_{n+1}) := \mathbb{P}^{\mathrm{R}}\left(\{B \geq B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1})\}\right), \\ (\alpha_{l+1}, \dots, \alpha_{n+1}) \in \mathbf{S}^{m+1}. \quad (10)$$

(Intuitively, large values of B indicate nonconformity.) This aggregating p-variable can then be used as an input to an IRP, and then we might say that this IRP is based on A (an inductive nonconformity measure) and B.

3 Binary inductive randomness predictors

In this section we will concentrate on *binary inductive randomness predictors* (BIRPs), for which the summary space is $\mathbf{S} := \{0, 1\}$. Intuitively, a summary of 0 means conformity, and 1 means lack of conformity. BIRPs are simple but they are less efficient than the sequential randomness predictors considered later in Sect. 5.

Let me give two examples of BIRPs, one for regression and another for binary classification.

Example 2. Here we are interested in a regression problem, so that $\mathbf{Y} = \mathbb{R}$. The inductive nonconformity measure A is defined as follows: to define $A(z_1, \dots, z_l, x, y)$, train a regression model $\hat{g} : \mathbf{X} \to \mathbb{R}$ on $z_1, \dots, z_l$ as training sequence and set

$$A(z_1, \dots, z_l, x, y) := \begin{cases} 1 & \text{if } |y - \hat{g}(x)| > \max_{i=1,\dots,l} |y_i - \hat{g}(x_i)| \\ 0 & \text{otherwise,} \end{cases} \quad (11)$$

where $z_i = (x_i, y_i)$, $i = 1, \dots, l$. As for B, we set

$$B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1}) := \max\left(\alpha_{n+1} - \frac{1}{n}\sum_{i=1}^{n} \alpha_i, 0\right). \quad (12)$$

Therefore, $B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1})$ is 0 if $\alpha_{n+1} = 0$ and is positive otherwise (unless $\alpha_1, \dots, \alpha_n$ are all zero).

Let us see how the IRP based on A and B can be applied in the context of inductive randomness prediction assuming that $A(z_1, \dots, z_l, z_i) = 0$ for some $i \in \{l+1, \dots, n\}$ (this excludes a very anomalous case with severe overfitting). Given a training sequence $z_1, \dots, z_n$, we start from training a regression model

$\hat{g} : \mathbf{X} \to \mathbb{R}$ on the proper training sequence $z_1, \ldots, z_l$. Next we compute the half-width $h := \max_{i=1,\ldots,l} |y_i - \hat{g}(x_i)|$ of the hedged prediction intervals output by the IRP. Given a test object x_{n+1}, we compute the prediction interval $[c-h, c+h]$ centred at the point prediction $c := \hat{g}(x_{n+1})$. The incertitude of this prediction interval will be computed in Proposition 4, as discussed after the statement of the proposition. Only the last step (computing the incertitude) involves the calibration sequence.

Example 3. Now set $\mathbf{Y} := \{-1, 1\}$, so that here we are interested in binary classification. To define $A(z_1, \ldots, z_l, x, y)$, consider the support vector machine (SVM) constructed from $z_1, \ldots, z_l$ as training sequence. Set $A(z_1, \ldots, z_l, x, y)$ to 1 if x is classified incorrectly (namely, as $-y$) by this SVM and x is outside its margin; set $A(z_1, \ldots, z_l, x, y)$ to 0 otherwise. A reasonable definition of B is still (12).

The hedged prediction set for a test object x_{n+1} will be $\{\hat{y}\}$ if x_{n+1} is outside the margin, where $\hat{y}$ is the SVM's prediction for the label of x_{n+1}. Otherwise (if x_{n+1} is inside the margin), the prediction set will be vacuous, $\{-1, 1\}$. This assumes, again, that $A(z_1, \ldots, z_l, z_i) = 0$ for some $i \in \{l + 1, \ldots, n\}$. The incertitude of this prediction set will be given after the statement of Proposition 4, and only this step uses the calibrating sequence.

An alternative definition would be to set $A(z_1, \ldots, z_l, x, y)$ to 1 if x is a support vector for the SVM constructed from $(z_1, \ldots, z_l, x, y)$ as training sequence and to set it to 0 otherwise, as in [9, Sect. 2]. However, the computational cost of such an IRP would be prohibitive, since it would require constructing a new SVM for each text object and each possible label for it.

Both IRPs described in Examples 2 and 3 output predictions sets that do not depend on the calibration sequence. This makes them inflexible as compared with typical conformal predictors, but on the positive side they can achieve very low incertitudes.

Let us now compute the p-values output by BIRPs based on the aggregating function (12). The following proposition gives the result of the computation, and after its statement we will discuss ways of using it.

Proposition 4. *Suppose that a binary sequence $\alpha_{l+1}, \ldots, \alpha_n$ contains $K < m$ 1s and that $\alpha_{n+1} = 1$. Then the aggregating function B defined by* (12) *leads to a p-value $P_B(\alpha_{l+1}, \ldots, \alpha_{n+1})$ of*

$$\max_{p \in [0,1]} \sum_{k=0}^{K} \binom{m}{k} p^{k+1} (1-p)^{m-k}. \tag{13}$$

In particular,

- *for $K = 0$, the p-value is*

$$\frac{m^m}{(m+1)^{m+1}} \sim \frac{\exp(-1)}{m} \approx \frac{0.37}{m}, \tag{14}$$

  where "$\sim$" holds as $m \to \infty$ (and we can replace "$\sim$" by "$\leq$"),

- *for $K = 1$, the p-value is asymptotically equivalent (as $m \to \infty$) to*

$$\frac{(\phi + \phi^2)\exp(-\phi)}{m} \approx \frac{0.84}{m}, \tag{15}$$

where $\phi := (1 + \sqrt{5})/2$ is the golden ratio,

- *for $K = 2$, the p-value is asymptotically equivalent to*

$$\frac{(c + c^2 + c^3/2)\exp(-c)}{m} \approx \frac{1.37}{m}, \tag{16}$$

where

$$c := \frac{1 + (37 - 3\sqrt{114})^{1/3} + (37 + 3\sqrt{114})^{1/3}}{3},$$

- *and for $K = 3$, the p-value is asymptotically equivalent to*

$$\frac{(c + c^2 + c^3/2 + c^4/6)\exp(-c)}{m} \approx \frac{1.94}{m}, \tag{17}$$

where

$$\begin{aligned} c := \frac{1}{4} + \frac{1}{4}\Big(4(\sqrt{778} - 7)^{1/3} - 36(\sqrt{778} - 7)^{-1/3} + 9\Big)^{1/2} \\ + \frac{1}{2}\Bigg(-(\sqrt{778} - 7)^{1/3} + 9(\sqrt{778} - 7)^{-1/3} + \frac{9}{2} \\ + \frac{61}{2\sqrt{4(\sqrt{778} - 7)^{1/3} - 36(\sqrt{778} - 7)^{-1/3} + 9}}\Bigg)^{1/2}. \end{aligned}$$

