Criteria of efficiency for conformal prediction*

Vladimir Vovk, Valentina Fedorova,
Ilia Nouretdinov, and Alex Gammerman

{volodya.vovk,alushaf}@gmail.com
{ilia,alex}@cs.rhul.ac.uk

March 20, 2016

Abstract

We study optimal conformity measures for various criteria of efficiency in an idealised setting. This leads to an important class of criteria of efficiency that we call probabilistic; it turns out that the most standard criteria of efficiency used in literature on conformal prediction are not probabilistic.

1 Introduction

Conformal prediction is a method of generating prediction sets that are guaranteed to have a prespecified coverage probability; in this sense conformal predictors have guaranteed validity. Different conformal predictors, however, widely differ in their efficiency, by which we mean the narrowness, in some sense, of their prediction sets. Empirical investigation of the efficiency of various conformal predictors is becoming a popular area of research: see, e.g., [1, 11] (and the COPA Proceedings, 2012–2015). This paper points out that the standard criteria of efficiency used in literature have a serious disadvantage, and we define a class of criteria of efficiency, called "probabilistic", that do not share this disadvantage. In two recent papers [3, 5] two probabilistic criteria have been introduced, and in this paper we introduce two more and argue that probabilistic criteria should be used in place of more standard ones. We concentrate on the case of classification only (the label space is finite).

Surprisingly few criteria of efficiency have been used in literature, and even fewer have been studied theoretically. We can speak of the efficiency of individual predictions or of the overall efficiency of predictions on a test sequence; the latter is usually (in particular, in this paper) defined by averaging the efficiency over the individual test examples, and so in this introductory section we only

*A preliminary version of this paper was published as Working Paper 11 of the On-line Compression Modelling project (New Series), http://alrw.net, in April 2014.

discuss the former. This section assumes that the reader knows the basic definitions of the theory of conformal prediction, but they will be given in Section 2, which can be consulted now.

The two criteria for efficiency of a prediction that have been used most often in literature (in, e.g., the references given above) are:

- The confidence and credibility of the prediction (see, e.g., [14], p. 96; introduced in [12]). This criterion does not depend on the choice of a significance level ϵ.
- Whether the prediction is a singleton (the ideal case), multiple (an inefficient prediction), or empty (a superefficient prediction) at a given significance level ϵ. This criterion was introduced in [10], Section 7.2, and used extensively in [14].

The other two criteria that have been used are the sum of the p-values for all potential labels (this does not depend on the significance level) and the size of the prediction set at a given significance level: see the papers [3] and [5].

In this paper we introduce six other criteria of efficiency (Section 2). We then discuss (in Sections 3–5) the conformity measures that optimise each of the ten criteria when the data-generating distribution is known; this sheds light on the kind of behaviour implicitly encouraged by the criteria even in the realistic case where the data-generating distribution is unknown. As we point out in Section 5, probabilistic criteria of efficiency are conceptually similar to "proper scoring rules" in probability forecasting [2, 4], and this is our main motivation for their detailed study in this paper. After that we briefly illustrate the empirical behaviour of two of the criteria for standard conformal predictors and a benchmark data set (Section 6).

We only consider the case of randomised ("smoothed") conformal predictors: the case of deterministic predictors may lead to packing problems without an explicit solution (this is the case, e.g., for the N criterion defined below). The situation here is analogous to the Neyman–Pearson lemma: cf. [7], Section 3.2.

2 Criteria of Efficiency for Conformal Predictors and Transducers

Let $\mathbf{X}$ be a measurable space (the *object space*) and $\mathbf{Y}$ be a finite set equipped with the discrete σ-algebra (the *label space*); the *example space* is defined to be $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$. A *conformity measure* is a measurable function A that assigns to every finite sequence $(z_1, \dots, z_n) \in \mathbf{Z}^*$ of examples a same-length sequence $(\alpha_1, \dots, \alpha_n)$ of real numbers and that is equivariant with respect to permutations: for any n and any permutation π of $\{1, \dots, n\}$,

$$(\alpha_1, \dots, \alpha_n) = A(z_1, \dots, z_n) \Longrightarrow \left(\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}\right) = A\left(z_{\pi(1)}, \dots, z_{\pi(n)}\right).$$

The *conformal predictor* determined by A is defined by

$$\Gamma^\epsilon(z_1, \dots, z_l, x) := \{y \mid p^y > \epsilon\}, \tag{1}$$

where $(z_1, \ldots, z_l) \in \mathbf{Z}^*$ is a training sequence, x is a test object, $\epsilon \in (0,1)$ is a given *significance level*, for each $y \in \mathbf{Y}$ the corresponding *p-value* p^y is defined by

$$p^y := \frac{1}{l+1} \left|\left\{i = 1, \ldots, l+1 \mid \alpha_i^y < \alpha_{l+1}^y\right\}\right| + \frac{\tau}{l+1} \left|\left\{i = 1, \ldots, l+1 \mid \alpha_i^y = \alpha_{l+1}^y\right\}\right|, \quad (2)$$

τ is a random number distributed uniformly on the interval $[0,1]$ (even conditionally on all the examples), and the corresponding sequence of *conformity scores* is defined by

$$(\alpha_1^y, \ldots, \alpha_l^y, \alpha_{l+1}^y) := A(z_1, \ldots, z_l, (x, y)).$$

Notice that the system of *prediction sets* (1) output by a conformal predictor is decreasing in ϵ, or *nested*.

The *conformal transducer* determined by A outputs the system of p-values $(p^y \mid y \in \mathbf{Y})$ defined by (2) for each training sequence $(z_1, \ldots, z_l)$ of examples and each test object x. (This is just a different representation of the conformal predictor.)

