CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Royal Holloway - Pure

Annals of Mathematics and Artificial Intelligence

Criteria of efficiency for set-valued classification

--Manuscript Draft--

Manuscript Number:	AMAI-D-16-00121R1
Full Title:	Criteria of efficiency for set-valued classification
Article Type:	S84 COPA-2016
Corresponding Author:	Volodya Vovk Royal Holloway University of London UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Royal Holloway University of London
Corresponding Author's Secondary Institution:	
First Author:	Volodya Vovk
First Author Secondary Information:	
Order of Authors:	Volodya Vovk
	Ilia Nouretdinov
	Valentina Fedorova
	Ivan Petej
	Alex Gammerman
Order of Authors Secondary Information:	
Funding Information:	Engineering and Physical Sciences Research Council (EP/K033344/1) — Prof. Volodya Vovk, Dr Ilia Nouretdinov, Professor Alex Gammerman
	Air Force Office of Scientific Research (grant "Semantic Completions") — Prof. Volodya Vovk
	EU Horizon 2020 Research and Innovation programme (671555) — Prof. Volodya Vovk, Professor Alex Gammerman
Abstract:	We study optimal conformity measures for various criteria of efficiency of classification 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 unless the problem of classification is binary. We consider both unconditional and label-conditional conformal prediction.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Manuscript Click here to download Manuscript vovk_etal_2.tex

Click here to view linked References

Annals of Mathematics and Artificial Intelligence manuscript No.
(will be inserted by the editor)

Criteria of efficiency for set-valued classification

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

Received: date / Accepted: date

Abstract We study optimal conformity measures for various criteria of efficiency of set-valued classification in an idealised setting. This leads to an important class of criteria of efficiency that we call probabilistic and argue for; it turns out that the most standard criteria of efficiency used in literature on conformal prediction are not probabilistic unless the problem of classification is binary. We consider both unconditional and label-conditional conformal prediction.

Keywords Conformal prediction · Label-conditional conformal prediction · Predictive efficiency · Informational efficiency

Mathematics Subject Classification (2010) MSC 68T05 · 68Q32 · 62G15

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, 14] (and the COPA Proceedings,

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. Its conference version [18] was published in the Proceedings of the Fifth Symposium on Conformal and Probabilistic Prediction and Their Applications (COPA 2016, Madrid, April 2016) under the title "Criteria of efficiency for conformal prediction". This journal version also incorporates (in Section 8) some material of our paper [20] in COPA 2014. 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).

V. Vovk
Computer Learning Research Centre, Department of Computer Science
Royal Holloway, University of London, Egham, Surrey, UK
Tel.: +44-1784-443426
Fax: +44-1784-439786
E-mail: v.vovk@rhul.ac.uk

2012–2016). 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 (see the discussion at the end of Section 5). 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 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 (and Section 8 for the label-conditional version), 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., [19], p. 96; introduced in [16]). 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 [13], Section 7.2, and used extensively in [19].

The other two criteria that had been used before the publication of the conference version [18] of this paper 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: see 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. In Section 6 we prove the results of Section 5. After that we briefly illustrate the empirical behaviour of two of the criteria for standard conformal predictors and a benchmark data set (Section 7). Sections 2–7 discuss the most standard unconditional conformal predictors. Section 8 defines label-conditional conformal predictors and discusses the analogues of the results of the previous sections for label-conditional predictors. Finally, Section 9 gives some directions of further research.

A version (with a different treatment of empty observations) of one of the new non-probabilistic criteria of efficiency that we discuss in this paper (the one that we call the E criterion) has been introduced independently in [15].

We only consider the case of randomised ("smoothed") conformal predictors: the case of deterministic (non-smoothed) predictors may lead to combinatorial 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. [8], 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 $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$. We will always assume that the label space $\mathbf{Y}$ is non-empty, and will usually assume that its size is at least 2. 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

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

where $(z_1, \dots, z_l) \in \mathbf{Z}^*$ 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

$$p^y = p^y(z_1, \dots, z_l, x_{l+1}) := \frac{1}{l+1} \left|\left\{i = 1, \dots, l+1 \mid \alpha_i^y < \alpha_{l+1}^y\right\}\right| + \frac{\tau}{l+1} \left|\left\{i = 1, \dots, l+1 \mid \alpha_i^y = \alpha_{l+1}^y\right\}\right|, \tag{2}$$

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

$$(\alpha_1^y, \dots, \alpha_l^y, \alpha_{l+1}^y) := A(z_1, \dots, z_l, (x, y)). \tag{3}$$

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 $(z_1, \dots, z_l)$ of examples and each test object x. (This is just a different representation of the conformal predictor.)

Notice that the p-values (2) (and, therefore, the corresponding conformal predictors and transducers) only depend on the *conformity order* corresponding to the given conformity measure: namely, on the way that the elements of a sequence $(z_1, \dots, z_n)$ are ordered by the values $(\alpha_1, \dots, \alpha_n)$ (with $z_i \preceq z_j$ defined to be $\alpha_i \le \alpha_j$). Therefore, to define conformal predictors and transducers we may define their conformity orders rather than conformity measures.

The standard property of validity for conformal transducers is that the p-values p^y are distributed uniformly on $[0, 1]$ when the examples $z_1, \dots, z_l, (x, y)$ are generated independently from the same probability distribution Q on $\mathbf{Z}$ and τ is generated independently from the uniform probability distribution on $[0, 1]$ (see, e.g., [19], Proposition 2.8). This implies that the probability of error, $y \notin \Gamma^\epsilon(z_1, \dots, z_l, x)$, for conformal predictors is ϵ at any significance level ϵ.

Suppose we are given a test sequence $(z_{l+1}, \dots, z_{l+k})$ and would like to use it to measure the efficiency of the predictions derived from the training sequence $(z_1, \dots, z_l)$. (Informally, by the efficiency of conformal predictors we mean that the prediction sets they output tend to be small, and by the efficiency of conformal

transducers we mean that the p-values they output tend to be small.) For each test example $z_i = (x_i, y_i)$, $i = l+1, \dots, l+k$, we have a nested family $(\Gamma_i^\epsilon \mid \epsilon \in (0,1))$ of subsets of $\mathbf{Y}$, where

$$\Gamma_i^\epsilon := \Gamma^\epsilon(z_1, \dots, z_l, x_i),$$

and a system of p-values $(p_i^y \mid y \in \mathbf{Y})$, where

$$p_i^y := p^y(z_1, \dots, z_l, x_i).$$

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 label y_i of the test example. 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
$$\frac{1}{k} \sum_{i=l+1}^{l+k} \sum_y p_i^y \tag{4}$$
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 [11] (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

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\min_y\max_{y'\neq y}p_i^{y'} \tag{5}$$

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 (5) 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, the criterion compares the average credibilities

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\max_y p_i^y, \tag{6}$$

where the *credibility* for a test object x_i is the largest p-value $\max_y p_i^y$; smaller values of (6) 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

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\left(\sum_y p_i^y-\max_y p_i^y\right), \tag{7}$$

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 (7) are preferable. If two conformal transducers lead to the same average fuzziness, the criterion compares the average credibilities (6), 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

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

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 (8) of multiple predictions is the same for two conformal predictors (which is a common situation: the percentage can well be zero when the data is clean and ϵ is not too demanding), the M criterion compares the percentages

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\mathbf{1}_{\{\Gamma_i^\epsilon=\emptyset\}} \tag{9}$$

of empty predictions (larger values are preferable). This is a widely used criterion; in particular, it was used in [19] 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 (9) of empty predictions; larger values are preferable.

A criterion that is very similar to the M and E criteria is used by Lei in [9] (Section 2.2); that paper considers the binary case, in which the difference between the M and E criteria disappears. The difference of the criterion used in [9] is that it prohibits empty predictions (an intermediate approach would be to prefer smaller values for the number (9) of empty predictions). Lei's criterion is extended to the multi-class case in [15], which proposes a modification of the E criterion with a different treatment of empty predictions.

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.,

$$\frac{1}{k}\sum_{i=l+1}^{l+k}\sum_{y\neq y_i} p_i^y; \tag{10}$$

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.

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)*

– 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.

3 Idealised Setting

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 7, 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

$$p^y = p(x, y) = p_A(x, y) = p_A(x, y, \tau) := Q\{z \in \mathbf{Z} \mid A(z) < A(x, y)\} + \tau Q\{z \in \mathbf{Z} \mid A(z) = A(x, y)\} \quad (11)$$

(it would have been 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 (11) for all (x,y).) The prediction set is

$$\Gamma^\epsilon(x) = \Gamma^\epsilon_A(x) = \Gamma^\epsilon_A(x,\tau) := \{y \in \mathbf{Y} \mid p(x,y) > \epsilon\}\,. \tag{12}$$

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 (11); in the idealised case we will usually use the alternative notation $p(x,y)$ for p^y.