In the context of Example 2, we can expect that $K = 0$ if the calibration sequence is much shorter than the proper training sequence and $\hat{g}$ does not involve too much overfitting. In this case the prediction interval output by the IRP based on (11) and (12) will be more confident than the identical prediction interval output by the ICP based on the same inductive nonconformity measure (11): the incertitude of of the former will be approximately $0.37/m$ for large m, whereas the incertitude of the latter will be approximately $1/m$. An advantage of ICPs is, of course, that their hedged prediction intervals can be much more adaptive and, moreover, their prediction p-functions do not have to be hedged prediction sets.

Even if $K = 1$, the incertitude for the IRP based on (11) and (12) is still close to $0.84/m$ (see (15)), which is better than the smallest p-value that can be achieved by any ICP on any training sequence.

In the context of Example 3, the definition of the nonconformity measure A was chosen so that K can be expected to be small. In this case the incertitude of the IRP based on (11) and (12) will be significantly better than the incertitude of the ICP based on (11) (we will discuss this further after the proof; cf. Table 1).

Proof of Proposition 4. The condition of the proposition implies that the inductive nonconformity measure A is a surjection. Let B_p be the Bernoulli probability measure on $\{0,1\}$ with parameter $p \in [0,1]$: $B_p(\{1\}) = p$. Since the sequence $\alpha_{l+1}, \dots, \alpha_{n+1}$ is IID, the p-value is the largest probability under B_p^{m+1} of the event of observing at most K 1s among $\alpha_{l+1}, \dots, \alpha_n$ and observing $\alpha_{n+1} = 1$. This gives the expression (13).

When $K = 0$, $\max_p p(1-p)^m$ is attained at $p = \frac{1}{m+1}$, which leads to (14). The inequality

$$\frac{m^m}{(m+1)^{m+1}} \le \frac{\exp(-1)}{m} \tag{18}$$

is equivalent to

$$\left(1 - \frac{1}{m+1}\right)^{m+1} \le \exp(-1)$$

and is easy to check.

When $K = 1$, solving the optimization problem

$$p(1-p)^m + mp^2(1-p)^{m-1} \to \max \tag{19}$$

leads to a quadratic equation with the solution in $[0,1]$ equal to

$$\frac{m - 2 + \sqrt{5m^2 - 4m}}{2(m^2 - 1)} \sim \frac{\phi}{m}.$$

Plugging this into the objective function (19) gives (15).

Now let us deal with an arbitrary (but fixed) K and let $m \to \infty$. The optimal value of p in (13) will be of the form $p \sim c/m$ for a constant c (as we will see later in the proof). Plugging $p \sim c/m$ into the expression following $\max_{p \in [0,1]}$ in (13), we can see that this expression is asymptotically equivalent to

$$\sum_{k=0}^{K} \frac{c^{k+1} \mathrm{e}^{-c}}{k!\, m}. \tag{20}$$

This gives the left-hand sides of (16) and (17). Setting the derivative of (20) to 0, we can check that the optimal c satisfies the equation

$$\sum_{k=0}^{K} \frac{c^k}{k!} = \frac{c^{K+1}}{K!}.$$

In the cases of $K = 2$ and $K = 3$, we obtain cubic and quartic equations, respectively, and their solutions are given in the statement of the proposition. □

Table 1 gives the numerators of asymptotic expressions such as (14)–(17) for a wide range of K. The IRP is based on (11) and (12), and the ICP is based on (11). The row labelled "IRP" gives the numerator itself, and the row labelled "ratio" gives the ratio of the numerator for the IRP to the numerator for the ICP. We can see that the ratio is substantially less than 1 even for $K = 7$, in which case we have $4.472/m$ for the IRP (approximately) and $0.125/m$ for the ICP; the growth of the ratio quickly slows down as K increases.

Table 1: The asymptotic numerators of the incertitudes for the IRP and ICP for various values of K: the asymptotic incertitude for the prediction set output by the IRP is a_K/m, where a_K is given in row "IRP", and the asymptotic incertitude for the ICP is $(K+1)/m$, with the numerator $K+1$ given in row "ICP". Row "ratio" reports $a_K/(K+1)$ showing by how much a_K/m is smaller.

K	0	1	2	3	4	5	6	7
IRP	0.368	0.840	1.371	1.942	2.544	3.168	3.812	4.472
ICP	1	2	3	4	5	6	7	8
ratio	0.368	0.420	0.457	0.486	0.509	0.528	0.545	0.559

4 Admissibility of inductive randomness predictors

Let us say that an IRP P_1 *dominates* an IRP P_2 if $P_1 \le P_2$ (the p-value output by P_1 never exceeds the p-value output by P_2 on the same data). The domination is *strict* if, in addition, $P_1(z_1,\dots,z_{n+1}) < P_2(z_1,\dots,z_{n+1})$ for some data sequence $z_1,\dots,z_{n+1}$.

An equivalent way to express the domination of P_2 by P_1 is to say that, at each significance level, the prediction set output by P_1 is a subset of (intuitively, is at least as precise as) the prediction set output by P_2. The strict domination means that sometimes the prediction set output by P_1 is more precise. An IRP (in particular, an ICP) is *inadmissible* if it is strictly dominated by another IRP. This is a special case of the standard notion of inadmissibility in statistics.

Proposition 5. *Any inductive conformal predictor is inadmissible.*

Proof. Let A be an inductive nonconformity measure; let us check that we can improve on the corresponding ICP Π_A and define an IRP P_A strictly dominating Π_A. If A takes only one value, Π_A always outputs 1 and so is clearly inadmissible (being strictly dominated by the ICP based on any inductive conformity measure taking at least two distinct values). So let us assume that A takes at least two distinct values, choose arbitrarily $a \in (\inf A, \sup A)$, and define P as

$$P(\alpha_{l+1},\dots,\alpha_{n+1}) := \begin{cases} \frac{m^m}{(m+1)^{m+1}} & \text{if } \alpha_{n+1} > a \text{ and } \alpha_i < a \text{ for all } i \in \{l+1,\dots,n\} \\ \Pi(\alpha_{l+1},\dots,\alpha_{n+1}) & \text{otherwise.} \end{cases}$$

By inequality (18), P can produce p-values that are impossible for ICPs.

