Conditional validity of inductive conformal predictors

Vladimir Vovk

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

On-line Compression Modelling Project (New Series)

Working Paper #5

First posted September 11, 2012. Last revised April 19, 2013.

Project web site:
http://alrw.net

Abstract

Conformal predictors are set predictors that are automatically valid in the sense of having coverage probability equal to or exceeding a given confidence level. Inductive conformal predictors are a computationally efficient version of conformal predictors satisfying the same property of validity. However, inductive conformal predictors have been only known to control unconditional coverage probability. This paper explores various versions of conditional validity and various ways to achieve them using inductive conformal predictors and their modifications. In particular, it discusses a convenient expression of one of the modifications in terms of ROC curves.

Contents

1 Introduction

This paper continues study of the method of conformal prediction, introduced in Vovk et al. (1999) and Saunders et al. (1999) and further developed in Vovk et al. (2005). An advantage of the method is that its predictions (which are set rather than point predictions) automatically satisfy a finite-sample property of validity. Its disadvantage is its relative computational inefficiency in many situations. A modification of conformal predictors, called inductive conformal predictors was proposed in Papadopoulos et al. (2002a,b) with the purpose of improving on the computational efficiency of conformal predictors. For further information on conformal predictors and inductive conformal predictors see, e.g., Balasubramanian et al. (2013) and Papadopoulos et al. (2013).

Most of the literature on conformal prediction studies the behaviour of set predictors in the online mode of prediction, perhaps because the property of validity can be stated in an especially strong form in the on-line mode (as first shown in Vovk 2002). The online mode, however, is much less popular in applications of machine learning than the batch mode of prediction. This paper follows the recent papers by Lei et al. (2013) and Lei and Wasserman (2013) studying properties of conformal prediction in the batch mode; we, however, concentrate on inductive conformal prediction. The performance of inductive conformal predictors in the batch mode is illustrated using the well-known `Spambase` data set; for earlier empirical studies of conformal prediction in the batch mode see, e.g., Vanderlooy et al. (2007). The conference version of this paper is published as Vovk (2012).

We will usually be making the *assumption of randomness*, which is standard in machine learning and nonparametric statistics: the available data is a sequence of *examples* generated independently from the same probability distribution Q. (In some cases we will make the weaker assumption of exchangeability; for some of our results even weaker assumptions, such as conditional randomness or exchangeability, would have been sufficient.) Each example consists of two components: an *object* and a *label*. We are given a *training set* of examples and a new object, and our goal is to predict the label of the new object. (If we have a whole *test set* of new objects, we can apply the procedure for predicting one new label to each of the objects in the test set.)

The two desiderata for inductive conformal predictors are their validity and efficiency: validity requires that the coverage probability of the prediction sets should be at least equal to a preset confidence level, and efficiency requires that the prediction sets should be as small as possible. However, there is a wide variety of notions of validity, since the "coverage probability" is, in general, conditional probability. The simplest case is where we condition on the trivial σ-algebra, i.e., the probability is in fact unconditional probability, but several other notions of conditional validity are depicted in Figure 1, where T refers to conditioning on the training set, O to conditioning on the test object, and L to conditioning on the test label. The arrows in Figure 1 lead from stronger to weaker notions of conditional validity; U is the sink and TOL is the source (the latter is not shown).

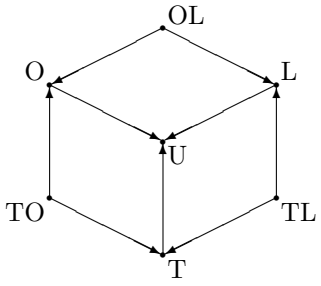

Figure 1: Eight notions of conditional validity. The visible vertices of the cube are U (unconditional), T (training conditional), O (object conditional), L (label conditional), OL (example conditional), TL (training and label conditional), TO (training and object conditional). The invisible vertex is TOL (and corresponds to conditioning on everything).

Inductive conformal predictors (slightly generalized as compared with the standard version) will be defined in Section 2. They are automatically valid, in the sense of unconditional validity. It should be said that, in general, the unconditional error probability is easier to deal with than conditional error probabilities; e.g., the standard statistical methods of cross-validation and bootstrap provide decent estimates of the unconditional error probability but poor estimates for the training conditional error probability: see Hastie et al. (2009), Section 7.12.