The standard property of validity for conformal predictors and transducers is that the p-values p^y are distributed uniformly on $[0,1]$ when the examples $z_1, \ldots, z_l, (x, y)$ are generated independently from the same probability distribution Q on $\mathbf{Z}$ (see, e.g., [14], Proposition 2.8). This implies that the probability of error, $y \notin \Gamma^\epsilon(z_1, \ldots, z_l, x)$, is ϵ at any significance level ϵ.

Suppose we are given a test sequence $(z_{l+1}, \ldots, z_{l+k})$ and would like to use it to measure the efficiency of the predictions derived from the training sequence $(z_1, \ldots, z_l)$. (The efficiency of conformal predictors means that the prediction sets they output tend to be small, and the efficiency of conformal transducers means that the p-values that they output tend to be small.) For each test example $z_i = (x_i, y_i)$, $i = l+1, \ldots, l+k$, we have a nested family $(\Gamma_i^\epsilon \mid \epsilon \in (0,1))$ of subsets of $\mathbf{Y}$ and a system of p-values $(p_i^y \mid y \in \mathbf{Y})$. In this paper we will discuss ten criteria of efficiency for such a family or a system, but some of them will depend, additionally, on the observed labels y_i of the test examples. We start from the *prior* criteria, which do not depend on the observed test labels.

2.1 Basic criteria

We will discuss two kinds of criteria: those applicable to the prediction sets Γ_i^ϵ and so depending on the significance level ϵ and those applicable to systems of p-values $(p_i^y \mid y \in \mathbf{Y})$ and so independent of ϵ. The simplest criteria of efficiency are:

- The *S criterion* (with "S" standing for "sum") measures efficiency by the average sum

$$\frac{1}{k} \sum_{i=l+1}^{l+k} \sum_y p_i^y \quad (3)$$

of the p-values; small values are preferable for this criterion. It is ϵ-free.

- The *N criterion* uses the average size

$$\frac{1}{k}\sum_{i=l+1}^{l+k}|\Gamma_i^\epsilon|$$

of the prediction sets ("N" stands for "number": the size of a prediction set is the number of labels in it). Small values are preferable. Under this criterion the efficiency is a function of the significance level ϵ.

Both these criteria are prior. The S criterion was introduced in [3] and the N criterion was introduced independently in [5] and [3], although the analogue of the N criterion for regression (where the size of a prediction set is defined to be its Lebesgue measure) had been used earlier in [9] (whose arXiv version was published in 2012).

2.2 Other prior criteria

A disadvantage of the basic criteria is that they look too stringent. Even for a very efficient conformal transducer, we cannot expect all p-values p^y to be small: the p-value corresponding to the true label will not be small with high probability; and even for a very efficient conformal predictor we cannot expect the size of its prediction set to be zero: with high probability it will contain the true label. The other prior criteria are less stringent. The ones that do not depend on the significance level are:

- The *U criterion* (with "U" standing for "unconfidence") uses the average unconfidence

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\min_y\max_{y'\neq y}p_i^{y'} \tag{4}$$

over the test sequence, where the *unconfidence* for a test object x_i is the second largest p-value $\min_y\max_{y'\neq y}p_i^{y'}$; small values of (4) are preferable. The U criterion in this form was introduced in [3], but it is equivalent to using the average confidence (one minus unconfidence), which is very common. If two conformal transducers have the same average unconfidence (which is presumably a rare event), the criterion compares the average credibilities

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\max_y p_i^y, \tag{5}$$

where the *credibility* for a test object x_i is the largest p-value $\max_y p_i^y$; smaller values of (5) are preferable. (Intuitively, a small credibility is a warning that the test object is unusual, and since such a warning presents useful information and the probability of a warning is guaranteed to be small, we want to be warned as often as possible.)

- The *F criterion* uses the average fuzziness

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\left(\sum_y p_i^y - \max_y p_i^y\right), \tag{6}$$

where the *fuzziness* for a test object x_i is defined as the sum of all p-values apart from a largest one, i.e., as $\sum_y p_i^y - \max_y p_i^y$; smaller values of (6) are preferable. If two conformal transducers lead to the same average fuzziness, the criterion compares the average credibilities (5), with smaller values preferable.

Their counterparts depending on the significance level are:

- The *M criterion* uses the percentage of objects x_i in the test sequence for which the prediction set Γ_i^ϵ at significance level ϵ is *multiple*, i.e., contains more than one label. Smaller values are preferable. As a formula, the criterion prefers smaller

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\mathbf{1}_{\{|\Gamma_i^\epsilon|>1\}}, \tag{7}$$

where $\mathbf{1}_E$ denotes the indicator function of the event E (taking value 1 if E happens and 0 if not). When the percentage (7) of multiple predictions is the same for two conformal predictors (which is a common situation: the percentage can well be zero), the M criterion compares the percentages

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\mathbf{1}_{\{\Gamma_i^\epsilon=\emptyset\}} \tag{8}$$

of empty predictions (larger values are preferable). This is a widely used criterion. (In particular, it was used in [14] and papers preceding it.)

- The *E criterion* (where "E" stands for "excess") uses the average (over the test sequence, as usual) amount the size of the prediction set exceeds 1. In other words, the criterion gives the average number of excess labels in the prediction sets as compared with the ideal situation of one-element prediction sets. Smaller values are preferable for this criterion. As a formula, the criterion prefers smaller

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\left(|\Gamma_i^\epsilon| - 1\right)^+,$$

where $t^+ := \max(t, 0)$. When these averages coincide for two conformal predictors, we compare the percentages (8) of empty predictions; larger values are preferable.

2.3 Observed criteria

The prior criteria discussed in the previous subsection treat the largest p-value, or prediction sets of size 1, in a special way. The corresponding criteria of this subsection attempt to achieve the same goal by using the observed label.

These are the observed counterparts of the non-basic prior ϵ-free criteria:

- The *OU* ("observed unconfidence") *criterion* uses the average observed unconfidence
$$\frac{1}{k}\sum_{i=l+1}^{l+k}\max_{y\neq y_i} p_i^y$$
over the test sequence, where the *observed unconfidence* for a test example (x_i, y_i) is the largest p-value p_i^y for the *false labels* $y \neq y_i$. Smaller values are preferable for this test.