We could have used the *idealised conformity order* when defining the p-values (11): $z \preceq z'$ is defined to mean $A(z) \le A(z')$. Let us say that two idealised conformity measures are *equivalent* if they lead to the same idealised conformity order; in other words, A and B are equivalent if, for all $z, z' \in \mathbf{Z}$, $A(z) \le A(z') \Leftrightarrow B(z) \le B(z')$.

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 and $\tau \in [0,1]$ is generated from the uniform distribution independently of (x,y), $p(x,y)$ 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. An idealised conformity measure A is:

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

$$\mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p_A(x,y) \le \mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p_B(x,y), \tag{13}$$

  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

$$A(x, y) = Q(y \mid x) = Q_{\mathbf{Y}|\mathbf{X}}(y \mid x) := \frac{Q(x, y)}{Q_{\mathbf{X}}(x)}. \tag{14}$$

(In this paper, we will invariably use the shorter notation $Q(y \mid x)$ instead of the more precise $Q_{\mathbf{Y}|\mathbf{X}}(y \mid x)$; we will never need $Q_{\mathbf{X}|\mathbf{Y}}$, which could be defined analogously.) 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 [11] (following [10] and literature on minimum volume prediction). We say that an idealised conformity measure A is a *refinement* of an idealised conformity measure B if

$$B(z_1) < B(z_2) \Longrightarrow A(z_1) < A(z_2) \tag{15}$$

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 always optimal for it. We will also use two modifications of this definition: an efficiency criterion is *strongly probabilistic* if any refinement of the CP idealised conformity measure is optimal for it, and it is *weakly probabilistic* if some refinement of the CP idealised conformity measure is optimal for it. We will say that it is *BW probabilistic* (or *binary-weakly probabilistic*) if some refinement of the CP idealised conformity measure is optimal for it whenever $|\mathbf{Y}| = 2$. Theorem 1 shows that four of our ten criteria are strongly 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 (they are only BW probabilistic). The intuition behind probabilistic criteria will be briefly discussed also in the next section.

Proof (of Theorem 1) 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 (over τ) of $y \notin \Gamma^\epsilon_A(x)$. It is clear from (11) and (12) that P takes at most three possible values (0, 1, and an intermediate value) and that

$$\sum_{x,y} Q(x,y)P(x,y) = \epsilon \tag{16}$$

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

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

for some A, $\mathbb{P}_\tau$ standing for the probability when $\tau \sim U$. 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

$$\mathbb{E}_{x,\tau} \left| \Gamma^\epsilon_A(x) \right| = |\mathbf{Y}| \left(1 - \sum_{(x,y) \in \mathbf{Z}} Q'(x,y)P(x,y) \right); \tag{17}$$

this expression should be minimised, i.e., $\sum_{(x,y)} Q'(x,y)P(x,y)$ should be maximised, under the restriction (16). Let us apply the Neyman–Pearson fundamental lemma ([8], Sect. 3.2, Theorem 1) using Q as the null and Q' as the alternative hypotheses. We can see that $\mathbb{E}_{x,\tau} \left| \Gamma^\epsilon_A(x) \right|$ takes its minimal 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 $\mathcal{O}(\mathrm{N}) = \mathcal{O}(\mathrm{S})$. The chain of equalities

$$\sum_{y \in \mathbf{Y}} p(x,y) = \sum_{y \in \mathbf{Y}} \int_0^1 \mathbf{1}_{\{p(x,y) > \epsilon\}} \, \mathrm{d}\epsilon$$
$$= \int_0^1 \sum_{y \in \mathbf{Y}} \mathbf{1}_{\{p(x,y) > \epsilon\}} \, \mathrm{d}\epsilon = \int_0^1 \left| \Gamma^\epsilon(x) \right| \, \mathrm{d}\epsilon \tag{18}$$

(which will be used as the model in several other proofs in the rest of this paper) implies, by Fubini's theorem,

$$\mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p(x,y) = \int_0^1 \mathbb{E}_{x,\tau} \left| \Gamma^\epsilon(x) \right| \, \mathrm{d}\epsilon. \tag{19}$$

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 $\mathbb{E}_{x,\tau} \left| \Gamma^\epsilon(x) \right|$ on the right-hand side of (19) 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 (19) 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 is proved in Lemma 1(b) below), and therefore, it does not minimise the right-hand side of (19) (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 (19).

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 [19]).

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 $\mathbb{P}_{(x,y),\tau}(y \in \Gamma^\epsilon(x)) = 1-\epsilon$, where $\mathbb{P}_{(x,y),\tau}$ refers to the probability when $(x,y) \sim Q$ and $\tau \sim U$ are independent. □

The following lemma was used in the proof of Theorem 1.

Lemma 1 *(a) The function $\Gamma^\epsilon(x) = \Gamma^\epsilon(x,\tau)$ of ϵ is right-continuous for fixed x and τ. (b) The function $\mathbb{E}_{x,\tau} |\Gamma^\epsilon(x)|$ is right-continuous in ϵ.*

Proof Let us first check (a). We have (i) $p(x,y,\tau) > \epsilon$ for all $y \in \Gamma^\epsilon(x,\tau)$, and (ii) $p(x,y,\tau) \le \epsilon$ for all $y \notin \Gamma^\epsilon(x,\tau)$. If we increase ϵ, (ii) will be still satisfied, and if the increase is sufficiently small, (i) will be also satisfied and, therefore, $\Gamma^\epsilon(x,\tau)$ will not change. As for (b), the right-continuity of $\Gamma^\epsilon(x,\tau)$ in ϵ implies the right-continuity of $|\Gamma^\epsilon(x,\tau)|$ in ϵ, which implies the right-continuity of $\mathbb{E}_{x,\tau} |\Gamma^\epsilon(x,\tau)|$ in ϵ by the Lebesgue dominated convergence theorem. □

Remark 1 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

$$\frac{1}{k} \sum_{i=l+1}^{l+k} \sum_y \phi(p_i^y) \text{ or } \mathbb{E}_{x,\tau} \sum_y \phi(p(x,y))$$

(instead of (4) or (13), respectively), 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 (18) 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

$$\mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_A(x, y') < \mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_B(x, y') \tag{20}$$

  or both

$$\mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_A(x, y') = \mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p_B(x, y') \tag{21}$$

  and

$$\mathbb{E}_{x,\tau} \max_y p_A(x, y) \leq \mathbb{E}_{x,\tau} \max_y p_B(x, y); \tag{22}$$

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

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| > 1) < \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_B(x)| > 1) \tag{23}$$

  or both

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| > 1) = \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_B(x)| > 1) \tag{24}$$

  and

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| = 0) \geq \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_B(x)| = 0); \tag{25}$$

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

$$\mathbb{E}_{x,\tau}\Big(\sum_y p_A(x, y) - \max_y p_A(x, y)\Big) < \mathbb{E}_{x,\tau}\Big(\sum_y p_B(x, y) - \max_y p_B(x, y)\Big) \tag{26}$$

  or both

$$\mathbb{E}_{x,\tau}\Big(\sum_y p_A(x, y) - \max_y p_A(x, y)\Big) = \mathbb{E}_{x,\tau}\Big(\sum_y p_B(x, y) - \max_y p_B(x, y)\Big) \tag{27}$$

  and (22);

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

$$\mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_A(x)| - 1)^+\big) < \mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_B(x)| - 1)^+\big) \tag{28}$$

  or both

$$\mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_A(x)| - 1)^+\big) = \mathbb{E}_{x,\tau}\big((|\Gamma^\epsilon_B(x)| - 1)^+\big) \tag{29}$$

  and (25);

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

$$\mathbb{E}_{(x,y),\tau} \max_{y' \neq y} p_A(x, y') \leq \mathbb{E}_{(x,y),\tau} \max_{y' \neq y} p_B(x, y'); \tag{30}$$

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

$$\mathbb{P}_{(x,y),\tau}(\Gamma^\epsilon_A(x) \setminus \{y\} \neq \emptyset) \leq \mathbb{P}_{(x,y),\tau}(\Gamma^\epsilon_B(x) \setminus \{y\} \neq \emptyset). \tag{31}$$

In the following three definitions we follow [19], Chapter 3. The *predictability* of $x \in \mathbf{X}$ is

$$f(x) := \max_{y \in \mathbf{Y}} Q(y \mid x). \tag{32}$$

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

$$\forall x \in \mathbf{X} : f(x) = Q(\hat{y}(x) \mid x). \tag{33}$$

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

Theorems 2–4 will be proved in Section 6 below.

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 $\mathcal{R}(\mathrm{CP})$ but using MCP idealised conformity measures rather than the CP idealised conformity measure.

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

Of course, Theorems 2 and 3 are equivalent when $|\mathbf{Y}| = 2$.