It is easy to check that P is a p-variable:

- when $\epsilon \ge \frac{1}{m+1}$, $Q^{m+1}(P \le \epsilon) \le \epsilon$ follows from $Q^{m+1}(\Pi \le \epsilon) \le \epsilon$ (since P improves on Π only when $\Pi = \frac{1}{m+1}$),

- when $\epsilon < \frac{1}{m+1}$, $Q^{m+1}(P \le \epsilon) \le \epsilon$ follows from the fact that the probability that B_p^{m+1} produces exactly one 1 and that the 1 is the last bit is given by the left-most expression in (14).

It is also clear that P_A strictly dominates Π_A. □

It should be noted that Proposition 5 can be demonstrated in "uninteresting" ways, which shows that our unrestricted requirement of admissibility is too strong. For example, if an IRP is *calibration-invariant*, i.e., invariant w.r. to the permutations of the calibration sequence (such as any ICP), it is easy to improve it (in the sense of sometimes getting smaller p-values, not in any really useful sense) by allowing dependence on the order of the calibration sequence. In our proof the dominating IRP is at least still calibration-invariant.

The phenomenon of inadmissibility of ICPs demonstrated by our proof of Proposition 5 is akin to the phenomenon of superefficiency in point estimation (see, e.g., [12] and [14, Sect. 2] for reviews). We are making an ICP superefficient at a nonconformity score that we choose arbitrarily, as in Hodges's example [12, Fig. 1]. In such situations the standard term "inadmissibility" also appears too harsh.

5 Sequential inductive randomness predictors

In this section we will assume that the summary space $\mathbf{S}$ is a closed interval (perhaps infinite in both directions, $\mathbb{R}$, or in one direction, $[c, \infty)$ or $(-\infty, c]$ for some $c \in \mathbb{R}$). Our results will be easiest to interpret if the reader assumes that the distribution of nonconformity scores $A(z_1, \dots, z_l, Z)$ (where $Z \sim Q$ and Q^{n+1} is the data-generating distribution) is continuous; in any case, this is what we will assume in informal discussions.

A *sequential inductive randomness predictor* (SIRP) is determined by an inductive nonconformity measure A and a 2D array (*threshold array*) $(c_{K,I})$, $K \in \{0, \dots, m-1\}$ and $I \in \mathbb{N}_1$, of real numbers in $\mathbf{S}$ that is dense for each $K \in \{0, \dots, m-1\}$:

$$\forall K \in \{0, \dots, m-1\} : \overline{\{c_{1,K}, c_{2,K}, \dots\}} = \mathbf{S}. \tag{21}$$

The SIRP is defined as follows. Let the conformal p-value, as defined in (1), be $(K+1)/(m+1)$. Now let I be the smallest index such that the test nonconformity score and exactly K calibration nonconformity scores are above $c_{K,I}$. Then the p-value output by the SIRP is

$$\mathrm{SIRP}(m, K, I) := \frac{K}{m+1} + \max_{(p_0, \dots, p_I) \in \Delta_I} \sum_{i=1}^{I} \sum_{k=0}^{K} \frac{m!}{(m-K)!(k+1)!(K-k)!} \left(\sum_{j=0}^{i-1} p_j \right)^{m-K} p_i^{k+1} \left(\sum_{j=i+1}^{I} p_j \right)^{K-k}, \quad K \in \{0, \dots, m-1\}, \ I \in \mathbb{N}_1, \tag{22}$$

Table 2: Some p-values $\mathrm{SIRP}(9,0,\cdot)$ for a calibration sequence of length $m=9$ corresponding to the smallest conformal p-value of 10% for various values of I.

I	1	2	3	4	5	6	7	8
SIRP	3.87%	5.53%	6.46%	7.07%	7.49%	7.81%	8.05%	8.25%
I	9	10	11	12	13	14	99	100
SIRP	8.41%	8.54%	8.65%	8.75%	8.83%	8.90%	9.82%	9.83%

where Δ_I is the standard I-simplex

$$\Delta_I := \left\{(p_0,\dots,p_I) \in [0,\infty)^{I+1} \mid p_0+\dots+p_I = 1\right\}.$$

If such I does not exist, set $\mathrm{SIRP}(m,K,I) := (K+1)/(m+1)$. The word "above" in this definition can be either inclusive or exclusive, so that "α_i is above $c_{K,I}$" may mean either $\alpha_i > c_{K,I}$ or $\alpha_i \ge c_{K,I}$. For concreteness, let us use the latter meaning. Then "α_i is below $c_{K,I}$" means $\alpha_i < c_{K,I}$.

The expression 0^0 in (22) is treated as 1. Therefore, the term in the sum $\sum_{i=1}^{I}$ corresponding to $i=I$ only contains the term corresponding to $k=K$ in the sum $\sum_{k=0}^{K}$, in which the factor $(\dots)^{K-k}$ can be ignored.

Remark 6. In principle, the definition (22) also works for $K := m$. However, this case does not bring anything interesting: even in the extreme case $I=1$, the second line of (22) is easily seen to be $1/(m+1)$ (set $k := K$, $p_0 := 0$, and $p_1 := 1$), and so $\mathrm{SIRP}(m,m,1) = 1$, which is a vacuous p-value.

Before studying SIRPs theoretically and even before proving that they are bona fide IRPs, let us see how they work. Table 2 shows the p-values (22) that they produce for a calibration sequence of length $m=9$ when the conformal p-value takes its smallest value 10%. In the case where the conformal p-value takes its smallest value $1/(m+1)$, I is the smallest index such that all calibration nonconformity scores are below $c_{0,I}$ and the test nonconformity score is above $c_{0,I}$, and the SIRP p-value (22) can then be written as

$$\mathrm{SIRP}(m,0,I) = \max_{(p_0,\dots,p_I)\in\Delta_I} \sum_{i=1}^{I} \left(\sum_{j=0}^{i-1} p_j\right)^m p_i. \tag{23}$$

The smallest possible p-value for SIRPs corresponds to $I=1$ and is 3.87%. All p-values in the table are below 10%, and later we will see that each SIRP strictly dominates the corresponding ICP.