- The *OF* ("observed fuzziness") *criterion* uses the average sum of the p-values for the false labels, i.e.,
$$\frac{1}{k}\sum_{i=l+1}^{l+k}\sum_{y\neq y_i} p_i^y; \tag{9}$$
smaller values are preferable.

The counterparts of the last group depending on the significance level ϵ are:

- The *OM criterion* uses the percentage of observed multiple predictions
$$\frac{1}{k}\sum_{i=l+1}^{l+k}\mathbf{1}_{\{\Gamma_i^\epsilon\setminus\{y_i\}\neq\emptyset\}}$$
in the test sequence, where an *observed multiple* prediction is defined to be a prediction set including a false label. Smaller values are preferable.

- The *OE criterion* (OE standing for "observed excess") uses the average number
$$\frac{1}{k}\sum_{i=l+1}^{l+k}|\Gamma_i^\epsilon\setminus\{y_i\}|$$
of false labels included in the prediction sets at significance level ϵ; smaller values are preferable.

The ten criteria used in this paper are given in Table 1. Half of the criteria depend on the significance level ϵ, and the other half are the respective ϵ-free versions.

In the case of binary classification problems, $|\mathbf{Y}| = 2$, the number of different criteria of efficiency in Table 1 reduces to six: the criteria not separated by a vertical or horizontal line (namely, U and F, OU and OF, M and E, and OM and OE) coincide.

Table 1: The ten criteria studied in this paper: the two basic ones in the upper section; the four other prior ones in the middle section; and the four observed ones in the lower section

ϵ-free	**ϵ-dependent**
S (sum of p-values)	*N (number of labels)*
U (unconfidence)	M (multiple)
F (fuzziness)	E (excess)
OU (observed unconfidence)	OM (observed multiple)
OF (observed fuzziness)	*OE (observed excess)*

3 Optimal Idealised Conformity Measures for a Known Probability Distribution

Starting from this section we consider the limiting case of infinitely long training and test sequences (and we will return to the realistic finitary case only in Section 6, where we describe our empirical studies). To formalise the intuition of an infinitely long training sequence, we assume that the prediction algorithm is directly given the data-generating probability distribution Q on $\mathbf{Z}$ instead of being given a training sequence. Instead of conformity measures we will use *idealised conformity measures*: functions $A(Q, z)$ of $Q \in \mathcal{P}(\mathbf{Z})$ (where $\mathcal{P}(\mathbf{Z})$ is the set of all probability measures on $\mathbf{Z}$) and $z \in \mathbf{Z}$. We will fix the data-generating distribution Q for the rest of the paper, and so write the corresponding conformity scores as $A(z)$. The *idealised conformal predictor* corresponding to A outputs the following prediction set $\Gamma^\epsilon(x)$ for each object $x \in \mathbf{X}$ and each significance level $\epsilon \in (0, 1)$. For each potential label $y \in \mathbf{Y}$ for x define the corresponding *p-value* as

$$p^y = p(x, y) := Q\{z \in \mathbf{Z} \mid A(z) < A(x, y)\} + \tau Q\{z \in \mathbf{Z} \mid A(z) = A(x, y)\} \quad (10)$$

(it would be more correct to write $A((x, y))$ and $Q(\{\ldots\})$, but we often omit pairs of parentheses when there is no danger of ambiguity), where τ is a random number distributed uniformly on $[0, 1]$. (The same random number τ is used in (10) for all (x, y).) The prediction set is

$$\Gamma^\epsilon(x) := \{y \in \mathbf{Y} \mid p(x, y) > \epsilon\}. \quad (11)$$

The *idealised conformal transducer* corresponding to A outputs for each object $x \in \mathbf{X}$ the system of p-values $(p^y \mid y \in \mathbf{Y})$ defined by (10); in the idealised case we will usually use the alternative notation $p(x, y)$ for p^y.

The standard properties of validity for conformal transducers and predictors mentioned in the previous section simplify in this idealised case as follows:

- If (x, y) is generated from Q, $p(x, y)$ is distributed uniformly on $[0, 1]$.

- Therefore, at each significance level ϵ the idealised conformal predictor makes an error with probability ϵ.

The test sequence being infinitely long is formalised by replacing the use of a test sequence in the criteria of efficiency by averaging with respect to the data-generating probability distribution Q. In the case of the top two and bottom two criteria in Table 1 (the ones set in italics) this is done as follows. Let us write $\Gamma^\epsilon_A(x)$ for the $\Gamma^\epsilon(x)$ in (11) and $p_A(x,y)$ for the $p(x,y)$ in (10) to indicate the dependence on the choice of the idealised conformity measure A. An idealised conformity measure A is:

- *S-optimal* if, for any idealised conformity measure B,

$$\mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p_A(x,y) \le \mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p_B(x,y),$$

  where the notation $\mathbb{E}_{x,\tau}$ refers to the expected value when x and τ are independent, $x \sim Q_{\mathbf{X}}$, and $\tau \sim U$; $Q_{\mathbf{X}}$ is the marginal distribution of Q on $\mathbf{X}$, and U is the uniform distribution on $[0,1]$;

- *N-optimal* if, for any idealised conformity measure B and any significance level ϵ,

$$\mathbb{E}_{x,\tau} \left|\Gamma^\epsilon_A(x)\right| \le \mathbb{E}_{x,\tau} \left|\Gamma^\epsilon_B(x)\right|;$$

- *OF-optimal* if, for any idealised conformity measure B,

$$\mathbb{E}_{(x,y),\tau} \sum_{y' \ne y} p_A(x,y') \le \mathbb{E}_{(x,y),\tau} \sum_{y' \ne y} p_B(x,y'),$$

  where the lower index (x,y) in $\mathbb{E}_{(x,y),\tau}$ refers to averaging over $(x,y) \sim Q$ (with (x,y) and τ independent);

- *OE-optimal* if, for any idealised conformity measure B and any significance level ϵ,

$$\mathbb{E}_{(x,y),\tau} \left|\Gamma^\epsilon_A(x) \setminus \{y\}\right| \le \mathbb{E}_{(x,y),\tau} \left|\Gamma^\epsilon_B(x) \setminus \{y\}\right|.$$

We will define the idealised versions of the other six criteria listed in Table 1 in Section 5.