The *modified signed predictability (MSP) idealised conformity measure* is defined by

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

where f is the predictability function (32); notice that this definition is unaffected by the choice of the choice function. Let $\mathcal{R}''(\mathrm{MSP})$ be the set of all refinements A of the MSP idealised conformity measure such that, for all $x \in \mathbf{X}$ and all $y_1, y_2 \in \mathbf{Y}$:

$$f(x) \ge 0.5 \ \&\ Q(y_1 \mid x) < 0.5 \ \&\ Q(y_2 \mid x) < 0.5 \Longrightarrow A(x, y_1) = A(x, y_2)$$
$$f(x) < 0.5 \Longrightarrow A(x, y_1) = A(x, y_2).$$

Table 2 Idealised conformity measures that are optimal for the ten criteria of efficiency given in Table 1; the arrangement of the criteria is the same as in Table 1

ϵ-free	ϵ-dependent
S: CP (Theorem 1)	*N: CP (Theorem 1)*
U: SP (Theorem 2)	M: SP (Theorem 2)
F: MCP (Theorem 3)	E: MCP (Theorem 3)
OU: MSP (Theorem 4)	OM: MSP (Theorem 4)
OF: CP (Theorem 1)	*OE: CP (Theorem 1)*

Theorem 4 $\mathcal{O}(\mathrm{OU}) = \mathcal{O}(\mathrm{OM}) = \mathcal{R}''(\mathrm{MSP})$.

Table 2 summarises the results given above. For each of the criteria listed in Table 1 it gives an optimal idealised conformity measure and cites the result asserting the optimality of that idealised conformity measure.

Theorems 2–4 show that the six criteria that are not set in italics in Table 1 are not probabilistic (however, we will see in Corollary 1 below that they are BW probabilistic). These are simple explicit examples (inevitably involving label spaces $\mathbf{Y}$ with $|\mathbf{Y}| > 2$) showing that they are not even weakly probabilistic:

– Let $\mathbf{X} = \{1\}$, $\mathbf{Y} = \{1, 2, 3\}$, and

$$Q_{\mathbf{X}}(1) = 1 \quad Q(1 \mid 1) = 0.2 \quad Q(2 \mid 1) = 0.3 \quad Q(3 \mid 1) = 0.5. \tag{34}$$

(Remember that, in this paper, $Q(y \mid x)$ always means $Q_{\mathbf{Y}|\mathbf{X}}(y \mid x)$.) In this case, all refinements of the CP idealised conformity measure are equivalent. The U criterion is not probabilistic since the expression

$$\mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p(x, y') \tag{35}$$

(cf. (20)) is 0.35 for the CP idealised conformity measure and is smaller, 0.25, for the SP idealised conformity measure. The M criterion is not probabilistic since at significance level $\epsilon = 0.2$ the CP idealised conformity measure gives the predictor $\Gamma^\epsilon(1) = \{2, 3\}$ (a.s.), and so

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_{\mathrm{CP}}(x)| > 1) = 1 > 0.6 = \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_{\mathrm{SP}}(x)| > 1)$$

(cf. (23)).

– Let $\mathbf{X} = \{1, 2\}$, $\mathbf{Y} = \{1, 2, 3\}$, and, for a small $\delta > 0$,

$$\begin{array}{llll} Q_{\mathbf{X}}(1) = 0.5 & Q(1 \mid 1) = 1/3 - \delta & Q(2 \mid 1) = 1/3 & Q(3 \mid 1) = 1/3 + \delta \\ Q_{\mathbf{X}}(2) = 0.5 & Q(1 \mid 2) = 1/3 - 5\delta & Q(2 \mid 2) = 1/3 + 2\delta & Q(3 \mid 2) = 1/3 + 3\delta. \end{array}$$

The CP idealised conformity measure again has only equivalent refinements. The F criterion is not probabilistic since the expression

$$\mathbb{E}_{x,\tau}\Big(\sum_y p(x, y) - \max_y p(x, y)\Big) \tag{36}$$

(cf. (26)) is $3/4 + O(\delta)$ for the CP idealised conformity measure and is smaller (provided δ is sufficiently small), $2/3 + O(\delta)$, for the MCP idealised conformity measure (which is unique). The E criterion is not probabilistic since at

significance level $\epsilon = 2/3$ the CP idealised conformity measure has a larger expected excess (for small δ) than the MCP idealised conformity measure (whose expected excess is zero):

$$\mathbb{E}_{x,\tau}\left(\left(|\Gamma^\epsilon_{\mathrm{CP}}(x)| - 1\right)^+\right) = 0.5 + O(\delta) > 0 = \mathbb{E}_{x,\tau}\left(\left(|\Gamma^\epsilon_{\mathrm{MCP}}(x)| - 1\right)^+\right)$$

(cf. (28)).

– Let us again set $\mathbf{X} = \{1\}$ and $\mathbf{Y} = \{1, 2, 3\}$, and define Q by (34). The OU criterion is not probabilistic since the expression

$$\mathbb{E}_{(x,y),\tau} \max_{y' \neq y} p(x, y') \tag{37}$$

(cf. (30)) is 0.55 for the CP idealised conformity measure and is smaller, 0.5, for the MSP idealised conformity measure. The OM criterion is not probabilistic since at significance level $\epsilon = 0.2$ the CP idealised conformity measure gives the predictor $\Gamma^\epsilon(1) = \{2, 3\}$ (a.s.), and so

$$\mathbb{P}_{(x,y),\tau}(\Gamma^{0.2}_{\mathrm{CP}}(x) \setminus \{y\} \neq \emptyset) = 1 > 0.8 = \mathbb{P}_{(x,y),\tau}(\Gamma^{0.2}_{\mathrm{MSP}}(x) \setminus \{y\} \neq \emptyset)$$

(cf. (31)).

Corollary 1 *All ten criteria of efficiency in Table 1 are BW probabilistic.*

Proof Criteria S, N, OF, and OE are BW probabilistic by Theorem 1. Criteria OU and OM are identical to OF and OE, respectively, in the binary case, and so are also BW probabilistic. Criteria F and E are identical to U and M, respectively, in the binary case, and so our task reduces to proving that U and M are BW probabilistic. By Theorem 2, it suffices to check $\mathcal{R}(\mathrm{CP}) \cap \mathcal{R}'(\mathrm{SP}) \neq \emptyset$, which is obvious: SP is in both $\mathcal{R}(\mathrm{CP})$ and $\mathcal{R}'(\mathrm{SP})$ when $|\mathbf{Y}| = 2$. □

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 in the case $|\mathbf{Y}| > 2$).

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 still turn out to be useful.

6 Proofs of Theorems 2–4

The proofs in this section will be slightly less formal than the proof of Theorem 1; in particular, all references to the Neyman–Pearson lemma will be implicit.

6.1 Proof of Theorem 2

We start from checking that $\mathcal{O}(\mathrm{M}) = \mathcal{R}'(\mathrm{SP})$ (essentially reproducing the argument given in the second parts of the proofs of Propositions 3.3 and 3.4 in [19]). We will analyze the requirements imposed by being M-optimal on the prediction set Γ^ϵ starting from small values of $\epsilon \in (0,1)$. (In this paper we only consider ϵ in the interval $(0,1)$, even if this restriction is not mentioned explicitly.)

Let $f_1 > f_2 > \cdots > f_n > 0$ be the list of the predictabilities (see (32)) of all objects $x \in \mathbf{X}$, with all duplicates removed and the remaining predictabilities sorted in the decreasing order. It is clear that an M-optimal idealised conformity measure will assign the lowest conformity to the group of examples (x, y) with $f(x) = f_1$ and $y \neq \hat{y}(x)$ for some choice function $\hat{y}$ (see (33)). The conformity of such examples can be different unless they contain the same object (in which case it must be the same); the conformity of any example in any other group must be higher than the conformity of the examples in this first group. If these conditions are satisfied for some idealised conformity measure A, A will satisfy (23) or (24) for any idealised conformity measure B and any

$$\epsilon \in (0, Q\{(x,y) \mid f(x) = f_1 \ \ \& \ \ y \neq \hat{y}(x)\}] \,.$$

The second least conforming group of examples consists of (x, y) with $f(x) = f_2$ and $y \neq \hat{y}(x)$ for some choice function $\hat{y}$. The conformity of examples in the second group can again be different unless they contain the same object. These and previous conditions ensure that A will satisfy (23) or (24) for any

$$\epsilon \in (0, Q\{(x,y) \mid f(x) \geq f_2 \ \ \& \ \ y \neq \hat{y}(x)\}] \,.$$

Continuing in such a way, we will obtain a choice function $\hat{y}$ and the conformity ordering for the examples whose label is not chosen by that choice function $\hat{y}$. All these examples are divided into n groups, and each elements of the ith group is coming before each element of the jth group when $i < j$; in the end we will get $2n$ groups satisfying this property. The first n groups take care of

$$\epsilon \in (0, Q\{(x,y) \mid y \neq \hat{y}(x)\}] \,.$$

The next, $(n+1)$th, group of examples are $(x, \hat{y}(x)) \in \mathbf{Z}$ with $f(x) = f_n$; they can be ordered in any way between themselves. If the conditions listed so far are satisfied for an idealised conformity measure A, A will satisfy (23)–(25) for any idealised conformity measure B and any

$$\epsilon \in (0, Q\{(x,y) \mid y \neq \hat{y}(x) \text{ or } (y = \hat{y}(x) \ \ \& \ \ f(x) = f_n)\}] \,.$$

The following, $(n+2)$th, group consists of $(x, \hat{y}(x)) \in \mathbf{Z}$ with $f(x) = f_{n-1}$. Continuing in the same way until all examples are exhausted, we will obtain a refinement of the SP idealised conformity measure that belongs to $\mathcal{R}'(\mathrm{SP})$.