Table 3 is the analogue of Table 2 for the second and third smallest conformal p-values, 20% and 30%. For the conformal p-value of $2/(m+1)$, I is the smallest index such that the test nonconformity score and exactly one calibration nonconformity score are above $c_{1,I}$, and the expression for the p-value can

Table 3: Some p-values $\mathrm{SIRP}(9,1,\cdot)$ and $\mathrm{SIRP}(9,2,\cdot)$ for $m=9$ corresponding to the conformal p-value of 20% (row labelled $K=1$) and 30% (row labelled $K=2$) for various values of I.

I	1	2	3	4	5	6	100
$K=1$	13.02%	14.59%	15.56%	16.23%	16.72%	17.10%	19.74%
$K=2$	22.67%	24.17%	25.14%	25.83%	26.34%	26.74%	29.70%

be slightly simplified to

$$\mathrm{SIRP}(m,1,I) = \frac{1}{m+1} + m \max_{(p_0,\dots,p_I)\in\Delta_I} \sum_{i=1}^{I} \left(\sum_{j=0}^{i-1} p_j\right)^{m-1} \left(\frac{p_i}{2} + \sum_{j=i+1}^{I} p_j\right) p_i. \tag{24}$$

To see how SIRPs could be used for prediction, let us consider a very simple and standard inductive nonconformity measure [17, (4.16)].

Example 7. Consider the problem of regression, $\mathbf{Y} := \mathbb{R}$, as in Example 2. Train a regression model $\hat{g} : \mathbf{X} \to \mathbb{R}$ (such as a neural network) on $z_1, \dots, z_l$ as training sequence. Use $A(z_1, \dots, z_l, x, y) := |y - \hat{g}(x)|$ as nonconformity measure, and set $\mathbf{S} := [0, \infty)$. Let $\alpha_i := |y_i - \hat{g}(x_i)|$, $i = l+1, \dots, n$, be the ith calibration nonconformity score. Arrange these nonconformity scores in the ascending order, $\alpha_{(1)} \le \dots \le \alpha_{(m)}$. Set $\hat{y}_{n+1} := \hat{g}(x_{n+1})$. These are the prediction intervals Γ^ϵ (see (8)) output by the SIRP based on a threshold array $(c_{K,I})$ in $\mathbf{S}$:

- First, it outputs all the conformal prediction intervals:

$$\Gamma^{\frac{K+1}{m+1}} = \left[\hat{y}_{n+1} - \alpha_{(m-K)}, \hat{y}_{n+1} + \alpha_{(m-K)}\right].$$

- Let $I_{0,1}$ be the smallest value of I such that $c_{0,I} \in (\alpha_{(m)}, \infty)$. (Such a value of I, here and later in this list, will usually exist in view of the density requirement (21).) Then the widest non-trivial prediction interval is

$$\Gamma^{\mathrm{SIRP}(m,0,I_{0,1})} = \left(\hat{y}_{n+1} - c_{0,I_{0,1}}, \hat{y}_{n+1} + c_{0,I_{0,1}}\right).$$

- For $j = 2, 3, \dots$, let $I_{0,j}$ be the smallest value of I such that $c_{0,I} \in (\alpha_{(m)}, c_{0,I_{0,j-1}})$. Then the following prediction intervals are

$$\Gamma^{\mathrm{SIRP}(m,0,I_{0,j})} = \left(\hat{y}_{n+1} - c_{0,I_{0,j}}, \hat{y}_{n+1} + c_{0,I_{0,j}}\right).$$

- For $K = 1, \dots, m-1$, let $I_{K,1}$ be the smallest value of I such that $c_{K,I} \in (\alpha_{(m-K)}, \alpha_{(m-K+1)})$. Then

$$\Gamma^{\mathrm{SIRP}(m,K,I_{K,1})} = \left(\hat{y}_{n+1} - c_{K,I_{K,1}}, \hat{y}_{n+1} + c_{K,I_{K,1}}\right).$$

- Finally, for $K = 1, \dots, m-1$ and $j = 2, 3, \dots$, let $I_{K,j}$ be the smallest value of I such that $c_{K,I} \in (\alpha_{(m-K)}, c_{K,I_{K,j-1}})$. Then the remaining prediction intervals are

$$\Gamma^{\mathrm{SIRP}(m,K,I_{K,j})} = \left(\hat{y}_{n+1} - c_{K,I_{K,j}}, \hat{y}_{n+1} + c_{K,I_{K,j}}\right).$$

If the required value I does not exist in any of these items, the corresponding prediction interval $\Gamma^{\mathrm{SIRP}(m,K,I_{K,j})}$ and any $\Gamma^{\mathrm{SIRP}(m,K,I_{K,j'})}$ for $j' > j$ are undefined.

In the context of Example 7, informal design principles for the threshold array $(c_{K,I})$ are: we would like $c_{0,I}$ to be concentrated right above the typical values of the largest calibration nonconformity score $\alpha_{(m)}$; we would like $c_{K,I}$, $K = 1, 2, \dots$, to be concentrated mostly inside a typical interval $(\alpha_{(m-K)}, \alpha_{(m-K+1)})$ and closer to $\alpha_{(m-K)}$ for small I. To calculate the likely intervals $(\alpha_{(m)}, \infty)$ and $(\alpha_{(m-K)}, \alpha_{(m-K+1)})$ we may use the proper training sequence.

Now let us check that any SIRP is a bona fide IRP.

Lemma 8. *Every SIRP is an IRP.*

Proof. We are required to check that (22) is a p-value. For a given inductive nonconformity measure A, we will:

- construct a nested family of events $E_{K,I}$ in $\mathbf{S}^{m+1}$ covering the whole sample space $\mathbf{S}^{m+1}$, with the lexicographic order on $(K, I) \in \{0, \dots, m-1\} \times \mathbb{N}_1$; formally, the requirement of being nested means that $E_{K,I} \subseteq E_{K',I'}$ whenever (K, I) comes earlier than (K', I') in the lexicographic order;
- upper bound each $\mathbb{P}^{\mathrm{R}}(E_{K,I})$ by $p_{K,I}$;
- and finally define the SIRP via the aggregating p-variable

$$P(\alpha_{l+1}, \dots, \alpha_{n+1}) := \begin{cases} p_{0,1} & \text{if } (\alpha_{l+1}, \dots, \alpha_{n+1}) \in E_{0,1} \\ p_{K,I} & \text{if } (\alpha_{l+1}, \dots, \alpha_{n+1}) \in E_{K,I} \setminus E_{K,I-1} \text{ for } I > 1 \text{ and any } K \\ p_{K,1} & \text{if } (\alpha_{l+1}, \dots, \alpha_{n+1}) \in E_{K,1} \setminus \cup_{I \in \mathbb{N}_1} E_{K-1,I} \\ & \qquad \text{for } I > 1 \text{ and any } K > 0. \end{cases}$$

(Using general linear orders of this kind when defining p-variables is discussed in detail in [10].)