4 Probabilistic Criteria of Efficiency

Our goal in this section is to characterise the optimal idealised conformity measures for the four criteria of efficiency that are set in italics in Table 1. We will assume in the rest of the paper that the set $\mathbf{X}$ is finite (from the practical point of view, this is not a restriction); since we consider the case of classification, $|\mathbf{Y}| < \infty$, this implies that the whole example space $\mathbf{Z}$ is finite. Without loss of generality, we also assume that the data-generating probability distribution

Q satisfies $Q_{\mathbf{X}}(x) > 0$ for all $x \in \mathbf{X}$ (we often omit curly braces in expressions such as $Q_{\mathbf{X}}(\{x\})$): we can always omit the xs for which $Q_{\mathbf{X}}(x) = 0$.

The *conditional probability (CP) idealised conformity measure* is

$$A(x, y) := Q(y \mid x) := \frac{Q(x, y)}{Q_{\mathbf{X}}(x)}. \tag{12}$$

This idealised conformity measure was introduced by an anonymous referee of the conference version of [3], but its non-idealised analogue in the case of regression had been used in [9] (following [8] and literature on minimum volume prediction). We say that an idealised conformity measure A is a *refinement* of an idealised conformity measure B if

$$B(z_1) < B(z_2) \Longrightarrow A(z_1) < A(z_2) \tag{13}$$

for all $z_1, z_2 \in \mathbf{Z}$. Let $\mathcal{R}(\mathrm{CP})$ be the set of all refinements of the CP idealised conformity measure. If C is a criterion of efficiency (one of the ten criteria in Table 1), we let $\mathcal{O}(C)$ stand for the set of all C-optimal idealised conformity measures.

Theorem 1. $\mathcal{O}(\mathrm{S}) = \mathcal{O}(\mathrm{OF}) = \mathcal{O}(\mathrm{N}) = \mathcal{O}(\mathrm{OE}) = \mathcal{R}(\mathrm{CP})$.

We say that an efficiency criterion is *probabilistic* if the CP idealised conformity measure is optimal for it. Theorem 1 shows that four of our ten criteria are probabilistic, namely S, N, OF, and OE (they are set in italics in Table 1). In the next section we will see that in general the other six criteria are not probabilistic. The intuition behind probabilistic criteria will be briefly discussed also in the next section.

Proof of Theorem 1. In this proof we partly follow [15], which simplified our original proof (considering, however, the case of label-conditional idealised conformal predictors and transducers).

We start from proving $\mathcal{R}(\mathrm{CP}) = \mathcal{O}(\mathrm{N})$. Let A be any idealised conformity measure. Fix for a moment a significance level ϵ. For each example $(x, y) \in \mathbf{Z}$, let $P(x, y)$ be the probability that the idealised conformal predictor based on A makes an error on the example (x, y) at the significance level ϵ, i.e., the probability of $y \notin \Gamma^\epsilon_A(x)$. It is clear from (10) and (11) that P takes at most three possible values (0, 1, and an intermediate value) and that

$$\sum_{x,y} Q(x, y) P(x, y) = \epsilon \tag{14}$$

(which just reflects the fact that the probability of error is ϵ). Vice versa, any P satisfying these properties will also satisfy

$$\forall (x, y) : P(x, y) = \mathbb{P}_{(x,y),\tau} \left(y \notin \Gamma^\epsilon_A(x) \right)$$

for some A. Let us see when we will have $A \in \mathcal{O}(\mathrm{N})$ (A is an N-optimal idealised conformity measure). Define Q' to be the probability measure on $\mathbf{Z}$ such that

$Q'_{\mathbf{X}} = Q_{\mathbf{X}}$ and $Q'(y \mid x) = 1/|\mathbf{Y}|$ does not depend on y. The N criterion at significance level ϵ for A can be evaluated as

$$\mathbb{E}_{x,\tau} |\Gamma^\epsilon_A(x)| = |\mathbf{Y}| \left(1 - \sum_{(x,y)} Q'(x,y) P(x,y) \right);$$

this expression should be minimised, i.e., $\sum_{(x,y)} Q'(x,y)P(x,y)$ should be maximised, under the restriction (14). Let us apply the Neyman–Pearson fundamental lemma ([7], Sect. 3.2, Theorem 1) using Q as the null and Q' as the alternative hypotheses. We can see that $\mathbb{E}_{x,\tau} |\Gamma^\epsilon_A(x)|$ takes its maximal value if and only if there exist thresholds $k_1 = k_1(\epsilon)$, $k_2 = k_2(\epsilon)$, and $k_3 = k_3(\epsilon)$ such that:

- $Q\{(x,y) \mid Q(y \mid x) < k_1\} < \epsilon \le Q\{(x,y) \mid Q(y \mid x) \le k_1\}$,
- $k_2 < k_3$,
- $A(x,y) < k_2$ if $Q(y \mid x) < k_1$,
- $k_2 < A(x,y) < k_3$ if $Q(y \mid x) = k_1$,
- $A(x,y) > k_3$ if $Q(y \mid x) > k_1$.

This will be true for all ϵ if and only if $Q(y \mid x)$ is a function of $A(x,y)$ (meaning that there exists a function F such that, for all (x,y), $Q(y \mid x) = F(A(x,y))$). This completes the proof of $\mathcal{R}(\mathrm{CP}) = \mathcal{O}(\mathrm{N})$.

