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 9, 2025.

Project web site:
http://alrw.net

Abstract

This paper introduces inductive randomness predictors, which form a superset of inductive conformal predictors. Its focus is on a very simple special case, binary inductive randomness predictors. It is interesting that binary inductive randomness predictors have an advantage over inductive conformal predictors, although they also have a serious disadvantage. This advantage will allow us to reach the surprising conclusion that non-trivial inductive conformal predictors are inadmissible in the sense of statistical decision theory.

Contents

1 Introduction

Randomness predictors were introduced and studied in [9]. 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., [7]), and studied (see, e.g., [1]). Both [9] and the follow-up paper [10] concentrate on negative results about randomness predictors, showing that the difference in predictive efficiency between conformal and randomness prediction is not great. (While [9] covers worst-case difference, [10] also treats difference on average.) This paper concentrates, instead, on positive results, giving examples of situations were randomness predictors have a clear 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 [12], is that the p-values that they output can never drop below $\frac{1}{n+1}$, where n is the size of the training set. 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. This paper introduces and studies inductive randomness predictors, which are also computationally efficient.

We will start in Sect. 2 from the main definitions, including that of inductive randomness predictors, and two examples of inductive randomness predictors. Section 3 is devoted to computing binary inductive randomness predictors. The topic of Sect. 4 is the inadmissibility of inductive conformal predictors as inductive randomness predictors. The short Sect. 5 concludes.

2 Definitions

The prediction problem considered in this paper is the same as in [9, 10]. We are given a training sequence $z_1, \dots, z_n$, where $z_i = (x_i, y_i)$ 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 [11, 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 $l \geq 1$ and $m \geq 1$. An *inductive nonconformity measure* is a measurable function $A : \mathbf{Z}^{l+1} \to \mathbb{R}$. The *inductive conformal predictor* (ICP) based on A outputs the prediction p-function

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

where the αs are defined by

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

$$\alpha_{n+1} := A(z_1, \dots, z_l, x_{n+1}, y) \tag{2}$$

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 [11, Sect. 9.1.1] as

$$\mathbb{P}^{\mathrm{R}}(E) := \sup_{Q \in \mathfrak{P}(\mathbf{Z})} Q^{n+1}(E), \tag{3}$$

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 (3), 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

$$\forall \epsilon \in (0,1) : \mathbb{P}^{\mathrm{R}}(\{P \leq \epsilon\}) \leq \epsilon. \tag{4}$$

A *randomness predictor*, as defined in [9, 10], is a p-variable $P : \mathbf{Z}^{n+1} \to [0,1]$, meaning that it is required to satisfy (4).

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

$$\alpha_j := A(z_1, \dots, z_l, z_j), \quad j = l+1, \dots, n+1. \tag{5}$$

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

$$f(y) = f(y; z_1, \ldots, z_n, x_{n+1}) := P_A(z_1, \ldots, z_n, x_{n+1}, y). \tag{6}$$

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

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

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 (6) 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

$$\Pi(\alpha_{l+1}, \ldots, \alpha_{n+1}) := \frac{|\{j = l+1, \ldots, n+1 \mid \alpha_j \geq \alpha_{n+1}\}|}{m+1}, \\ (\alpha_{l+1}, \ldots, \alpha_{n+1}) \in \mathbf{S}^{m+1}. \tag{7}$$

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

$$P_B(\alpha_{l+1}, \ldots, \alpha_{n+1}) := \mathbb{P}^{\mathrm{R}}\left(\{B \geq B(\alpha_{l+1}, \ldots, \alpha_n, \alpha_{n+1})\}\right), \\ (\alpha_{l+1}, \ldots, \alpha_{n+1}) \in \mathbf{S}^{m+1}. \tag{8}$$

(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.