The innermost nested set $E_{0,1}$ is defined as the event that $c_{0,1}$ separates the test nonconformity score from the calibration nonconformity scores: $\alpha_{n+1} \geq c_{0,1}$ while $\alpha_i < c_{0,1}$ for all $i \in \{l+1, \dots, n\}$. The probability of this event under randomness is $p_0^m p_1$, where p_0 is the probability that $A(z_1, \dots, z_l, Z) \in (-\infty, c_{0,1})$ and p_1 is the probability that $A(z_1, \dots, z_l, Z) \in [c_{0,1}, \infty)$. This allows us to define

$$\mathrm{SIRP}(m, 0, 1) = \max_{(p_0,p_1) \in \Delta_1} p_0^m p_1,$$

in agreement with (22).

For a given $I \in \mathbb{N}_1$, the event $E_{0,I}$ is defined as one of $c_{0,1}, \dots, c_{0,I}$ separating the test nonconformity score from the calibration nonconformity scores. Let $c_{(1)}, \dots, c_{(I)}$ be the sequence $c_{0,1}, \dots, c_{0,I}$ sorted in the ascending order; we extend it by setting $c_{(0)} := -\infty$ and $c_{(I+1)} := \infty$. The probability of the conjunction of the separation and the test nonconformity score lying in $[c_{(i)}, c_{(i+1)})$ is equal to $(p_0 + \dots + p_{i-1})^m p_i$, where p_j is the probability of $A(z_1, \dots, z_l, Z) \in [c_{(j)}, c_{(j+1)})$. This allows us to set

$$\mathrm{SIRP}(m, 0, I) = \max_{(p_0, \dots, p_I) \in \Delta_I} \left(p_0^m p_1 + (p_0 + p_1)^m p_2 + \dots + (p_0 + \dots + p_{I-1})^m p_I \right),$$

which again agrees with (22).

Now we assume $K \geq 1$. Let us say that $c \in \mathbb{R}$ *K-separates* the test nonconformity score α_{n+1} from the calibration nonconformity scores $\alpha_{l+1}, \dots, \alpha_n$ if $\alpha_{n+1} \geq c$ and there are exactly K α_i, $i \in \{l+1, \dots, n\}$, such that $\alpha_i \geq c$ (so that separation, as defined above, corresponds to 0-separation). The event $E_{K,I}$ is defined as the disjunction of the conformal p-value being at most $K/(m+1)$ and the test nonconformity score being K-separated from the calibration nonconformity scores by an element of the set $\{c_{K,1}, \dots, c_{K,I}\}$.

We proceed by induction in K, assuming that (22) works for $K' < K$ in place of K. We also assume that the nonconformity score $A(z_1, \dots, z_l, Z)$ has a continuous distribution; this does not lead to any loss of generality, as will be explained later. The event $\cup_I E_{K-1,I}$ coincides with the conformal p-value being at most $K/(m+1)$, since the threshold array was supposed to be dense in $\mathbf{S}$ for each K.

The first addend in (22) corresponds to the probability of $\cup_I E_{K-1,I}$ under any continuous power probability measure Q^{n+1} on $\mathbf{S}^{m+1}$. Let us check that the term in the second line of (22) corresponds to the probability of the event

$$E'_{K,I} := E_{K,I} \setminus \cup_I E_{K-1,I}$$

that an element of $c_{K,1}, \dots, c_{K,I}$ (or equivalently, of $c_{(1)}, \dots, c_{(I)}$, which are $c_{K,1}, \dots, c_{K,I}$ rearranged in the ascending order, as above) K-separates the test nonconformity score (from the calibration nonconformity scores). The index i in (22) stands for the part of $E'_{K,I}$ corresponding to $\alpha_{n+1} \in [c_{(i)}, c_{(i+1)})$, and the index k stands for the part of that part corresponding to there being exactly k calibration nonconformity scores α_j, $j \in \{l+1, \dots, n\}$, such that $\alpha_j \in [c_{(i)}, c_{(i+1)})$ and $\alpha_j \geq \alpha_{n+1}$. The second line of (22) is obtained by the multiplication of several terms:

- the probability that exactly $m - K$ calibration nonconformity scores are below $c_{(i)}$ is
$$\binom{m}{m-K} \left(\sum_{j=0}^{i-1} p_j \right)^{m-K};$$

- the probability that exactly $K-k$ of the remaining K calibration nonconformity scores are above $c_{(i+1)}$ is

$$\binom{K}{K-k}\left(\sum_{j=i+1}^{I} p_j\right)^{K-k};$$

- the probability that the remaining k calibration nonconformity scores and the test nonconformity score are in $[c_{(i)}, c_{(i+1)})$ is p_i^{k+1};
- the conditional probability (given the event in the previous item) that all those k calibration nonconformity scores are above the test nonconformity score is

$$\int_0^1 x^k \,\mathrm{d}x = \frac{1}{k+1}.$$

Let us check that we can make the assumption of continuity of the probability measure generating nonconformity scores without loss of generality. By [8, Lemma A.23] the calibration and test nonconformity scores, as long as we are interested in their joint distribution, can be assumed to be obtained by applying the same increasing function to IID random variables $\xi_{l+1}, \dots, \xi_{n+1}$ distributed uniformly in $[0,1]$. We can compute the p-values from $\xi_{l+1}, \dots, \xi_{n+1}$ in place of $\alpha_{l+1}, \dots, \alpha_{n+1}$, in which case the values (22) can only decrease. Since even these smaller values are p-values, the original values are p-values as well.

The simplified expression (24) follows from (22), and it can also be interpreted directly. □

To apply a SIRP predictor, we need the function SIRP of three variables, m, K, and I, defined by (22). Hopefully, for sizeable m the dependence on m will be very predictable; we find a few asymptotic expressions in the following proposition. If SIRPs are ever used in practice, it makes sense to make the sequence $c_{K,1}, c_{K,2}, \dots$ finite and short for each K. We can say least about the dependence on K.