In Section 3 we explore training conditional validity of inductive conformal predictors. Our simple results (Theorem 1 and Corollaries 1 and 2) are of the PAC type, involving two parameters: the target training conditional coverage probability $1 - \epsilon$ and the probability $1 - \delta$ with which $1 - \epsilon$ is attained. They show that inductive conformal predictors achieve training conditional validity automatically (whereas for other notions of conditional validity the method has to be modified). We give a self-contained proof of Theorem 1, but Appendix A explains how its significant part can be deduced from classical results about tolerance regions.

In the following section, Section 4, we introduce a conditional version of inductive conformal predictors and explain, in particular, how it achieves label conditional validity. Label conditional validity is important as it allows the learner to control the set-prediction analogues of false positive and false negative rates. Section 5 is about object conditional validity and its main result (a version of a lemma in Lei and Wasserman 2013) is negative: precise object conditional validity cannot be achieved in a useful way unless the test object has a positive probability. Whereas precise object conditional validity is usually not achievable, we should aim for approximate and asymptotic object conditional validity when given enough data (cf. Lei and Wasserman 2013).

Section 6 reports on the results of empirical studies for the standard `Spambase` data set (see, e.g., Hastie et al. 2009, Chapter 1, Example 1, and

Section 9.1.2). Section 7 discusses close connections between an important class of label conditional ICPs and ROC curves. Section 8 concludes the main part of the paper, and two appendixes are devoted to related approaches to set prediction. Appendix A discusses connections with the classical theory of tolerance regions (in particular, it explains how part of Theorem 1 can be deduced from classical results about the training conditional validity of tolerance regions). Appendix B discusses training conditional validity of conformal predictors.

2 Inductive conformal predictors

The example space will be denoted $\mathbf{Z}$; it is the Cartesian product $\mathbf{X} \times \mathbf{Y}$ of two measurable spaces, the object space $\mathbf{X}$ and the label space $\mathbf{Y}$. In other words, each example $z \in \mathbf{Z}$ consists of two components: $z = (x, y)$, where $x \in \mathbf{X}$ is its object and $y \in \mathbf{Y}$ is its label. Two important special cases are the problem of *classification*, where $\mathbf{Y}$ is a finite set (equipped with the discrete σ-algebra), and the problem of *regression*, where $\mathbf{Y}$ is the real line $\mathbb{R}$.

Various predictors defined and discussed in this paper are randomized: they depend, in addition to the data, on an element $\omega \in \bar{\Omega}$ of a measurable space $\bar{\Omega}$ equipped with a probability distribution R (the "coin-tossing" distribution). This is important to cover various predictors based on the MART procedure, which is randomized and used in our computational experiments in Section 6.

Let $(z_1, \dots, z_l)$ be the training set, $z_i = (x_i, y_i) \in \mathbf{Z}$. We split it into two parts, the *proper training set* $(z_1, \dots, z_m)$ of size $m < l$ and the *calibration set* of size $n := l - m$. An *inductive conformity m-measure* is a measurable function $A : \mathbf{Z}^m \times \mathbf{Z} \times \bar{\Omega} \to \mathbb{R}$; the idea behind the *conformity score* $A((z_1, \dots, z_m), z, \omega)$ is that it should measure how well z conforms to the proper training set. We omit "m-" when it is clear from the context. A standard choice of an inductive conformity measure is

$$A((z_1, \dots, z_m), (x, y), \omega) := \Delta(y, f(x)), \tag{1}$$

where $f : \mathbf{X} \to \mathbf{Y}'$ is a prediction rule found (perhaps using a randomized procedure) from $(z_1, \dots, z_m)$ as the training set and $\Delta : \mathbf{Y} \times \mathbf{Y}' \to \mathbb{R}$ is a measure of similarity between a label and a prediction. Allowing $\mathbf{Y}'$ to be different from $\mathbf{Y}$ (often $\mathbf{Y}' \supset \mathbf{Y}$) may be useful when the underlying prediction method gives additional information to the predicted label; e.g., the MART procedure used in Section 6 gives the logit of the predicted probability that the label is 1.