This proof of $\mathcal{O}(\mathrm{M}) = \mathcal{R}'(\mathrm{SP})$ demonstrates the following property of M-optimal idealised conformity measures.

Corollary 2 *If $A \in \mathcal{O}(\mathrm{M})$,*

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| > 1)\, \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| = 0) = 0$$

at each significance level ϵ.

Let us now check that $\mathcal{O}(\mathrm{U}) = \mathcal{O}(\mathrm{M})$. Analogously to (18) and (19), we have, for a given idealised conformity measure A (omitted from our notation),

$$\mathbb{E}_{x,\tau} \min_y \max_{y' \neq y} p(x, y', \tau) = \mathbb{E}_{x,\tau} \int_0^1 \mathbf{1}_{\{\min_y \max_{y' \neq y} p(x,y',\tau) > \epsilon\}} \, \mathrm{d}\epsilon$$
$$= \mathbb{E}_{x,\tau} \int_0^1 \mathbf{1}_{\{|\Gamma^\epsilon(x)|>1\}} \, \mathrm{d}\epsilon = \int_0^1 \mathbb{P}_{x,\tau} \left(|\Gamma^\epsilon(x)| > 1\right) \mathrm{d}\epsilon. \quad (38)$$

Similarly, we have

$$\mathbb{E}_{x,\tau} \max_y p(x, y, \tau) = \mathbb{E}_{x,\tau} \int_0^1 \mathbf{1}_{\{\max_y p(x,y,\tau) > \epsilon\}} \, \mathrm{d}\epsilon$$
$$= \mathbb{E}_{x,\tau} \int_0^1 \mathbf{1}_{\{|\Gamma^\epsilon(x)|>0\}} \, \mathrm{d}\epsilon = \int_0^1 \mathbb{P}_{x,\tau} \left(|\Gamma^\epsilon(x)| > 0\right) \mathrm{d}\epsilon$$
$$= 1 - \int_0^1 \mathbb{P}_{x,\tau} \left(|\Gamma^\epsilon(x)| = 0\right) \mathrm{d}\epsilon. \quad (39)$$

Our argument will also use the following continuity property for idealised conformal predictors. (For now, we only need parts (a) and (b).)

Corollary 3 *The functions*

(a) $\mathbb{P}_{x,\tau} \left(|\Gamma^\epsilon(x)| > 1\right)$
(b) $\mathbb{P}_{x,\tau} \left(|\Gamma^\epsilon(x)| = 0\right)$
(c) $\mathbb{E}_{x,\tau} \left((|\Gamma^\epsilon(x)| - 1)^+\right)$
(d) $\mathbb{P}_{(x,y),\tau} \left(\Gamma^\epsilon(x) \setminus \{y\} \neq \emptyset\right)$

are right-continuous in ϵ.

Proof All these statements can be deduced from part (a) of Lemma 1 in the same way as in the proof of part (b) of that lemma. The right-continuity of the function $\Gamma^\epsilon(x, \tau)$ implies the right-continuity of $\mathbf{1}_{\{|\Gamma^\epsilon(x)|>1\}}$ (remember that $|\Gamma^\epsilon(x)|$ takes only integer values). Therefore, the right-continuity of $\mathbb{P}_{x,\tau} \left(|\Gamma^\epsilon(x)| > 1\right)$ follows by the Lebesgue dominated convergence theorem. This proves (a), and proofs of (b)–(d) are analogous. □

First suppose that A is M-optimal. Let B be any idealised conformity measure. From (38), it is clear that (20) holds with $<$ replaced by $\leq$. If, furthermore, we have (21): by Corollary 3 we also have (24) for all ϵ; therefore, we also have (25) for all ϵ; in combination with (39), we obtain (22). Therefore, A is U-optimal.

Now suppose that A is U-optimal. Let B be the SP idealised conformity measure, which we know to be not only M-optimal but also U-optimal (as shown in the previous paragraph). By the definition ((20)–(22)) of U-optimality, we have (21) and (22) with $=$ in place of $\leq$. This implies that (24) holds for all ϵ (had the equality been violated for some $\epsilon \in (0, 1)$, it would have been violated for a range of ϵ by Corollary 3, which would have contradicted (21)). In the same way, it implies that (25) holds (even with $=$ in place of $\geq$) for all ϵ. Therefore, A is M-optimal.

6.2 Proof of Theorem 3

Our argument for $\mathcal{O}(\mathrm{E}) = \mathcal{R}(\mathrm{MCP})$ will be similar to the argument for $\mathcal{O}(\mathrm{M}) = \mathcal{R}'(\mathrm{SP})$ given in the previous subsection; we will again analyze the requirements imposed by being E-optimal starting from small values of $\epsilon \in (0,1)$. Let $g_1 < g_2 < \dots < g_n$ be the list of the conditional probabilities $Q(y \mid x)$ of all examples $(x,y) \in \mathbf{Z}$, with all duplicates removed and the remaining conditional probabilities sorted in the increasing order. All examples will be split into $2n$ groups, with the examples in the ith and $(n+i)$th groups satisfying $Q(y \mid x) = g_i$, $i = 1, \dots, n$. Initially the ith group, $i = 1, \dots, n$, contains all examples satisfying $Q(y \mid x) = g_i$, and the other groups are empty. (Later some of the examples will be moved into the groups numbered $n+1, n+2, \dots$, and as a result some of the first n groups may become empty.) It will be true that each element of the ith group will be coming before each element of the jth group when $1 \le i < j \le 2n$.

Any F-optimal idealised conformity measure will assign the lowest conformity to the first group of examples, perhaps except for examples (x,y) for which $Q(y \mid x) = \max_{y'} Q(y' \mid x)$. If for some $x \in \mathbf{X}$, the first group contains (x,y) with $Q(y \mid x) = \max_{y'} Q(y' \mid x)$, we choose one such (x,y) for each such x and move it to the $(n+1)$th group. The rest of the examples in the group can be ordered in their conformity in any way (with ties allowed). The examples in the $(n+1)$th group can also be ordered arbitrarily. Process the 2nd, 3rd,..., nth groups in the same way. It is clear that in the end we will obtain a refinement of an MCP idealised conformity measure.

Next we prove $\mathcal{O}(\mathrm{E}) = \mathcal{O}(\mathrm{F})$. Defining a *p-choice function* $\tilde{y} : \mathbf{X} \to \mathbf{Y}$ (for a given idealised conformity measure) by the requirement

$$p(x, \tilde{y}(x)) = \max_y p(x,y),$$

we have the following analogue of (18):

$$\sum_{y \in \mathbf{Y}} p(x,y) - \max_{y \in \mathbf{Y}} p(x,y) = \sum_{y \in \mathbf{Y} \setminus \{\tilde{y}(x)\}} p(x,y) = \sum_{y \in \mathbf{Y} \setminus \{\tilde{y}(x)\}} \int_0^1 \mathbf{1}_{\{p(x,y) > \epsilon\}} \, \mathrm{d}\epsilon$$
$$= \int_0^1 \sum_{y \in \mathbf{Y} \setminus \{\tilde{y}(x)\}} \mathbf{1}_{\{p(x,y) > \epsilon\}} \, \mathrm{d}\epsilon = \int_0^1 \left(|\Gamma^\epsilon(x)| - 1 \right)^+ \mathrm{d}\epsilon.$$

This implies, similarly to (19),

$$\mathbb{E}_{x,\tau} \left(\sum_{y \in \mathbf{Y}} p(x,y) - \max_{y \in \mathbf{Y}} p(x,y) \right) = \int_0^1 \mathbb{E}_{x,\tau} \left(\left(|\Gamma^\epsilon(x)| - 1 \right)^+ \right) \mathrm{d}\epsilon. \tag{40}$$

Suppose that A is E-optimal, and let B be any idealised conformity measure. From (40), it is clear that (26) holds with $<$ replaced by $\le$. If, furthermore, we have (27): by Corollary 3(c) we also have (29) for all ϵ; therefore, we also have (25) for all ϵ; in combination with (39), we obtain (22). Therefore, A is F-optimal.