Proposition 9. *The function* $\mathrm{SIRP}(m,K,I)$ *defined by* (22) *has the following properties.*

- *It is increasing in* K *and* I*,*

$$\mathrm{SIRP}(m,K,I) \in \left(\frac{K}{m+1}, \frac{K+1}{m+1}\right], \text{ and} \tag{25}$$

$$\mathrm{SIRP}(m,K,\infty) = \frac{K+1}{m+1}. \tag{26}$$

- *Finally,*

$$\mathrm{SIRP}(m,0,2) \sim \frac{\exp(\mathrm{e}^{-1}-1)}{m} \approx \frac{0.531}{m} \text{ as } m \to \infty. \tag{27}$$

The limit (26) as $I \to \infty$ corresponds to ignoring the parameter I in sequential inductive randomness prediction, i.e., to inductive conformal prediction. The approximation 0.531 in (27) roughly agrees with the value 5.53% given in Table 2 (when $m = 19$, that value becomes 5.42%, and so the agreement becomes better).

Proof of Proposition 9. The monotonicity of SIRP is obvious. The first equality, (26), follows from our density assumption (21).

Let us check (27). For $K = 0$ and $I = 2$ our optimization problem (23) can be written as

$$p_0^m(1 - p_0 - p_2) + (1 - p_2)^m p_2 \to \max \tag{28}$$

(after substituting $1 - p_0 - p_2$ for p_1). Setting the partial derivatives of the objective function in p_0 and p_2 to 0 we obtain

$$p_0 = 1 + \frac{\mathrm{e}^{-1} - 2}{m} + O(m^{-2}), \quad p_2 = \frac{1 - \mathrm{e}^{-1}}{m} + O(m^{-2})$$

(so that $p_0 + p_2 < 1$ asymptotically, as it should). Plugging this into the objective function in (28) gives

$$\begin{aligned}
\mathrm{SIRP}(m, 0, 2) &= \left(1 + \frac{\mathrm{e}^{-1} - 2}{m} + O(m^{-2})\right)^m \left(\frac{1}{m} + O(m^{-2}\right) \\
&\quad + \left(1 + \frac{\mathrm{e}^{-1} - 1}{m} + O(m^{-2})\right)^m \left(\frac{1 - \mathrm{e}^{-1}}{m} + O(m^{-2}\right) \\
&= \frac{\exp(\mathrm{e}^{-1} - 2)}{m} + \frac{\exp(\mathrm{e}^{-1} - 1)(1 - \mathrm{e}^{-1})}{m} + O(m^{-2}) \\
&= \frac{\exp(\mathrm{e}^{-1} - 1)}{m} + O(m^{-2}). \quad \square
\end{aligned}$$

Now let us state formally that the SIRP based on an inductive nonconformity measure A dominates the ICP based on A as corollary of Proposition 9. It is then obvious than the domination is usually strict, which once again demonstrates the inadmissibility of typical ICPs.

Corollary 10. *Let A be an inductive nonconformity measure. The SIRP based on A dominates the ICP based on A.*

Proof. The statement of the corollary follows from (25). $\square$

However, even SIRPs are typically inadmissible and strictly dominated by a calibration-invariant IRP. Indeed, take any SIRP and any sequence $\alpha_{l+1}, \dots, \alpha_{n+1}$ of distinct nonconformity scores such that α_{n+1} is the largest number in this sequence and $c_{0,1}$ separates it from the calibration nonconformity scores. The maximum power probability Q^{n+1} of the set

$$\left\{(\alpha_{\pi(l+1)}, \dots, \alpha_{\pi(n)}, \alpha_{n+1}) \mid \pi \in \mathrm{Sym}(\{l+1, \dots, n\})\right\} \subseteq \mathbf{S}^{m+1} \tag{29}$$

is

$$\frac{m!}{(m+1)^{m+1}} \sim \sqrt{2\pi/m}\, \mathrm{e}^{-m-1},$$

which is much smaller, for a large m, than the smallest p-value attainable by a SIRP. Therefore, we can improve the given SIRP by redefining the p-value on the set (29).

Remark 11. The non-trivial second addend in the function (22) is defined as the maximum of a homogenous polynomial of degree $m+1$ over the unit simplex Δ_I. This polynomial is not convex in general, as can be seen by differentiating the polynomial $p_0^2 p_1$ that is maximized in $\mathrm{SIRP}(2,0,1)$ (for simplicity, replace p_1 by $1-p_0$). Despite the lack of convexity, this is a well-studied problem. The problem is NP-complete already for quadratic polynomials, but there are PTAS (polynomial-time approximation schemes) for a fixed m. (See [5–7].)

6 Conclusion

In this paper we have defined inductive randomness predictors and started their study. Whereas inductive conformal predictors are inadmissible and are dominated by SIRPs, it remains unclear whether SIRPs, or other dominating inductive randomness predictors, can be more useful in practice. A related theoretical question is how to define a suitable weakened notion of admissibility and apply it usefully to dominating inductive randomness predictors.

A cheap way to improve on ICPs is to use randomization; the resulting *smoothed ICPs* [17, Sect. 4.2.1] dominate the corresponding ICPs. We will discuss smoothed IRPs in Appendix B, but in this version of the paper we will concentrate on smoothed BIRPs. Allowing randomization in SIRPs is potentially more interesting.

Acknowledgments

Many thanks to Alexander Shen for his advice. Computational experiments in this paper used Wolfram Mathematica and the SciPy Python library.

References

[1] Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. Technical Report arXiv: 2411.11824 [math.ST], arXiv.org e-Print archive, November 2024. Prepublication version of a book to be published by Cambridge University Press.

[2] Henrik Boström. Conformal prediction in Python with crepes. *Proceedings of Machine Learning Research*, 230:236–249, 2024. COPA 2024.

[3] Thibault Cordier, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, and Nicolas Brunel. Flexible and systematic uncertainty estimation with conformal prediction via the MAPIE library. *Proceedings of Machine Learning Research*, 204:549–581, 2023. COPA 2023.

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

[5] Etienne de Klerk, Monique Laurent, and Pablo A. Parrilo. A PTAS for the minimization of polynomials of fixed degree over the simplex. *Theoretical Computer Science*, 361:210–225, 2006.

[6] Etienne de Klerk, Monique Laurent, and Zhao Sun. An error analysis for polynomial optimization over the simplex based on the multivariate hypergeometric distribution. *SIAM Journal on Optimization*, 25:1498–1514, 2015.