Remark. The idea behind the term "calibration set" is that this set allows us to calibrate the conformity scores of test examples by translating them into a probability-type scale.

The *inductive conformal predictor* (ICP) corresponding to A is defined as the set predictor

$$\Gamma^\epsilon(z_1, \dots, z_l, x, \omega) := \{y \mid p^y > \epsilon\}, \tag{2}$$

where $\epsilon \in (0,1)$ is the chosen *significance level* ($1-\epsilon$ is known as the *confidence level*), the *p-values* p^y, $y \in \mathbf{Y}$, are defined by

$$p^y := \frac{|\{i = m+1, \ldots, l \mid \alpha_i \le \alpha^y\}| + 1}{l - m + 1}, \tag{3}$$

and

$$\alpha_i := A((z_1, \ldots, z_m), z_i, \omega), \quad i = m+1, \ldots, l,$$
$$\alpha^y := A((z_1, \ldots, z_m), (x, y), \omega) \tag{4}$$

are the conformity scores. Given the training set and a new object x the ICP predicts its label y; it *makes an error* if $y \notin \Gamma^\epsilon(z_1, \ldots, z_l, x, \omega)$. All predictors considered in this paper are randomized, and so we omit the word "randomized".

We consider a canonical probability space Δ whose elements are all possible sequences $z_i = (x_i, y_i)$, $i = 1, \ldots, l+1$, of $l+1$ examples and which is equipped with a probability distribution P. Random variables $Z_i = (X_i, Y_i)$, $i = 1, \ldots, l+1$, are projections of this probability space onto its ith coordinate: $Z_i(z_1, \ldots, z_{l+1}) := z_i$, $X_i(z_1, \ldots, z_{l+1}) := x_i$, and $Y_i(z_1, \ldots, z_{l+1}) := y_i$. We often let x_i, y_i, and z_i stand for realizations of the random variables X_i, Y_i, and Z_i, respectively. Our overall probability space is $\Delta \times \bar{\Omega} \times [0,1]$, and it is equipped with the product measure $P \times R \times U$, where R is the coin-tossing distribution mentioned above and U is the uniform probability distribution on $[0,1]$ (we will need U in the definition of "smoothed" ICP below). The generic element of $\Delta \times \bar{\Omega} \times [0,1]$ will usually be denoted $(z_1, \ldots, z_{l+1}, \omega, \theta)$, and the projections onto the last two components will be denoted $\Omega(z_1, \ldots, z_{l+1}, \omega, \theta) := \omega$ and $\Theta(z_1, \ldots, z_{l+1}, \omega, \theta) := \theta$; Z_i will also be regarded as random variables on the overall probability space that ignore the last two coordinates. In cases where θ is irrelevant we will also consider the probability space $\Delta \times \bar{\Omega}$ equipped with the probability distribution $P \times R$. It will always be clear from the context which of the three probability spaces we are talking about.

Smoothed inductive conformal predictors are defined as ICPs except that (3) is replaced by

$$p^y := \frac{|\{i = m+1, \ldots, l \mid \alpha_i < \alpha^y\}| + \theta\,(|\{i = m+1, \ldots, l \mid \alpha_i = \alpha^y\}| + 1)}{l - m + 1}; \tag{5}$$

therefore, Γ^ϵ now depends on θ as well (remember that θ stands for values taken by the random variable Θ distributed uniformly on $[0,1]$).

Remark. The smoothed inductive conformal predictors defined in this section are more general than the corresponding smoothed predictors considered in Vovk et al. (2005): the former involve not only the tie-breaking random variable Θ but also randomized conformity measures. However, this generalization is straightforward: we get it essentially for free.

Proposition 1 (Vovk et al., 2005, Proposition 4.1)**.** *Let random examples $Z_{m+1}, \ldots, Z_l, Z_{l+1} = (X_{l+1}, Y_{l+1})$ be exchangeable (i.e., their distribution P is invariant under permutations). The probability of error*

$Y_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l, X_{l+1}, \Omega)$ *does not exceed* ϵ *for any* ϵ *and any inductive conformal predictor* Γ. *The probability of error* $Y_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l, X_{l+1}, \Omega, \Theta)$ *is equal to* ϵ *for any* ϵ *and any smoothed inductive conformal predictor* Γ.