Now suppose that A is F-optimal. Let B be any MCP idealised conformity measure, which we know to be both E-optimal and F-optimal. By the definition of F-optimality, we have (27) and (22) with $=$ in place of $\le$. As in the previous subsection, this implies that (29) holds for all ϵ, and also implies that (25) holds (even with $=$ in place of $\ge$) for all ϵ. Therefore, A is E-optimal.

6.3 Proof of Theorem 4

The proof is similar to the proofs in the previous two subsections. First we check that $\mathcal{O}(\mathrm{OM}) = \mathcal{R}''(\mathrm{MSP})$, analyzing the requirement of OM-optimality starting from small values of $\epsilon \in (0,1)$. Let $f_1 > f_2 > \cdots > f_n > 0.5$ be the list of the predictabilities of all objects $x \in \mathbf{X}$ whose predictability exceeds 0.5, with all duplicates removed and the remaining predictabilities sorted in the decreasing order. All examples are split into $2n+1$ groups (perhaps some of them empty) in such a way that each element of the ith group is coming before each element of the jth group when $1 \le i < j \le 2n+1$. The ith group, $i = 1, \dots, n$, contains all examples (x, y) with predictability f_i and $Q(y \mid x) < 1/2$, the $(n+1)$th group contains all examples with predictability 0.5 or less, and the $(n+1+i)$th group, $i = 1, \dots, n$, contains all examples (x, y) with $Q(y \mid x) = f_i$ (there is, however, at most one such example); it is possible that $n = 0$.

Any OM-optimal idealised conformity measure will assign the lowest conformity to the first group of examples (assuming $n \ge 1$), and those examples can be ordered arbitrarily in their conformity, except that any examples sharing their objects should have the same conformity. This group takes care of the values

$$\epsilon \in \left(0, Q\left\{(x,y) \mid f(x) = f_1 \ \ \& \ \ Q(y \mid x) \neq f_1\right\}\right].$$

Proceed in the same way through groups $2, \dots, n$. The $(n+1)$th group is most complicated (when non-empty). It contains the following kinds of examples:

- Examples whose predictability is less than 0.5. All such examples should have the same conformity if they share the same object.
- Examples (x, y) whose predictability is exactly 0.5 and which satisfy $Q(y \mid x) < 0.5$. All such examples should have the same conformity if they share the same object.
- Examples (x, y) whose predictability is exactly 0.5 and which satisfy $Q(y \mid x) = 0.5$.

Otherwise, the examples in the $(n+1)$th group can be ordered arbitrarily in their conformity. Groups $n+2, \dots, 2n+1$ are singletons or empty and do not cause any problems. Therefore, an idealised conformity measure is OM-optimal if and only if it is in $\mathcal{R}''(\mathrm{MSP})$.

Next we check that $\mathcal{O}(\mathrm{OU}) = \mathcal{O}(\mathrm{OM})$. Similarly to (38), we have, for a given idealised conformity measure,

$$\begin{aligned}\mathbb{E}_{(x,y),\tau} \max_{y' \neq y} p(x, y', \tau) &= \mathbb{E}_{(x,y),\tau} \int_0^1 \mathbf{1}_{\{\max_{y' \neq y} p(x,y',\tau) > \epsilon\}} \, \mathrm{d}\epsilon \\ &= \mathbb{E}_{(x,y),\tau} \int_0^1 \mathbf{1}_{\{\Gamma^\epsilon(x) \setminus \{y\} \neq \emptyset\}} \, \mathrm{d}\epsilon = \int_0^1 \mathbb{P}_{x,\tau}\left(\Gamma^\epsilon(x) \setminus \{y\} \neq \emptyset\right) \mathrm{d}\epsilon. \quad (41)\end{aligned}$$

By (41), OM-optimality immediately implies OU-optimality.

Now suppose that A is OU-optimal. Let B be the MSP idealised conformity measure, which is both OM-optimal and OU-optimal. If (31) is violated for some ϵ, it is violated for a range of ϵ (by Corollary 3(d)), which, by (41), contradicts the OU-optimality of A. Therefore, A is OM-optimal.

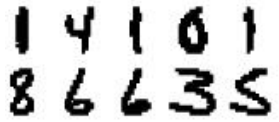

Fig. 1 Examples of hand-written digits in the USPS data set.

7 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 ([7]; examples of such digits are given in Figure 1, which is a subset of Figure 2 in [7]). 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. (14)), 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, $\hat{y}_i \in \arg\max_y(N_i^y/K)$

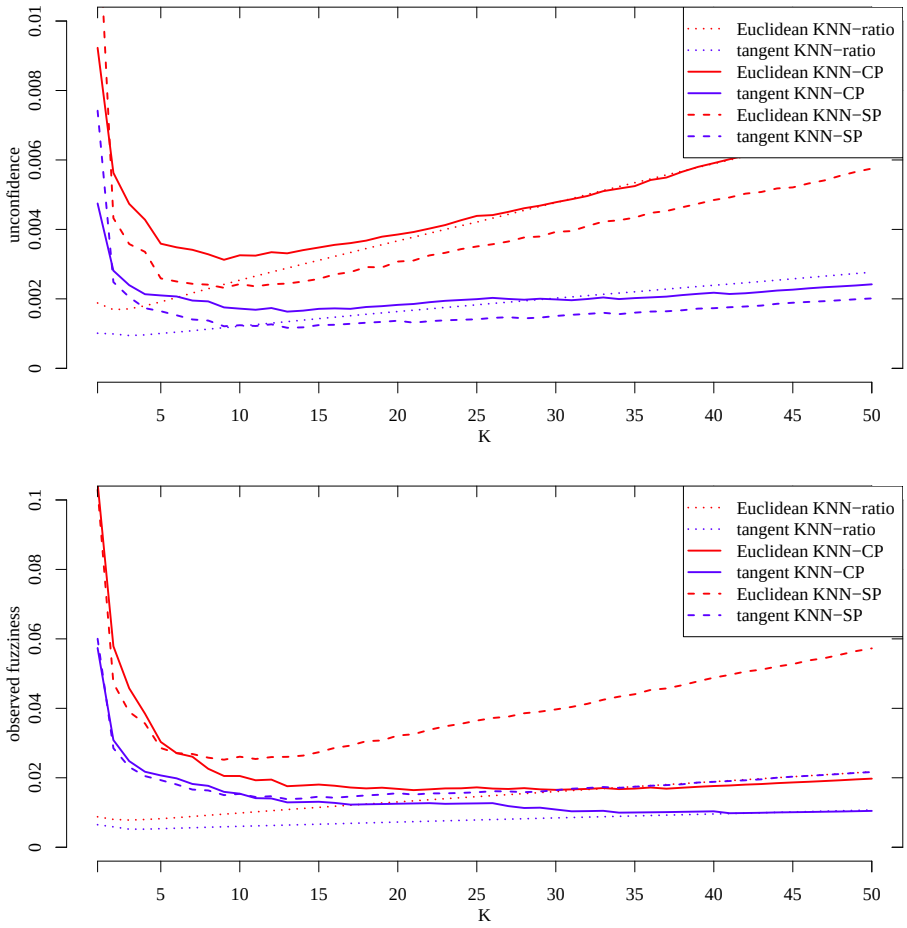

Fig. 2 Top plot: average unconfidence for the USPS data set (for different values of parameters). Bottom plot: average observed fuzziness for the USPS data set. In black-and-white the lines of the same type (dotted, solid, or dashed) corresponding to Euclidean and tangent distances can always be distinguished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

(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. We use both metrics in order to test the performance of our criteria for different kinds of underlying algorithms: both more efficient (represented by tangent metric) and less efficient (represented by Euclidean metric).

Figure 2 gives the average unconfidence (5) (top panel) and the average observed fuzziness (10) (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).

Our conclusion is that whereas some conformal predictors (such as the KNN-ratio ones in our experiments) can perform well under different criteria of efficiency, the performance of other conformal predictors depends very much on the criterion of efficiency used to evaluate it.

8 Efficiency of Label-conditional Conformal Predictors and Transducers

Conformal predictors, as defined in Section 2, only guarantee the overall coverage probability, averaged over all labels. Sometimes we want to have a guarantee for the coverage probability for each label $y \in \mathbf{Y}$ separately, and in this case one should use label-conditional conformal predictors, which are studied in this section.