[7] Etienne de Klerk, Monique Laurent, Zhao Sun, and Juan C. Vera. On the convergence rate of grid search for polynomial optimization over the simplex. *Optimization Letters*, 11:597–608, 2017.

[8] Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter, Berlin, fourth edition, 2016.

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

[10] Yuri Gurevich and Vladimir Vovk. Test statistics and p-values. *Proceedings of Machine Learning Research*, 105:89–104, 2019. COPA 2019.

[11] Alexander Shen, Vladimir A. Uspensky, and Nikolai Vereshchagin. *Kolmogorov Complexity and Algorithmic Randomness*. American Mathematical Society, Providence, RI, 2017.

[12] Stephen M. Stigler. The epic story of maximum likelihood. *Statistical Science*, 22:598–620, 2007.

[13] Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström, and Lars Carlsson, editors. *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*. PMLR, 2024.

[14] Vladimir Vovk. Superefficiency from the vantage point of computability. *Statistical Science*, 24:73–86, 2009.

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

[16] Vladimir Vovk. Set and functional prediction: randomness, exchangeability, and conformal. Technical Report arXiv:2502.19254 [cs.LG], arXiv.org e-Print archive, February 2025.

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

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

A Universal threshold array

It is not clear how to choose the threshold array for use in sequential inductive randomness prediction, and in this appendix we will discuss the universal choice adapting the notion of Kolmogorov complexity (see, e.g., [11, Sect. 1.1]). The usual notion of Kolmogorov complexity is binary, but here it will be more convenient to use the unary version (binary complexity coincides, to within an additive constant, with the binary logarithm of unary complexity).

A *description mode* is a computably enumerable set D of quintuples $(m, K, I, a, b) \in \mathbb{N}_1 \times \mathbb{N}_0 \times \mathbb{N}_1 \times \mathbb{Q}^2$ such that $K < m$ and $a < b$, where $\mathbb{Q}$ is the set of rational numbers. The *unary (Kolmogorov) D-complexity* $U_D(c \mid m, K)$ of a real number c given m and K is the smallest I such that

$$\{c\} = \cap\{(a, b) \mid (m, K, I, a, b) \in D\}. \tag{30}$$

There is a description mode D (called *universal*) such that for any other description mode D' there exists $B > 0$ such that

$$\forall c \in \mathbb{R}, m \in \mathbb{N}_1, K \in \{0, \dots, m-1\} : U_D(c \mid m, K) \le BU_{D'}(c \mid m, K).$$

Let us fix a universal description mode D and call $U(c \mid m, K) := U_D(c \mid m, K)$ the *unary complexity* of c given m and K. For any set $C \subseteq \mathbb{R}$ of real numbers define

$$U(C \mid m, K) := \inf_{c \in C} U(c \mid m, K).$$

For simplicity, let us fix m. With the universal description mode D we can associate the threshold array $(c_{K,I})$ (*universal threshold array*) whose elements are allowed to take value ∞; namely, we define $c_{K,I}$ as the only element of the set on the right-hand side of (30) if that set is a singleton; otherwise, we set $c_{K,I} := \infty$. Let $\alpha_{l+1}, \dots, \alpha_{n+1}$ be a sequence of calibration nonconformity scores extended by a test nonconformity score. Then the SIRP based on $(c_{K,I})$ and fed with these nonconformity scores outputs $\mathrm{SIRP}(m, K, U((\alpha_{n+1}, \alpha'])))$ as its p-value, where

$$\begin{aligned} K &:= \left|\{i \in \{1, \dots, n\} \mid \alpha_i \ge \alpha_{n+1}\}\right|, \\ \alpha' &:= \min\{\alpha_i \mid i \in \{l+1, \dots, n\} \,\&\, \alpha_i \ge \alpha_{n+1}\} \end{aligned}$$

(if $\alpha' = \alpha_{n+1}$, the SIRP p-value is simply the conformal p-value $(K+1)/(m+1)$).

The expression $\mathrm{SIRP}(m, K, U((\alpha_{n+1}, \alpha']))$ illustrates the difference between the conformal and SIRP p-values. The conformal p-value $(K+1)/(m+1)$ only depends on the order of nonconformity scores. The SIRP p-value depends, additionally, on how easy it is to separate the test nonconformity score from the K largest calibration nonconformity scores. The size of the margin of separation $(\alpha_{n+1}, \alpha']$ is measured by its unary complexity.

We can regard approximating the universal threshold array to be an informal design principle for threshold arrays. Ideally, in the definition of $U(c \mid m, K)$ we should condition, in addition to the length m of the calibration sequence, on all other known relevant features of our prediction problem, such as the proper training sequence and the chosen inductive nonconformity measure. The universal threshold array will then satisfy the informal design principles discussed after Example 7.

B Smoothed BIRPs

In the main part of the paper we only discussed deterministic predictors, while randomized ("smoothed") conformal predictors [17, Sect. 2.2.6] produce smaller p-values and, therefore, are more predictively efficient. Adding randomization to prediction procedures is often regarded as objectionable, and so discussing randomized predictors is relegated to this appendix. Randomization significantly complicates discussions of predictive efficiency and admissibility.

The *smoothed inductive conformal predictor* (SICP) based on an inductive nonconformity measure A outputs the prediction p-function

$$f(y) := \frac{|\{j = l+1, \dots, n+1 \mid \alpha_j > \alpha_{n+1}\}|}{m+1} + \tau \frac{|\{j = l+1, \dots, n+1 \mid \alpha_j = \alpha_{n+1}\}|}{m+1} \in [0,1],$$

where the αs are defined as before, by (2) and (3), and $\tau \sim U$ is a random number generated from the uniform probability measure U on $[0,1]$. This will be a special case of smoothed inductive randomness predictors, which we define next.

A *randomized aggregating p-variable* is a measurable function $P : [0,1] \times \mathbf{S}^{m+1} \to [0,1]$ such that

$$\forall \epsilon \in (0,1)\ \forall Q \in \mathfrak{P}(\mathbf{S}) : (U \times Q^{m+1})(\{P \le \epsilon\}) \le \epsilon.$$

The *smoothed inductive randomness predictor* (SIRP) based on an inductive nonconformity measure A and a randomized aggregating p-variable P is defined, similarly to the IRP, by

$$P_A(\tau, z_1, \dots, z_{n+1}) := P(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}),$$

where the αs are defined by (6). Instead of (7), the SIRP P_A outputs the prediction p-function

$$f(y) = f(y; \tau, z_1, \dots, z_n, x_{n+1}) := P_A(\tau, z_1, \dots, z_n, x_{n+1}, y),$$

where $\tau \sim U$. To embed the class of SICPs into the class of SIRPs, we set, analogously to (9),

$$\Pi(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}) := \frac{|\{j = l+1, \dots, n+1 \mid \alpha_j > \alpha_{n+1}\}|}{m+1} + \tau \frac{|\{j = l+1, \dots, n+1 \mid \alpha_j = \alpha_{n+1}\}|}{m+1}.$$

Then the SICP based on A is identical to the SIRP Π_A.