Next we show that $\mathcal{O}(\mathrm{N}) = \mathcal{O}(\mathrm{S})$. We will use the equality between the extreme terms of

$$\sum_{y \in \mathbf{Y}} p(x,y) = \sum_{y \in \mathbf{Y}} \int_0^1 \mathbf{1}_{\{p(x,y) > \epsilon\}} \, \mathrm{d}\epsilon$$
$$= \int_0^1 \sum_{y \in \mathbf{Y}} \mathbf{1}_{\{p(x,y) > \epsilon\}} \, \mathrm{d}\epsilon = \int_0^1 |\Gamma^\epsilon(x)| \, \mathrm{d}\epsilon, \quad (15)$$

which implies

$$\mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p(x,y) = \int_0^1 \mathbb{E}_{x,\tau} |\Gamma^\epsilon(x)| \, \mathrm{d}\epsilon. \quad (16)$$

We can see that $A \in \mathcal{O}(\mathrm{S})$ whenever $A \in \mathcal{O}(\mathrm{N})$: indeed, any N-optimal idealised conformity measure minimises the expectation $\mathbb{E}_{x,\tau} |\Gamma^\epsilon(x)|$ on the right-hand side of (16) for all ϵ simultaneously, and so minimises the whole right-hand-side, and so minimises the left-hand-side. On the other hand, $A \notin \mathcal{O}(\mathrm{S})$ whenever $A \notin \mathcal{O}(\mathrm{N})$: indeed, if an idealised conformity measure fails to minimise the expectation $\mathbb{E}_{x,\tau} |\Gamma^\epsilon(x)|$ on the right-hand side of (16) for some ϵ, it fails to do so for all ϵ in a non-empty open interval (because of the right-continuity of $\mathbb{E}_{x,\tau} |\Gamma^\epsilon(x)|$ in ϵ, which follows from the Lebesgue dominated convergence

theorem and the right-continuity of $|\Gamma^\epsilon(x)| = |\Gamma^\epsilon(x,\tau)|$ in ϵ for a fixed τ), and therefore, it does not minimise the right-hand side of (16) (any N-optimal idealised conformity measure, such as the CP idealised conformity measure, will give a smaller value), and therefore, it does not minimise the left-hand side of (16).

The equality $\mathcal{O}(\mathrm{S}) = \mathcal{O}(\mathrm{OF})$ follows from

$$\mathbb{E}_{x,\tau} \sum_y p(x,y) = \mathbb{E}_{(x,y),\tau} \sum_{y' \neq y} p(x,y') + \frac{1}{2},$$

where we have used the fact that $p(x,y)$ is distributed uniformly on $[0,1]$ when $((x,y),\tau) \sim Q \times U$ (see [14]).

Finally, we notice that $\mathcal{O}(\mathrm{N}) = \mathcal{O}(\mathrm{OE})$. Indeed, for any significance level ϵ,

$$\mathbb{E}_{x,\tau} |\Gamma^\epsilon(x)| = \mathbb{E}_{(x,y),\tau} |\Gamma^\epsilon(x) \setminus \{y\}| + (1-\epsilon),$$

again using the fact that $p(x,y)$ is distributed uniformly on $[0,1]$ and so $\mathbb{P}_{(x,y),\tau}(y \in \Gamma^\epsilon(x)) = 1 - \epsilon$. □

Remark. The statement $\mathcal{O}(\mathrm{S}) = \mathcal{R}(\mathrm{CP})$ of Theorem 1 can be generalised to the criterion S_ϕ preferring small values of

$$\frac{1}{k} \sum_{i=l+1}^{l+k} \sum_y \phi(p_i^y)$$

(instead of (3)), where $\phi : [0,1] \to \mathbb{R}$ is a fixed continuously differentiable strictly increasing function, not necessarily the identity function. Namely, we still have $\mathcal{O}(\mathrm{S}_\phi) = \mathcal{R}(\mathrm{CP})$. Indeed, we can assume, without loss of generality, that $\phi(0) = 0$ and $\phi(1) = 1$ and replace (15) by

$$\begin{aligned}\sum_{y \in \mathbf{Y}} \phi(p(x,y)) = \sum_{y \in \mathbf{Y}} \int_0^1 \mathbf{1}_{\{\phi(p(x,y)) > \epsilon\}} \, \mathrm{d}\epsilon = \int_0^1 \sum_{y \in \mathbf{Y}} \mathbf{1}_{\{p(x,y) > \phi^{-1}(\epsilon)\}} \, \mathrm{d}\epsilon \\ = \int_0^1 \left|\Gamma^{\phi^{-1}(\epsilon)}(x)\right| \, \mathrm{d}\epsilon = \int_0^1 \left|\Gamma^{\epsilon'}(x)\right| \phi'(\epsilon') \, \mathrm{d}\epsilon',\end{aligned}$$

where ϕ' is the (continuous) derivative of ϕ, and then use the same argument as before.