This simple proposition of validity is proved in Vovk et al. (2005) for inductive conformal predictors based on deterministic inductive conformity measures, but integration over $\bar{\Omega}$ immediately yields Proposition 1. In practice the probability of error is usually close to ϵ even for unsmoothed ICPs (as we will see in Section 6 and Appendix B).

In conclusion of this section, let me give two specific examples of ICPs. Since an ICP is determined by its inductive conformity measure, it suffices to specify the latter.

- In the case of regression, $\mathbf{Y} = \mathbb{R}$, we can define the inductive conformity measure by (1) where $\Delta(y, f(x)) := -|y - f(x)|$ and f is the prediction rule found by using ridge regression from $(z_1, \dots, z_m)$ as the training set. This ICP is the inductive counterpart of the Ridge Regression Confidence Machine (Vovk et al. 2005, Section 2.3).

- An example not covered by the scheme (1) is the *1-Nearest Neighbour ICP*, whose inductive conformity measure is

$$A((z_1, \dots, z_m), (x, y), \omega) := \frac{\min_{i=1,\dots,m:y_i \neq y} d(x, x_i)}{\min_{i=1,\dots,m:y_i = y} d(x, x_i)}, \tag{6}$$

  where d is a distance on $\mathbf{X}$. Intuitively, an example conforms to the proper training set if it is closer to the examples labelled in the same way than to those labelled differently. In the case of classification, this ICP will be called the *1-Nearest Neighbour ICP*.

Another example, based on boosting, will be given in Section 6. For numerous other examples, see Vovk et al. (2005), Section 4.2.

3 Training conditional validity

As discussed in Section 1, the standard property of validity of inductive conformal predictors is unconditional. The property of training conditional validity can be formalized using a PAC-type 2-parameter definition. It will be convenient to represent the ICP (2) in a slightly different form downplaying the structure (x_i, y_i) of z_i. Define $\Gamma^\epsilon(z_1, \dots, z_l, \omega) := \{(x, y) \mid p^y > \epsilon\}$, where p^y is defined, as before, by (3) and (4) (therefore, p^y depends implicitly on x). In this notation the first part of Proposition 1 can be restated by saying that the probability of error $Z_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l, \Omega)$ does not exceed ϵ provided $Z_1, \dots, Z_{l+1}$ are exchangeable. We will also use similar conventions in the smoothed case.

A set predictor Γ (outputting a subset of $\mathbf{Z}$ given l examples and measurable in the sense of the set $\{Z_{l+1} \in \Gamma(Z_1, \dots, Z_l, \Omega, \Theta)\}$ being measurable) is (ϵ, δ)-*valid with respect to* a probability distribution Q on $\mathbf{Z}$ if

$$(Q^{l+1} \times R \times U)\left(Q(\Gamma(Z_1, \dots, Z_l, \Omega, \Theta)) \geq 1 - \epsilon\right) \geq 1 - \delta$$

(we will apply this definition to both smoothed and unsmoothed ICPs, even though the latter in fact do not depend on θ). We say that Γ is (ϵ, δ)-*valid* if it is (ϵ, δ)-valid with respect to any probability distribution Q on $\mathbf{Z}$. Our next result (Theorem 1 below) says that ICPs satisfy this property for suitable ϵ and δ; we will see, however, that this is not true for smoothed ICPs in general. Some conditions in the statement of Theorem 1 are not straightforward to interpret; for more explicit conditions, see Corollaries 1 and 2.

Let Z be the random variable $Z(z) := z$ on the measurable space $\mathbf{Z}$ (equipped with a probability distribution usually denoted Q). We will say that an inductive conformity measure is *continuous* under a probability distribution Q on $\mathbf{Z}$ if, for Q^m-almost all $(z_1, \dots, z_m) \in \mathbf{Z}^m$ and R-almost all $\omega \in \bar{\Omega}$, the random variable $A((z_1, \dots, z_m), Z, \omega)$ on the probability space $(\mathbf{Z}, Q)$ is continuous.

Theorem 1. *Let* $\mathrm{bin}_{n,E}$ *be the cumulative binomial distribution function with* n *trials and probability of success* E*; set* $\mathrm{bin}_{n,E}(-1) := 0$.