8.1 Label-conditional conformal predictors and transducers

The *label-conditional conformal predictor* determined by a conformity measure A is defined by (1) where the *label-conditional p-values* p^y are defined by

$$p^y := \Big(\left| \{ i = 1, \dots, l \mid y_i = y \ \& \ \alpha_i^y < \alpha_{l+1}^y \} \right| + \tau \left| \{ i = 1, \dots, l \mid y_i = y \ \& \ \alpha_i^y = \alpha_{l+1}^y \} \right| + \tau \Big) \Big/ \left(\left| \{ i = 1, \dots, l \mid y_i = y \} \right| + 1 \right) \quad (42)$$

(instead of (2)); as before, τ is a random number distributed uniformly on the interval $[0, 1]$ (conditionally on all the examples), and the conformity scores are defined by (3).

The *label-conditional conformal transducer* determined by A outputs the system of p-values $(p^y \mid y \in \mathbf{Y})$ defined by (42) for each training sequence $(z_1, \dots, z_l)$ of examples and each test object x. The property of validity for label-conditional conformal predictors and transducers is that the p-values p^y are distributed uniformly on $[0, 1]$ given y when the examples $z_1, \dots, z_l, (x, y)$ are generated independently from the same probability distribution Q on $\mathbf{Z}$ (see, e.g., [19],

Proposition 4.10). This implies that the conditional probability of error, $y \notin \Gamma^\epsilon(z_1, \dots, z_l, x)$, given y is ϵ at any significance level ϵ.

The p-values (42), and the corresponding conformal predictors and transducers, only depend on the conformity order within each class: now we define $(x_i, y_i) \preceq (x_j, y_j)$ to mean $y_i = y_j$ and $\alpha_i \le \alpha_j$ (with (x_i, y_i) and (x_j, y_j) such that $y_i \ne y_j$ being incomparable).

The definitions of all ten criteria of efficiency introduced in Section 2 and listed in Table 1 carry over to the case of label-conditional conformal transducers and predictors.

8.2 Idealised setting

As before, we assume that the object space $\mathbf{X}$ is finite and $Q_{\mathbf{X}}(x) > 0$ for all $x \in \mathbf{X}$. We also assume $Q_{\mathbf{Y}}(y) > 0$ for all $y \in \mathbf{Y}$, where $Q_{\mathbf{Y}}$ is the marginal distribution of Q on the label space $\mathbf{Y}$.

Let A be an idealised conformity measure. For each potential label $y \in \mathbf{Y}$ for an object x define the corresponding *label-conditional p-value* as

$$p^y = p(x, y) := \frac{Q\{(x', y) \in \mathbf{Z} \mid A(x', y) < A(x, y)\}}{Q_{\mathbf{Y}}(y)} + \tau \frac{Q\{(x', y) \in \mathbf{Z} \mid A(x', y) = A(x, y)\}}{Q_{\mathbf{Y}}(y)}, \quad (43)$$

analogously to (11), with the same random number $\tau \in [0, 1]$ used for all (x, y). The *label-conditional idealised conformal predictor* is defined by (12) for the new definition of $p(x, y)$ and the *label-conditional idealised conformal transducer* corresponding to the idealised conformity measure A outputs for each object $x \in \mathbf{X}$ the system of p-values $(p^y \mid y \in \mathbf{Y})$ defined by (43).

The idealised p-values (43), and the corresponding idealised conformal predictors and transducers, also depend only on the conformity order within each class: we can define $(x, y) \preceq (x', y')$ to mean $y = y'$ and $A(x, y) \le A(x', y')$. Two idealised conformity measures are *equivalent within classes* if they lead to the same order $\preceq$; in this section we will consider only this notion of equivalence (without mentioning it explicitly).

The properties of validity now become conditional:

- If (x, y) is generated from Q and τ is generated independently from the uniform probability distribution on $[0, 1]$, $p(x, y)$ is distributed uniformly on $[0, 1]$ even if we condition on y.
- Therefore, at each significance level ϵ the idealised conformal predictor makes an error with conditional probability ϵ given y.

8.3 Probabilistic criteria of efficiency

Label-conditionally S-optimal, *N-optimal*, *OF-optimal*, and *OE-optimal* idealised conformity measures are defined exactly as S-optimal, N-optimal, OF-optimal, and OE-optimal idealised conformity measures at the end of Section 3 but with the label-conditional definitions of the p-values and prediction sets.

Let us say that an idealised conformity measure A is a *label-conditional refinement* of an idealised conformity measure B if

$$B(x_1, y) < B(x_2, y) \Longrightarrow A(x_1, y) < A(x_2, y)$$

for all $x_1, x_2 \in \mathbf{X}$ and all $y \in \mathbf{Y}$. Notice that the notion of label-conditional refinement is weaker than that of refinement (as defined by (15)): if A is a refinement of B, then A is a label-conditional refinement of B (but not vice versa, in general). Let $\mathcal{R}_{\mathrm{lc}}(\mathrm{CP})$ be the set of all label-conditional refinements of the CP idealised conformity measure. If C is a criterion of efficiency (one of the ten criteria in Table 1), we let $\mathcal{O}_{\mathrm{lc}}(C)$ stand for the set of all label-conditionally C-optimal idealised conformity measures. We have the following simple corollary of Theorem 1.

Theorem 5 $\mathcal{O}_{\mathrm{lc}}(\mathrm{S}) = \mathcal{O}_{\mathrm{lc}}(\mathrm{OF}) = \mathcal{O}_{\mathrm{lc}}(\mathrm{N}) = \mathcal{O}_{\mathrm{lc}}(\mathrm{OE}) = \mathcal{R}_{\mathrm{lc}}(\mathrm{CP})$.

Proof The proof is a modification of the proof of Theorem 1. In the case of $\mathcal{O}_{\mathrm{lc}}(\mathrm{N})$, for each label $y \in \mathbf{Y}$ we have a separate optimization problem. Now the constraint becomes

$$\sum_x Q(x, y)P(x, y) = \epsilon Q_{\mathbf{Y}}(y)$$

(in place of (16)), and the objective becomes to maximise $\sum_x Q'(x, y)P(x, y)$ (since maximising the sum over (x, y) in (17) can be achieved by maximizing the sum over x for each y separately). Now an application of the Neyman–Pearson lemma, as in the proof of Theorem 1, shows that $\mathcal{O}_{\mathrm{lc}}(\mathrm{N}) = \mathcal{R}_{\mathrm{lc}}(\mathrm{CP})$.

The same argument as in the proof of Theorem 1 (the last three paragraphs) shows that $\mathcal{O}_{\mathrm{lc}}(\mathrm{N}) = \mathcal{O}_{\mathrm{lc}}(\mathrm{S}) = \mathcal{O}_{\mathrm{lc}}(\mathrm{OF}) = \mathcal{O}_{\mathrm{lc}}(\mathrm{OE})$, and so we have the formula in Theorem 5. □

We say that an efficiency criterion is *label-conditionally probabilistic* if the CP idealised conformity measure is label-conditionally optimal for it; we add the qualifier *weakly* if this is true for some (label-conditional) refinement of CP and *strongly* if this is true for an arbitrary (label-conditional) refinement of CP. We can see that the four criteria that are set in italics in Table 1 are still optimal in the label-conditional setting.

8.4 Other criteria of efficiency

Using the label-conditional definitions of the p-values and prediction sets, we define *label-conditionally U-optimal*, *M-optimal*, *F-optimal*, *E-optimal*, *OU-optimal*, and *OM-optimal* idealised conformity measures in exactly the same way as their unconditional counterparts at the beginning of Section 5. The label-conditional U and M criteria are standard, and the label conditional E criterion (with a different treatment of empty observations) has been introduced and explored in [15].

We do not give label-conditional analogues of Theorems 2–4, since the label-conditionally U-, M-, F-, E-, OU-, and OM-optimal idealised conformity measures are unlikely to have explicit expressions (cf. our remark about deterministic conformal predictors on p. 2), unless $|\mathbf{Y}| = 2$. The following theorem says that all of these criteria are BW probabilistic (and the examples that we will give after its proof will show that they are not probabilistic).

Theorem 6 *If $|\mathbf{Y}| = 2$, each of the sets*

$$\mathcal{O}_{\mathrm{lc}}(\mathrm{U}), \mathcal{O}_{\mathrm{lc}}(\mathrm{M}), \mathcal{O}_{\mathrm{lc}}(\mathrm{F}), \mathcal{O}_{\mathrm{lc}}(\mathrm{E}), \mathcal{O}_{\mathrm{lc}}(\mathrm{OU}), \mathcal{O}_{\mathrm{lc}}(\mathrm{OM}) \tag{44}$$

contains a refinement of the CP idealised conformity measure.