5 Criteria of Efficiency that are not Probabilistic

Now we define the idealised analogues of the six criteria that are not set in italics in Table 1. An idealised conformity measure A is:

- *U-optimal* if, for any idealised conformity measure B, we have either

$$\mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_A(x, y') < \mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_B(x, y')$$

  or both

$$\mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_A(x, y') = \mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_B(x, y')$$

  and

$$\mathbb{E}_{x,\tau} \max_y p_A(x, y) \leq \mathbb{E}_{x,\tau} \max_y p_B(x, y);$$

- *M-optimal* if, for any idealised conformity measure B and any significance level ϵ, we have either

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| > 1) < \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_B(x)| > 1)$$

  or both

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| > 1) = \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_B(x)| > 1)$$

  and

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| = 0) \geq \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_B(x)| = 0);$$

- *F-optimal* if, for any idealised conformity measure B, we have either

$$\mathbb{E}_{x,\tau}\Big(\sum_y p_A(x, y) - \max_y p_A(x, y)\Big) < \mathbb{E}_{x,\tau}\Big(\sum_y p_B(x, y) - \max_y p_B(x, y)\Big)$$

  or both

$$\mathbb{E}_{x,\tau}\Big(\sum_y p_A(x, y) - \max_y p_A(x, y)\Big) = \mathbb{E}_{x,\tau}\Big(\sum_y p_B(x, y) - \max_y p_B(x, y)\Big)$$

  and

$$\mathbb{E}_{x,\tau} \max_y p_A(x, y) \leq \mathbb{E}_{x,\tau} \max_y p_B(x, y);$$

- *E-optimal* if, for any idealised conformity measure B and any significance level ϵ, we have either

$$\mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_A(x)| - 1)^+\big) < \mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_B(x)| - 1)^+\big)$$

  or both

$$\mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_A(x)| - 1)^+\big) = \mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_B(x)| - 1)^+\big)$$

  and

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| = 0) \geq \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_B(x)| = 0);$$

- *OU-optimal* if, for any idealised conformity measure B,

$$\mathbb{E}_{(x,y),\tau} \max_{y' \neq y} p_A(x, y') \leq \mathbb{E}_{(x,y),\tau} \max_{y' \neq y} p_B(x, y');$$

- *OM-optimal* if, for any idealised conformity measure B and any significance level ϵ,

$$\mathbb{P}_{(x,y),\tau}(\Gamma^\epsilon_A(x) \setminus \{y\} \neq \emptyset) \leq \mathbb{P}_{(x,y),\tau}(\Gamma^\epsilon_B(x) \setminus \{y\} \neq \emptyset).$$

In the following three definitions we follow [14], Chapter 3. The *predictability* of $x \in \mathbf{X}$ is

$$f(x) := \max_{y \in \mathbf{Y}} Q(y \mid x).$$

A *choice function* $\hat{y} : \mathbf{X} \to \mathbf{Y}$ is defined by the condition

$$\forall x \in \mathbf{X} : f(x) = Q(\hat{y}(x) \mid x).$$

Define the *signed predictability idealised conformity measure* corresponding to $\hat{y}$ by

$$A(x,y) := \begin{cases} f(x) & \text{if } y = \hat{y}(x) \\ -f(x) & \text{if not;} \end{cases}$$

a *signed predictability (SP) idealised conformity measure* is the signed predictability idealised conformity measure corresponding to some choice function.

For the following two theorems we will need to modify the notion of refinement. Let $\mathcal{R}'(\mathrm{SP})$ be the set of all idealised conformity measures A such that there exists an SP idealised conformity measure B that satisfies both (13) and

$$B(x,y_1) = B(x,y_2) \Longrightarrow A(x,y_1) = A(x,y_2)$$

for all $x \in \mathbf{X}$ and $y_1, y_2 \in \mathbf{Y}$.

Theorem 2. $\mathcal{O}(\mathrm{U}) = \mathcal{O}(\mathrm{M}) = \mathcal{R}'(\mathrm{SP})$.

We omit the proofs of Theorems 2–4 in this version of the paper.

Define the *MCP (modified conditional probability) idealised conformity measure* corresponding to a choice function $\hat{y}$ by

$$A(x,y) := \begin{cases} Q(y \mid x) & \text{if } y = \hat{y}(x) \\ Q(y \mid x) - 1 & \text{if not;} \end{cases}$$

an *MCP idealised conformity measure* is an idealised conformity measure corresponding to some choice function; $\mathcal{R}'(\mathrm{MCP})$ is defined analogously to $\mathcal{R}'(\mathrm{SP})$ but using MCP rather than SP idealised conformity measures.

Theorem 3. $\mathcal{O}(\mathrm{F}) = \mathcal{O}(\mathrm{E}) = \mathcal{R}'(\mathrm{MCP})$.

The *modified signed predictability idealised conformity measure* is defined by

$$A(x,y) := \begin{cases} f(x) & \text{if } f(x) > 1/2 \text{ and } y = \hat{y}(x) \\ 0 & \text{if } f(x) \leq 1/2 \\ -f(x) & \text{if } f(x) > 1/2 \text{ and } y \neq \hat{y}(x), \end{cases}$$

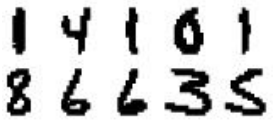

Figure 1: Examples of hand-written digits in the USPS data set.

where f is the predictability function; notice that this definition is unaffected by the choice of the choice function. Somewhat informally and assuming $|\mathbf{Y}| > 2$ (we are in the situation of Theorem 1 when $|\mathbf{Y}| = 2$), we define a set $\mathcal{R}''(\mathrm{MSP})$ in the same way as $\mathcal{R}'(\mathrm{MSP})$ (analogously to $\mathcal{R}'(\mathrm{SP})$) except that for $A \in \mathcal{R}''(\mathrm{MSP})$, $f(x) = 1/2$, and $y \neq \hat{y}(x)$ we allow $A(x, y) < A(x, \hat{y}(x))$.

Theorem 4. *If* $|\mathbf{Y}| > 2$, $\mathcal{O}(\mathrm{OU}) = \mathcal{O}(\mathrm{OM}) = \mathcal{R}''(\mathrm{MSP})$.

Theorems 2–4 show that the six criteria that are not set in italics in Table 1 are not probabilistic (except for OU and OM when $|\mathbf{Y}| = 2$, of course). Criteria of efficiency that are not probabilistic are somewhat analogous to "improper scoring rules" in probability forecasting (see, e.g., [2] and [4]). The optimal idealised conformity measures for the criteria of efficiency given in this paper that are not probabilistic have clear disadvantages, such as:

- They depend on the arbitrary choice of a choice function. In many cases there is a unique choice function, but the possibility of non-uniqueness is still awkward.
- They encourage "strategic behaviour" (such as ignoring the differences, which may be very substantial, between potential labels other than $\hat{y}(x)$ for a test object x when using the M criterion).