In this paper we will concentrate mainly on *binary inductive randomness predictors*, for which the summary space is $\mathbf{S} := \{0, 1\}$. Intuitively, a summary of 0 means conformity, and 1 means lack of conformity. Let me give two examples of binary IRPs, 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

$$A(z_1, \dots, z_l, x, y) := \begin{cases} 1 & \text{if } |y - \hat{g}(x)| > \max_{i=1,\dots,l} |y_i - \hat{g}(x_i)| \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

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

$$B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1}) := \max \left(\alpha_{n+1} - \frac{1}{n} \sum_{i=1}^{n} \alpha_i, 0 \right). \tag{10}$$

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 $z_1, \dots, z_l$. Next we compute the half-width $h := \max_{i=1,\dots,l} |y_i - \hat{g}(x_i)|$ 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 $A(z_1, \dots, z_l, x, y)$, consider the support vector machine (SVM) constructed from $z_1, \dots, z_l$ as training set. Set $A(z_1, \dots, z_l, x, y)$ to 1 if x is classified incorrectly (namely, as $-y$) by this SVM and x is outside its margin; set $A(z_1, \dots, z_l, x, y)$ to 0 otherwise. A reasonable definition of B is still (10).

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 $A(z_1, \dots, z_l, z_i) = 0$ for some $i \in \{l + 1, \dots, n\}$. 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 $A(z_1, \dots, z_l, x, y)$ to 1 if x is a support vector for the SVM constructed from $(z_1, \dots, z_l, x, y)$ as training set and to set it to 0 otherwise, as in [5, 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.

3 Binary inductive randomness predictors

In this section we will compute the p-values output by binary IRPs based on the aggregating function (10). 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 $\alpha_{l+1}, \dots, \alpha_n$ contains $k < m$ 1s and that $\alpha_{n+1} = 1$. Then the aggregating function B defined by* (10) *leads to a p-value $P_B(\alpha_{l+1}, \dots, \alpha_{n+1})$ of*

$$\max_{p \in [0,1]} \sum_{i=0}^{k} \binom{m}{i} p^{i+1} (1-p)^{m-i}. \tag{11}$$

In particular,

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

$$\frac{m^m}{(m+1)^{m+1}} \sim \frac{\exp(-1)}{m} \approx \frac{0.37}{m}, \tag{12}$$

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

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

$$\frac{(\phi + \phi^2) \exp(-\phi)}{m} \approx \frac{0.84}{m}, \tag{13}$$

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

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

$$\frac{(c + c^2 + c^3/2) \exp(-c)}{m} \approx \frac{1.37}{m}, \tag{14}$$

  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*

$$\frac{(c + c^2 + c^3/2 + c^4/6)\exp(-c)}{m} \approx \frac{1.94}{m}, \tag{15}$$

where

$$c := \frac{1}{4} + \frac{1}{4}\Bigl(4(\sqrt{778} - 7)^{1/3} - 36(\sqrt{778} - 7)^{-1/3} + 9\Bigr)^{1/2} + \frac{1}{2}\Biggl(-(\sqrt{778} - 7)^{1/3} + 9(\sqrt{778} - 7)^{-1/3} + \frac{9}{2} + \frac{61}{2\sqrt{4(\sqrt{778} - 7)^{1/3} - 36(\sqrt{778} - 7)^{-1/3} + 9}}\Biggr)^{1/2}.$$

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 (9) and (10) will be more confident than the identical prediction interval output by the ICP based on the same inductive nonconformity measure (9): 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 (9) and (10) is still close to $0.84/m$ (see (13)), 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 (9) and (10) will be significantly better than the incertitude of the ICP based on (9) (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 (11).

When $k = 0$, $\max_p p(1 - p)^m$ is attained at $p = \frac{1}{m+1}$, which leads to (12). The inequality

$$\frac{m^m}{(m + 1)^{m+1}} \le \frac{\exp(-1)}{m} \tag{16}$$

is equivalent to

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

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

and is easy to check.