(a) Let Γ *be an inductive conformal predictor. Suppose that* $\epsilon, \delta, E \in (0,1)$ *satisfy*

$$\delta \geq \mathrm{bin}_{n,E}\left(\lfloor \epsilon(n+1) - 1 \rfloor\right), \tag{7}$$

where $n := l - m$ *is the size of the calibration set. The set predictor* Γ^ϵ *is then* (E, δ)*-valid. Moreover, for any probability distribution* Q *on* $\mathbf{Z}$*, any proper training set* $(z_1, \dots, z_m) \in \mathbf{Z}^m$*, and any* $\omega \in \bar{\Omega}$*,*

$$Q^{l+1}\left(Q(\Gamma^\epsilon(z_1, \dots, z_m, Z_{m+1}, \dots, Z_l, \omega)) \geq 1 - E\right) \geq 1 - \delta. \tag{8}$$

If Γ *is based on an inductive conformity measure that is continuous under* Q*,* Γ^ϵ *is* (E, δ)*-valid with respect to* Q *if and only if (7) holds.*

(b) Let Q *be a probability distribution on* $\mathbf{Z}$ *and* Γ *be a smoothed inductive conformal predictor based on an inductive conformity measure continuous under* Q*. Suppose* $\epsilon, \delta, E \in (0,1)$ *satisfy*

$$\delta \geq \mathrm{bin}_{n,E}\left(\lfloor \epsilon(n+1) \rfloor\right). \tag{9}$$

The set predictor Γ^ϵ *is* (E, δ)*-valid with respect to* Q*. Moreover, for* Q^m*-almost all proper training sets* $(z_1, \dots, z_m) \in \mathbf{Z}^m$*,* R*-almost all* ω*, and all* $\theta \in [0,1]$*,*

$$Q^{l+1}\left(Q(\Gamma^\epsilon(z_1, \dots, z_m, Z_{m+1}, \dots, Z_l, \omega, \theta)) \geq 1 - E\right) \geq 1 - \delta. \tag{10}$$

The set predictor Γ^ϵ *is not* (E, δ)*-valid with respect to* Q *unless* ϵ, δ, E *satisfy (7).*

In the case of smoothed ICPs there is a gap between the sufficient condition (9) and the necessary condition (7), but it does not appear excessive. More worrying is the requirement that the inductive conformity measure be continuous under the unknown data-generating distribution Q. Unfortunately, without this or similar requirement there are no meaningful guarantees of training conditional validity. Indeed, consider the trivial smoothed ICP based on the

inductive conformity measure identically equal to 0. At significance level ϵ, it has coverage probability 1 with probability $1-\epsilon$ and coverage probability 0 with probability ϵ. Therefore, it cannot be (E,δ)-valid for $E<1$ unless $\delta\geq\epsilon$. This contrasts with the case of unsmoothed ICPs where very small δ are achievable: see, e.g., Figure 8 below. Another natural way to define smoothed ICPs is to use different random variables Θ when computing p^y for different labels $y\in\mathbf{Y}$; however, this version also encounters similar problems with training conditional validity when the inductive conformity measure is not required to be continuous under Q.

Proof of Theorem 1. We start from part (a), namely, from proving (8). By (2) and (3), the set predictor Γ^ϵ makes an error, $z_{l+1}\notin\Gamma^\epsilon(z_1,\dots,z_l,\omega)$, if and only if the number of $i=m+1,\dots,l$ such that $\alpha_i\leq\alpha^y$ is at most $\lfloor\epsilon(n+1)-1\rfloor$; in other words, if and only if $\alpha^y<\alpha_{(k)}$, where $\alpha_{(k)}$ is the kth smallest α_i and $k:=\lfloor\epsilon(n+1)-1\rfloor+1$. (Formally, $\alpha_{(k)}$ is defined by the requirement that $\left|\left\{i\,|\,\alpha_i<\alpha_{(k)}\right\}\right|<k\leq\left|\left\{i\,|\,\alpha_i\leq\alpha_{(k)}\right\}\right|$; in other words, $\alpha_{(k)}$ is the kth order statistic.) Therefore, the Q-probability of the complement of $\Gamma^\epsilon(z_1,\dots,z_l,\omega)$ is $Q(A((z_1,\dots,z_m),Z,\omega)<\alpha_{(k)})$, where A is the inductive conformity measure. Set

$$\begin{aligned}\alpha^*&:=\inf\{\alpha\mid Q(A((z_1,\dots,z_m),Z,\omega)<\alpha)>E\}\\&=\inf\{\alpha\mid Q(A((z_1,\dots,z_m),Z,\omega)\leq\alpha)>E\}\\E'&:=Q(A((z_1,\dots,z_m),Z,\omega)<\alpha^*)\\E''&:=Q(A((z_1,\dots,z_m),Z,\omega)\leq\alpha^*).\end{aligned}$$