However, we do not use the terminology "proper/improper" in the case of criteria of efficiency for conformal prediction since it is conceivable that some non-probabilistic criteria of efficiency may turn out to be useful.

6 Empirical Study

In this section we demonstrate differences between two of our ϵ-free criteria, OF (probabilistic) and U (standard but not probabilistic) on the USPS data set of hand-written digits ([6]; examples of such digits are given in Figure 1, which is a subset of Figure 2 in [6]). We use the original split of the data set into the training and test sets. Our programs are written in R, and the results presented in the figures below are for the seed 0 of the R random number generator; however, we observe similar results in experiments with other seeds.

The problem is to classify hand-written digits, the labels are elements of $\{0, \dots, 9\}$, and the objects are elements of $\mathbb{R}^{256}$, where the 256 numbers represent the brightness of pixels in 16×16 pictures. We normalise each object by

applying the same affine transformation (depending on the object) to each of its pixels making the mean brightness of the pixels in the picture equal to 0 and making its standard deviation equal to 1. The sizes of the training and test sets are 7291 and 2007, respectively.

We evaluate six conformal predictors using the two criteria of efficiency. Fix a metric on the object space $\mathbb{R}^{256}$; in our experiments we use tangent distance (as implemented by Daniel Keysers) and Euclidean distance. Given a sequence of examples $(z_1, \dots, z_n)$, $z_i = (x_i, y_i)$, we consider the following three ways of computing conformity scores: for $i = 1, \dots, n$,

- $\alpha_i := \sum_{j=1}^{K} d_j^{\neq} / \sum_{j=1}^{K} d_j^{=}$, where $d_j^{\neq}$ are the distances, sorted in the increasing order, from x_i to the objects in $(z_1, \dots, z_n)$ with labels different from y_i (so that $d_1^{\neq}$ is the smallest distance from x_i to an object x_j with $y_j \neq y_i$), and $d_j^{=}$ are the distances, sorted in the increasing order, from x_i to the objects in $(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n)$ labelled as y_i (so that $d_1^{=}$ is the smallest distance from x_i to an object x_j with $j \neq i$ and $y_j = y_i$). We refer to this conformity measure as the *KNN-ratio conformity measure*; it has one parameter, K, whose range is $\{1, \dots, 50\}$ in our experiments (so that we always have $K \ll n$).

- $\alpha_i := N_i / K$, where N_i is the number of objects labelled as y_i among the K nearest neighbours of x_i (when $d_K = d_{K+1}$ in the ordered list $d_1, \dots, d_{n-1}$ of the distances from x_i to the other objects, we choose the nearest neighbours randomly among z_j with $y_j = y_i$ and with x_j at a distance of d_K from x_i). This conformity measure is a KNN counterpart of the CP idealised conformity measure (cf. (12)), and we will refer to it as the *KNN-CP conformity measure*; its parameter K is in the range $\{2, \dots, 50\}$ in our experiments.

- finally, we define $f_i := \max_y (N_i^y / K)$, where N_i^y is the number of objects labelled as y among the K nearest neighbours of x_i, $\hat{y}_i \in \arg\max_y (N_i^y / K)$ (chosen randomly from $\arg\max_y (N_i^y / K)$ if $|\arg\max_y (N_i^y / K)| > 1$), and

$$\alpha_i := \begin{cases} f_i & \text{if } y_i = \hat{y}_i \\ -f_i & \text{otherwise;} \end{cases}$$

  this is the *KNN-SP conformity measure*.

The three kinds of conformity measures combined with the two metrics (tangent and Euclidean) give six conformal predictors.

Figure 2 gives the average unconfidence (4) (top panel) and the average observed fuzziness (9) (bottom panel) over the test sequence (so that $k = 2007$) for a range of the values of the parameter K. Each of the six lines corresponds to one of the conformal predictors, as shown in the legends; in black-and-white the lines of the same type (dotted, solid, or dashed) corresponding to Euclidean and tangent distances can always be distinguished by their position: the former is above the latter.

The best results are for the KNN-ratio conformity measure combined with tangent distance for small values of the parameter K. For the two other types of conformity measures their relative evaluation changes depending on the kind of a criterion used to measure efficiency: as expected, the KNN-CP conformal predictors are better under the OF criterion, whereas the KNN-SP conformal predictors are better under the U criterion (cf. Theorems 1 and 2), if we ignore small values of K (when the probability estimates N_i^y/K are very unreliable).

7 Conclusion

This paper investigates properties of various criteria of efficiency of conformal prediction in the case of classification. It would be interesting to transfer, to the extent possible, this paper's results to the cases of:

Regression. The sum of p-values (as used in the S criterion) now becomes the integral under the p-value as function of the label y of the text example, and the size of a prediction set becomes its Lebesgue measure (considered, as already mentioned, in [9] in the non-idealised case). Whereas the latter is typically finite, ensuring the convergence of the former is less straightforward.

Anomaly detection. A first step in this direction is made in [13], which considers the average p-value as its criterion of efficiency.

Other natural directions of further research include:

- Extensions of our results to infinite, including non-discrete, $\mathbf{X}$.
- Extensions to Mondrian conformal predictors. In the case of label-conditional conformal predictors and probabilistic criteria, this was started in [15].
- Extensions to non-idealised conformal predictors.

Acknowledgments

We are grateful to the reviewers for helpful comments. This work was partially supported by EPSRC (grant EP/K033344/1), the Air Force Office of Scientific Research (grant "Semantic Completions"), and the EU Horizon 2020 Research and Innovation programme (grant 671555).