A convenient way to generate randomized aggregating p-variables is to use aggregating functions $B : \mathbf{S}^{m+1} \to \mathbb{R}$, as defined earlier. The corresponding aggregating p-variable will be the following variation on (10):

$$P_B(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}) := \sup_{Q \in \mathfrak{P}(\mathbf{S})} \Big(Q^{m+1}(\{B > B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1})\}) + \tau Q^{m+1}(\{B = B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1})\}) \Big). \quad (31)$$

Proposition 12. *The function P_B defined by* (31) *is a randomized aggregating p-variable.*

Proof. Let us define a function $g : \mathbb{R} \times [0,1] \to [0,1]$ by

$$g(b, \tau) := \sup_{Q \in \mathfrak{P}(\mathbf{S})} \left(Q^{m+1}(\{B > b\}) + \tau Q^{m+1}(\{B = b\}) \right)$$

(cf. (31)). It is clear that $g(b, \tau)$ is decreasing in b and increasing in τ. It also satisfies the following two useful properties.

Lemma 13. *For all b, $g(b, 0) = \sup_{b' > b} g(b', 1) = \sup_{b' > b} g(b', 0)$.*

Proof. Take any $\delta > 0$. Choose $Q \in \mathfrak{P}(\mathbf{S})$ such that $Q^{m+1}(\{B > b\}) > g(b, 0) - \delta$. Then, for some $b' > b$, $Q^{m+1}(\{B \geq b'\}) > g(b, 0) - \delta$. Finally, the last inequality implies $g(b', 1) > g(b, 0) - \delta$. □

Lemma 14. *As function of τ, $g(b, \tau)$ is continuous.*

Proof. Suppose $g(b, \cdot)$ makes a jump at some point $\tau_0 \in [0, 1]$. For an arbitrarily small $\delta > 0$, take any $\tau_1 \in [\tau_0, \tau_0 + \delta]$ and choose $Q \in \mathfrak{P}(\mathbf{S})$ satisfying

$$Q^{m+1}(\{B > b\}) + \tau_1 Q^{m+1}(\{B = b\}) > g(b, \tau_1) - \delta.$$

Then, for any $\tau_2 \in [\tau_0 - \delta, \tau_0]$,

$$g(b, \tau_2) \geq Q^{m+1}(\{B > b\}) + \tau_2 Q^{m+1}(\{B = b\})$$

$$\begin{aligned}
&\geq Q^{m+1}(\{B > b\}) + \tau_1 Q^{m+1}(\{B = b\}) - 2\delta \\
&> g(b, \tau_1) - 3\delta.
\end{aligned}$$

Since δ can be arbitrarily small, the inequality between the extreme terms of this chain leads to a contradiction. □

Now we can prove the statement of the proposition. Fix $\epsilon \in (0, 1)$ and set

$$b := \inf\{b' \mid g(b', 0) \leq \epsilon\}.$$

By Lemma 13, $g(b, 0) \leq \epsilon$, and we know that $g(b', 0) > \epsilon$ for all $b' < b$. Let us consider two cases.

First we consider the presumably typical case where $g(b, 0) \leq \epsilon \leq g(b, 1)$. Choose τ_0 satisfying $g(b, \tau_0) = \epsilon$. Make it as large as possible if such τ_0 is not unique (this step uses Lemma 14). Then the set $\{P_B \leq \epsilon\}$ consists of $(\tau, \alpha_{l+1}, \dots, \alpha_{n+1})$ at which $B > b$ or both $B = b$ and $\tau \leq \tau_0$. The supremum $U \times Q^{m+1}$-probability of this set is $g(b, \tau_0) = \epsilon$.

It remains to consider the case $g(b, 0) \leq g(b, 1) < \epsilon$. Then the set $\{P_B \leq \epsilon\}$ consists of $(\tau, \alpha_{l+1}, \dots, \alpha_{n+1})$ at which $B \geq b$. The supremum $U \times Q^{m+1}$-probability of this set is $g(b, 1) < \epsilon$. □

Remark 15. Proposition 12 is applicable to any statistical model, not just the *randomness model* $\{Q^{m+1} \mid Q \in \mathfrak{P}(\mathbf{S})\}$.

We will say that the SIRP $P_{A,B} := (P_B)_A$ is *based on A and B*, where A is an inductive nonconformity measure and B is an aggregating function.

Proposition 4 can be generalized to the smoothed case, but the calculations become messier for $K > 1$.

Proposition 16. *Suppose that a binary sequence $\alpha_{l+1}, \dots, \alpha_n$ contains $K < m$ 1s and that $\alpha_{n+1} = 1$. Then the aggregating function B defined by* (12) *leads to the smoothed p-value*

$$P_B(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}) = \max_{p \in [0,1]} \left(\sum_{k=0}^{K-1} \binom{m}{k} p^{k+1} (1-p)^{m-k} + \tau \binom{m}{K} p^{K+1} (1-p)^{m-K} \right). \quad (32)$$

In particular, for $K = 0$, the smoothed p-value (32) *is*

$$\tau \frac{m^m}{(m+1)^{m+1}} \sim \tau \frac{\exp(-1)}{m} \approx \frac{0.37\tau}{m}.$$

Let us check that SICPs are inadmissible. As in the proof of Proposition 5, we can improve Π_A to P_A, where

$$P(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}) := \begin{cases} \tau \frac{m^m}{(m+1)^{m+1}} & \text{if } \alpha_{n+1} > a \text{ and } \alpha_i < a \text{ for all } i \in \{l+1, \dots, n\} \\ \Pi(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}) & \text{otherwise.} \end{cases}$$

Checking the domination reduces to checking the inequality

$$\tau \frac{m^m}{(m+1)^{m+1}} < \frac{\tau}{m+1},$$

which is obvious.

We have not discussed smoothed SIRPs, and, as mentioned in Sect. 6, this is an interesting direction of further research.