Proof Assume, without loss of generality, that $\mathbf{Y} = \{0, 1\}$. And let us assume, for simplicity, that the values $Q(1 \mid x)$ are all different for different $x \in \mathbf{X}$ (if this condition is not satisfied, the theorem still holds, but finding a suitable refinement becomes, in general, a difficult combinatorial problem). In this case it is easy to see that each of the sets in (44) is the equivalence class of the CP idealised conformity measure: we can construct the optimal idealised conformity measure gradually starting from small values of ϵ, as in the proofs of Theorems 2–4. □

The following examples show that none of the criteria considered in this subsection is probabilistic (or even weakly probabilistic):

- Let $\mathbf{X} = \{1, 2\}$, $\mathbf{Y} = \{1, 2, 3, 4\}$, and

$$\begin{array}{lllll} Q_{\mathbf{X}}(1) = 0.5 & Q(1 \mid 1) = 0.2 & Q(2 \mid 1) = 0.3 & Q(3 \mid 1) = 0.2 & Q(4 \mid 1) = 0.3 \\ Q_{\mathbf{X}}(2) = 0.5 & Q(1 \mid 2) = 0.3 & Q(2 \mid 2) = 0.2 & Q(3 \mid 2) = 0.3 & Q(4 \mid 2) = 0.2 \end{array} \tag{45}$$

  ($Q(y \mid x)$ meaning $Q_{\mathbf{Y}|\mathbf{X}}(y \mid x)$, as usual). All refinements of the CP idealised conformity measure are equivalent (as for different labels y the two conditional probabilities $Q(y \mid x)$, $x = 1, 2$, are different), and so all of them will lead to the same p-values. Let A be any idealised conformity measure that makes all observations containing object 1 less conforming than all observations containing object 2. The U criterion is not probabilistic since the expression (35) is 0.7 for the CP idealised conformity measure and is smaller, 0.55, for the idealised conformity measure A. The M criterion is not probabilistic since at significance level $\epsilon = 0.4$ the CP idealised conformity measure gives the predictor $\Gamma^\epsilon(1) = \{2, 4\}$ and $\Gamma^\epsilon(2) = \{1, 3\}$ (a.s.), and so

$$\mathbb{P}_{x,\tau}(|\Gamma^\epsilon_{\mathrm{CP}}(x)| > 1) = 1 > 2/3 = \mathbb{P}_{x,\tau}(|\Gamma^\epsilon_A(x)| > 1)$$

  (cf. (23)).
- Let $\mathbf{X} = \{1, 2, 3\}$, $\mathbf{Y} = \{1, 2, 3\}$, and, for a small $\delta > 0$,

$$\begin{array}{llll} Q_{\mathbf{X}}(1) = 1/3 & Q(1 \mid 1) = 1/3 + \delta & Q(2 \mid 1) = 1/3 - 2\delta & Q(3 \mid 1) = 1/3 + \delta \\ Q_{\mathbf{X}}(2) = 1/3 & Q(1 \mid 2) = 1/3 - \delta & Q(2 \mid 2) = 1/3 + 2\delta & Q(3 \mid 2) = 1/3 - \delta \\ Q_{\mathbf{X}}(3) = 1/3 & Q(1 \mid 3) = 1/3 & Q(2 \mid 2) = 1/3 & Q(3 \mid 3) = 1/3. \end{array}$$

  All refinements of the CP idealised conformity measure are equivalent, and so the choice of the refinement does not affect the p-values. Let A be an idealised conformity measure satisfying

$$A(1,2) < A(2,1) = A(2,3) < A(3,1) = A(3,2) < A(1,1) = A(1,3) < A(2,2) < A(3,3)$$

  (in other words, A is the CP idealised conformity measure modified in such a way that that it assigns to $(3, 3)$ the highest conformity score). The F criterion

is not probabilistic since the expression (36) is $7/9 + O(\delta)$ for the CP idealised conformity measure and is smaller (for sufficiently small δ), $2/3 + O(\delta)$, for A. The E criterion is not probabilistic since at significance level $\epsilon = 2/3$ the idealised conformity measure A gives a predictor whose excess is always 0, whereas the CP idealised conformity measure will have expected excess $1/3 + O(\delta)$.
- Let $\mathbf{X} = \{1, 2\}$, $\mathbf{Y} = \{1, 2, 3, 4\}$, and Q be defined by (45). Let A be any idealised conformity measure that makes all observations containing object 1 less conforming than all observations containing object 2. The OU criterion is not probabilistic since the expression (37) is 0.7 for the CP idealised conformity measure and is smaller, 0.55, for the idealised conformity measure A. The OM criterion is not probabilistic since at significance level $\epsilon = 0.4$ the CP idealised conformity measure produces an observed multiple prediction a.s., whereas the idealised conformity measure A produces an observed multiple prediction with probability 2/3.

9 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 of the p-value as function of the label y of the test example, and the size of a prediction set becomes its Lebesgue measure (considered, as already mentioned, in [11] 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 [17], which considers the average p-value as its criterion of efficiency.
- Infinite, including non-discrete, object spaces $\mathbf{X}$.
- Non-idealised conformal predictors.
- Significance levels $\epsilon = \epsilon_y$ that depend on the label $y \in \mathbf{Y}$ in the label-conditional case.

The main part of this paper merely mentions what we called "combinatorial problems" (see pages 2 and 24). It would be interesting to explore them systematically. As an example, let us consider the N criterion of efficiency for deterministic idealised conformal predictors (with τ set to 1 rather than being random) in the trivial case $|\mathbf{Y}| = 1$ (which we did not allow in the main part of the paper; in this case, there is no difference between unconditional and label-conditional idealised conformal predictors; computational difficulties can be expected to become more severe in less trivial cases). The problem of finding an N-optimal idealised conformity measure then becomes the Subset-Sum Problem, known to be NP-hard: see, e.g., [12], Chapter 4 (a special case of this problem, Partition, was already one of Karp's original 21 NP-complete problems [6]). There are, however, efficient polynomial approximation schemes for this problem. It would be interesting, in particular, to find such schemes for general deterministic idealised conformal predictors and transducers and for smoothed idealised conformal predictors and transducers for non-probabilistic criteria of efficiency in the label-conditional case.

Acknowledgements We are grateful to the reviewers of the conference and journal versions of this paper for their helpful comments.

References

1. Balasubramanian, V.N., Ho, S.S., Vovk, V. (eds.): Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications. Elsevier, Amsterdam (2014)
2. Dawid, A.P.: Probability forecasting. In: S. Kotz, N. Balakrishnan, C.B. Read, B. Vidakovic, N.L. Johnson (eds.) Encyclopedia of Statistical Sciences, vol. 10, second edn., pp. 6445–6452. Wiley, Hoboken, NJ (2006)
3. Fedorova, V., Gammerman, A., Nouretdinov, I., Vovk, V.: Hypergraphical conformal predictors. International Journal on Artificial Intelligence Tools **24**(6), 1560,003 (2015). COPA 2013 Special Issue
4. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**, 359–378 (2007)
5. Johansson, U., König, R., Löfström, T., Boström, H.: Evolved decision trees as conformal predictors. In: L.G. de la Fraga (ed.) Proceedings of the 2013 IEEE Conference on Evolutionary Computation, vol. 1, pp. 1794–1801. Cancun, Mexico (2013)
6. Karp, R.M.: Reducibility among combinatorial problems. In: R.E. Miller, J.W. Thatcher (eds.) Complexity of Computer Computations, pp. 85–103. Plenum Press, New York (1972)
7. Le Cun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: D.S. Touretzky (ed.) Advances in Neural Information Processing Systems 2, pp. 396–404. Morgan Kaufmann, San Francisco, CA (1990)
8. Lehmann, E.L.: Testing Statistical Hypotheses, second edn. Springer, New York (1986)
9. Lei, J.: Classification with confidence. Biometrika **101**, 755–769 (2014)
10. Lei, J., Robins, J., Wasserman, L.: Distribution free prediction sets. Journal of the American Statistical Association **108**, 278–287 (2013)
11. Lei, J., Wasserman, L.: Distribution free prediction bands for nonparametric regression. Journal of the Royal Statistical Society B **76**, 71–96 (2014)
12. Martello, S., Toth, P.: Knapsack Problems: Algorithms and Computer Implementations. Wiley, Chichester (1990)
13. Melluish, T., Saunders, C., Nouretdinov, I., Vovk, V.: Comparing the Bayes and typicalness frameworks. In: L. De Raedt, P.A. Flach (eds.) Proceedings of the Twelfth European Conference on Machine Learning, *Lecture Notes in Computer Science*, vol. 2167, pp. 360–371. Springer, Heidelberg (2001)
14. Papadopoulos, H., Gammerman, A., Vovk, V. (eds.): Special Issue of the *Annals of Mathematics and Artificial Intelligence* on Conformal Prediction and its Applications, vol. 74(1–2). Springer (2015)
15. Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Tech. Rep. `arXiv:1609.00451v1 [stat.ME]`, `arXiv.org` e-Print archive (2016)
16. Saunders, C., Gammerman, A., Vovk, V.: Transduction with confidence and credibility. In: T. Dean (ed.) Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, vol. 2, pp. 722–726. Morgan Kaufmann, San Francisco, CA (1999)
17. Smith, J., Nouretdinov, I., Craddock, R., Offer, C., Gammerman, A.: Anomaly detection of trajectories with kernel density estimation by conformal prediction. In: L. Iliadis, I. Maglogiannis, H. Papadopoulos, S. Sioutas, C. Makris (eds.) AIAI Workshops, COPA 2014, *IFIP Advances in Information and Communication Technology*, vol. 437, pp. 271–280 (2014)
18. Vovk, V., Fedorova, V., Nouretdinov, I., Gammerman, A.: Criteria of efficiency for conformal prediction. In: A. Gammerman, Z. Luo, J. Vega, V. Vovk (eds.) Proceedings of the Fifth International Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2016), *Lecture Notes in Artificial Intelligence*, vol. 9653, pp. 23–39. Springer, Switzerland (2016). To appear in the *Annals of Mathematics and Artificial Intelligence* (COPA 2016 Special Issue)
19. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
20. Vovk, V., Petej, I., Fedorova, V.: From conformal to probabilistic prediction. In: L. Iliadis, I. Maglogiannis, H. Papadopoulos, S. Sioutas, C. Makris (eds.) AIAI Workshops, COPA 2014, *IFIP Advances in Information and Communication Technology*, vol. 437, pp. 221–230 (2014)