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] A. Philip Dawid. Probability forecasting. In Samuel Kotz, N. Balakrishnan, Campbell B. Read, Brani Vidakovic, and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 10, pages 6445–6452. Wiley, Hoboken, NJ, second edition, 2006.

[3] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Conformal prediction under hypergraphical models. In Harris Papadopoulos, Andreas S. Andreou, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Intelligence Applications and Innovations. Second Workshop on Conformal Prediction and Its Applications (COPA 2013)*, pages 371–383, Heidelberg, 2013. Springer. Journal version: *International Journal on Artificial Intelligence Tools* **24**(6), 1560003 (2015).

[4] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[5] Ulf Johansson, Rikard König, Tuve Löfström, and Henrik Boström. Evolved decision trees as conformal predictors. In Luis Gerardo de la Fraga, editor, *Proceedings of the 2013 IEEE Conference on Evolutionary Computation*, volume 1, pages 1794–1801, Cancun, Mexico, 2013.

[6] Yann Le Cun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan Kaufmann, San Francisco, CA, 1990.

[7] Erich L. Lehmann. *Testing Statistical Hypotheses*. Springer, New York, second edition, 1986.

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

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

[10] Thomas Melluish, Craig Saunders, Ilia Nouretdinov, and Vladimir Vovk. Comparing the Bayes and typicalness frameworks. In Luc De Raedt and Peter A. Flach, editors, *Proceedings of the Twelfth European Conference on Machine Learning*, volume 2167 of *Lecture Notes in Computer Science*, pages 360–371, Heidelberg, 2001. Springer.

[11] Harris Papadopoulos, Alex Gammerman, and Vladimir Vovk, editors. *Special Issue of the* Annals of Mathematics and Artificial Intelligence *on Conformal Prediction and its Applications*, volume 74(1–2). Springer, 2015.

[12] Craig Saunders, Alex Gammerman, and Vladimir Vovk. Transduction with confidence and credibility. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 722–726, San Francisco, CA, 1999. Morgan Kaufmann.

[13] James Smith, Ilia Nouretdinov, Rachel Craddock, Charles Offer, and Alexander Gammerman. Anomaly detection of trajectories with kernel density estimation by conformal prediction. In Lazaros Iliadis, Ilias Maglogiannis, Harris Papadopoulos, Spyros Sioutas, and Christos Makris, editors, *AIAI Workshops, COPA 2014*, volume 437 of *IFIP Advances in Information and Communication Technology*, pages 271–280, 2014.

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

[15] Vladimir Vovk, Ivan Petej, and Valentina Fedorova. From conformal to probabilistic prediction. In Lazaros Iliadis, Ilias Maglogiannis, Harris Papadopoulos, Spyros Sioutas, and Christos Makris, editors, *AIAI Workshops, COPA 2014*, volume 437 of *IFIP Advances in Information and Communication Technology*, pages 221–230, 2014.

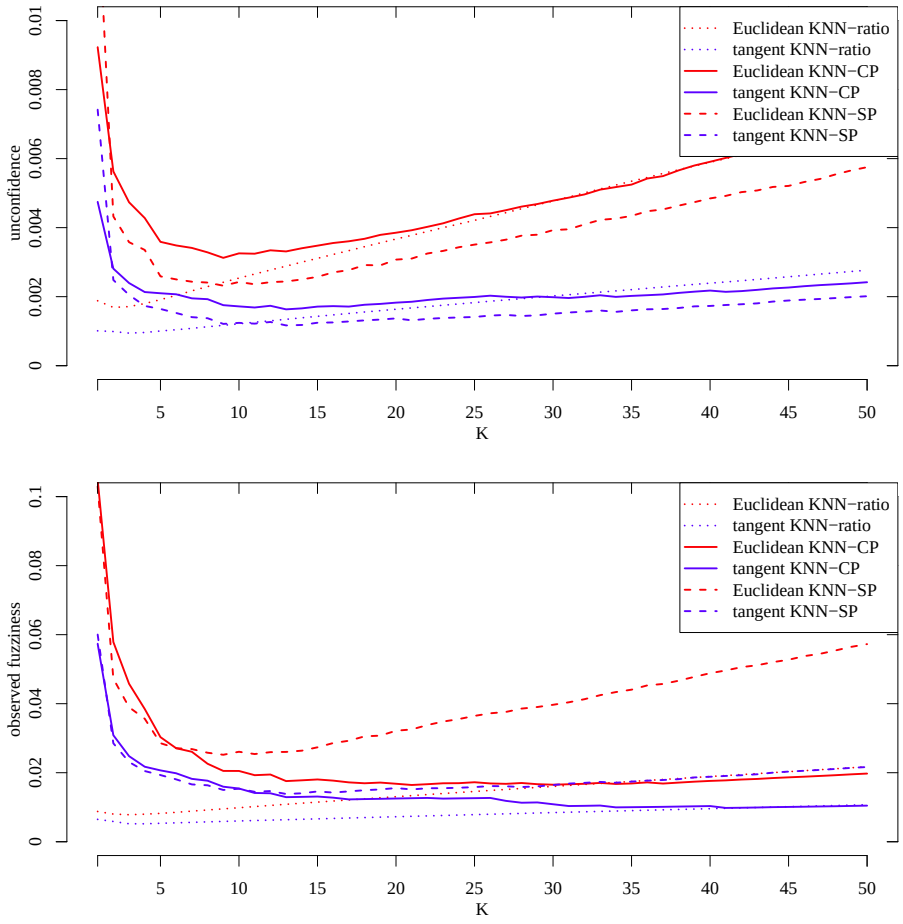

Figure 2: Top plot: average unconfidence for the USPS data set (for different values of parameters). Bottom plot: average observed fuzziness for the USPS data set. In black-and-white the lines of the same type (dotted, solid, or dashed) corresponding to Euclidean and tangent distances can always be distinguished by their position: the former is above the latter.