When $k = 1$, solving the optimization problem

$$p(1-p)^m + mp^2(1-p)^{m-1} \to \max \tag{17}$$

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 (17) gives (13).

Now let us deal with an arbitrary (but fixed k) and let $m \to \infty$. The optimal value of p in (11) 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 (11), we can see that this expression is asymptotically equivalent to

$$\sum_{i=0}^{k} \frac{c^{i+1}\mathrm{e}^{-c}}{i!\,m}. \tag{18}$$

This gives the left-hand sides of (14) and (15). Setting the derivative of (18) to 0, we can check that the optimal c satisfies the equation

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

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 (12)–(15) for a wide range of k. The IRP is based on (9) and (10), and the ICP is based on (9). 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.

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 (16), 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 (12).

It is also clear that P_A strictly dominates Π_A. □

The phenomenon of inadmissibility of ICPs demonstrated by this proof is akin to the phenomenon of superefficiency in point estimation (see, e.g., [6] and [8, Sect. 2] for reviews). We are making an ICP superefficient at a nonconformity score that we choose arbitrarily, as in Hodges's example [6, Fig. 1]. It seems that in such situations the standard term "inadmissibility" becomes too harsh.

5 Conclusion

In this paper we have defined inductive randomness predictors and started their study. Whereas inductive conformal predictors are strongly inadmissible, it remains an open question whether there are dominating inductive randomness predictors that are clearly more useful. If there are, can we make such inductive randomness predictors weakly admissible?

Acknowledgments

Computational experiments in this paper used Wolfram Mathematica.

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] 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.

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

[7] 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.

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

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

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

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

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

A Smoothed predictors

In the main part of the paper we only discussed deterministic predictors, while randomized ("smoothed") conformal predictors [11, 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 (1) and (2), 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 (5). Instead of (6), 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 (7),

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

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 (8):

$$P_B(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}) := \sup_{Q \in \mathfrak{P}(\mathbf{S})} \Big(Q^{m+1}\left(\{B > B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1})\}\right) + \tau Q^{m+1}\left(\{B = B(\alpha_{l+1}, \dots, \alpha_n, \alpha_{n+1})\}\right) \Big). \quad (19)$$

Proposition 6. *The function P_B defined by* (19) *is a randomized aggregating p-variable.*

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

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

(cf. (19)). It is clear that $g(b, \tau)$ is decreasing in b and increasing in τ. It also satisfies the following two useful properties.

Lemma 7. *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 8. *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 $\tau_2 \in [\tau_0 - \delta, \tau_0]$,

$$\begin{aligned} g(b, \tau_2) &\geq Q^{m+1}(\{B > b\}) + \tau_2 Q^{m+1}(\{B = b\}) \\ &\geq Q^{m+1}(\{B > b\}) + \tau_1 Q^{m+1}(\{B = b\}) - 2\delta \\ &> g(b, \tau_1) - 3\delta. \end{aligned}$$

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

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

By Lemma 7, $g(b,0) \le \epsilon$, 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 $g(b,0) \le \epsilon \le g(b,1)$. 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 8). Then the set $\{P_B \le \epsilon\}$ consists of $(\tau, \alpha_{l+1}, \dots, \alpha_{n+1})$ at which $B > b$ or both $B = b$ and $\tau \le \tau_0$. The supremum $U \times Q^{m+1}$-probability of this set is $g(b,\tau_0) = \epsilon$.

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

Remark 9. Proposition 6 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 10. *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* (10) *leads to the smoothed p-value*

$$P_B(\tau, \alpha_{l+1}, \dots, \alpha_{n+1}) = \max_{p \in [0,1]} \left(\sum_{i=0}^{k-1} \binom{m}{i} p^{i+1} (1-p)^{m-i} + \tau \binom{m}{k} p^{k+1} (1-p)^{m-k} \right). \quad (20)$$

In particular, for $k = 0$, the smoothed p-value (20) *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.