Authors' Response to Reviewers' Comments
Click here to download Authors' Response to Reviewers' Comments changes2.tex

January 13, 2017

Dr Zhiyuan Luo
Guest Editor, *Annals of Mathematics and Artificial Intelligence*

Dear Dr Luo

We are very grateful to you, the Springer staff, and the referees for the time and effort all of you invested in reading and commenting on our paper. In this letter we will describe the changes that we have made to the paper in response to the referees' comments; when quoting them, we use the *slanted font*. By the "old version" of the paper we will mean the version that was refereed, and the "new version" is the attached draft taking all comments into account; by default, the page and line numbers refer to the old version of the paper.

Changes made in response to Referee #1's comments

These are our replies and changes:

Firstly, I have a general comment: I do not think the references to Sadinle, Lei, and Wasserman (2016, henceforth SLW) are accurate. The efficiency criterion in SLW does not correspond to the E criterion but to the N criterion. In fact, the optimal conformity measure in SLW is the conditional probability $p(y \mid x)$, so it's probabilistic (following the definition in Section 4). ➳ We agree that it looks as if the criterion used in SLW is the N criterion: e.g., on page 2 SLW state their criterion as minimizing the ambiguity

$$\mathbb{A}(H) = \mathbb{E}\{|H(X)|\}. \tag{A}$$

According to this criterion, the ideal predictions are those for which $|H(X)| = 0$, i.e., empty predictions. However, the authors (SLW) do not take the expression (A) seriously. In the next paragraph they declare their desire to avoid empty predictions ("a potentially undesirable property of the optimal classifiers"). Their real optimization problem is different from the one they announce (minimizing the ambiguity). In Section 3 they struggle with eliminating empty predictions. Therefore, their criterion is much closer to our E criterion than to our N criterion. However, their criterion goes further: whereas the E criterion does not care about empty predictions, the real SLW criterion involves eliminating them. We did not give the SLW definitions in our paper (an interested reader can easily check them) and tried to make our statements vague enough so that they do not contradict [15]; in particular, this is what we say about the SLW criterion:

> A version (with a different treatment of empty observations) of one of the new non-probabilistic criteria of efficiency that we discuss in this paper (the one that we call the E criterion) has been introduced independently in [15].

1. The abstract could be more informative. For example mention the conclusion of p. 22 line 14. ➳ That conclusion is local to Section 7, and the abstract would be too prominent a place for it. We have made the abstract more informative by mentioning that we argue for probabilistic criteria. (We also made "classification" more specific by prefixing it with "set-valued".)

2. P. 2, first sentence. Could you briefly mention here the disadvantage of the standard criteria? I think you only mention this again much later in the paper, at the end of Section 5. ➳ The disadvantage is difficult to describe in the introduction, so we have just added a forward reference to the end of Section 5.

3. P. 2, line 48. Could you explain the meaning of "deterministic predictor"? ➳ We have added "non-smoothed" in parentheses; is it clearer now?

4. P. 4. This is just a comment: the criteria based on p-values are difficult to interpret. For example, think of a perfect scenario where there's no ambiguity in the predictions. In such case each point gets only one prediction label and therefore it's easy to find what the value of the criteria based on the predictors should be. It is however difficult to see what the value should be when using the p-values. ➳ This appears to be a valuable comment, but we do not understand it, sorry. If you get a chance, could you please elaborate on your idea?

5. P. 4, line 47. "even for a very efficient conformal predictor we cannot expect the size of its prediction set to be zero." Why would we consider having a prediction set with size zero to be efficient? ➳ We will split this question (consisting of three parts) into three questions. One possible meaning of efficiency is just prediction regions being small, and then, formally, empty is efficient. (But we should worry about efficiency only for valid predictors.)

What is the meaning of an empty prediction set? ➳ We discuss this in the answer to the first question on page 296 of [19].

This is related to Equation (9): why'd we consider larger numbers of empty predictions as something good? ➳ We do not regard empty predictions as being good; we just say that they contribute to efficiency (potentially endangering validity). On page 2, line 23, we refer to empty predictions as superefficient, and superefficiency is not regarded, as far as we know, a good property in statistics.

6. Consider transforming Section 6 into an appendix. ➳ We have thought about this, but realized that transforming Section 6 into an appendix would make the structure of the paper too complicated, since the proofs of Section 6 are used in Section 8. (Moving the proofs in Section 8 to the same appendix is much less natural since those proofs are so simple, and then it is not clear what to do with the examples in that section.)

7. P. 22, eq. (42): there's an extra τ in the numerator of p^y. Is this a typo? if not, could you explain why it appears here? ➳ The second τ corresponds to $i = l + 1$ in (2).

8. P. 26. line 30. The case of label-dependent significance levels was developed in SLW. ➳ This case is very standard. Label-dependent significance levels were used in, e.g., [19], the bottom of page 115 (and probably in earlier papers as well).

9. P. 26, line 37. Why is the case $|Y| = 1$ interesting? in such case there's no need for predictions. ➳ This is not an interesting case per se, but already this extremely simple case demonstrates the computational difficulties. We have revised our description to make our meaning clearer.

10. Should the title be changed to "Criteria of efficiency for conformal set-valued classification"? ➳ Have you seen `changes.pdf`? (We know that the journal system is not working properly; it took a while for us to get hold of your pdf report.) We say there:

> We have changed the title of the conference paper to make it more general: many of our results are applicable outside the theory of conformal prediction.

Changes made in response to Referee #2's comments

Thank you for your positive comments. These are the changes made in response to your minor critical comments:

First, it is not clear which efficiency measures are new and which have been introduced earlier. Although the authors provide references in the text, it would be very helpful if they can add remarks in Table 1, to indicate which papers (if any) first proposed each efficiency measure. ➳ This is not easy to do, since the criteria used in various papers are often modifications of our criteria: see, e.g., p. 2, line 46 and p. 6, lines 14–20; it is not clear how to treat the previous version of this paper and our other papers that refer to this paper (such as [3]). Finally, taking into account the importance of Table 1, we have decided not to overload it with extra information.

Second, the authors mention in the introduction that the non-probabilistic efficiency measures have some disadvantages. It is not clear what these advantages are. Please elaborate, both theoretically and empirically. ➳ This is explained at the end of Section 5, and we have added a forward reference to it at the top of page 2 in Section 1.

Third, this paper mostly focuses on ideal conformity scores. In practice, how would these be estimated? Are some scores easier to approximate than others? ➳ We do not propose to estimate the ideal conformity scores. Our idea is that researchers will be implicitly encouraged to use the conditional probabilities (or a monotonic function thereof) when the performance of their prediction algorithms is measured using probabilistic criteria of efficiency.

Changes made in response to Referee #3's comments

We are grateful for your positive comments; these are our responses to your suggestions:

Authors state the conformal predictors and transducer must take into account conformity order rather than conformity measurements. Does it mean that it depends on the p-values? Therefore, are they still e-free? ➳ We do not think it can be said that conformal predictors and transducers must take into account conformity order rather than conformity scores (measurements). In fact conformity orders and conformity measure are equivalent for our purposes, leading to the same p-values. The distinction between ϵ-free and ϵ-dependent criteria is orthogonal to using conformity orders or conformity scores: both conformity orders and conformity scores can be used in defining both kinds of criteria.

Figure 2 shows the average values of some criteria for two distances: tangent and Euclidean. Why those definitions of distance? ➳ This is now discussed on page 21 after the itemized list.

Are there other examples of real applications of these criteria in the literature? ➳ There are plenty of real applications for non-probabilistic criteria. We would like to see more application of probabilistic ones. References are given throughout the paper.

Sincerely yours

The Authors

encl: New version of the paper