The σ-additivity of measures implies that $E'\leq E\leq E''$, and $E'=E=E''$ unless α^* is an atom of the distribution of $A((z_1,\dots,z_m),Z,\omega)$. Both when $E'=E$ and when $E'<E$, the probability of error will exceed E if an only if $\alpha_{(k)}>\alpha^*$. In other words, if only if we have at most $k-1$ of the α_i below or equal to α^*. The probability that at most $k-1=\lfloor\epsilon(n+1)-1\rfloor$ values of the α_i are below or equal to α^* equals $\mathbb{P}(B''_n\leq\lfloor\epsilon(n+1)-1\rfloor)\leq\mathbb{P}(B_n\leq\lfloor\epsilon(n+1)-1\rfloor)$, where $B''_n\sim\mathrm{bin}_{n,E''}$, $B_n\sim\mathrm{bin}_{n,E}$, and $\mathrm{bin}_{n,p}$ is also allowed to stand for the binomial distribution with parameters (n,p). (For the inequality, see Lemma 1 below.) This completes the proof of (8) and, therefore, the first two statements of part (a). And the last statement of part (a) follows from the fact that $E''=E$ unless α^* is an atom of the distribution of $A((z_1,\dots,z_m),Z,\omega)$.

Let us now prove part (b), starting from (10). We will assume that the distribution of $A((z_1,\dots,z_m),Z,\omega)$ is continuous (we can do so since (10) is required to hold only for almost all proper training sets and ω). By (5), the set predictor Γ^ϵ can make an error only if the number of $i=m+1,\dots,l$ such that $\alpha_i<\alpha^y$ is at most $\lfloor\epsilon(n+1)\rfloor$ (set $\theta:=0$ in (5) and combine this with $p^y\leq\epsilon$); in other words, only if $\alpha^y\leq\alpha_{(k)}$, where $\alpha_{(k)}$ is the kth smallest α_i and $k:=\lfloor\epsilon(n+1)\rfloor+1$. Therefore, the Q-probability of the complement of $\Gamma^\epsilon(z_1,\dots,z_l,\omega,\theta)$ is at most $Q(A((z_1,\dots,z_m),Z,\omega)\leq\alpha_{(k)})$. Define α^*,E',E'' as before; now we know that $E'=E=E''$. The probability of error can exceed E only if $\alpha_{(k)}>\alpha^*$. In other words, only if we have at most $k-1$ of the α_i

below or at α^*. The probability that at most $k-1 = \lfloor \epsilon(n+1) \rfloor$ values of the α_i are below or at α^* equals $\mathbb{P}(B_n \le \lfloor \epsilon(n+1) \rfloor)$, where $B_n \sim \mathrm{bin}_{n,E}$. This proves (10).

The last statement of part (b) follows immediately from what we have already proved. □

In the proof of Theorem 1 we used the first statement of the following lemma.

Lemma 1. *Fix the number of trials n. The distribution function $\mathrm{bin}_{n,p}(K)$ of the binomial distribution is decreasing in the probability of success p for a fixed $K \in \{0, \dots, n\}$. It is strictly decreasing unless $K = n$.*

Proof. For the first statement of the lemma, it suffices to check that

$$\frac{d \, \mathrm{bin}_{n,p}(K)}{dp} = \frac{d}{dp} \sum_{k=0}^{K} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{K} \frac{k - np}{p(1-p)} \binom{n}{k} p^k (1-p)^{n-k}$$

is nonpositive for $p \in (0,1)$. The last sum has the same sign as the mean of the function $f(k) := k - np$ over the set $k \in \{0, \dots, K\}$ with respect to the binomial distribution, and so it remains to notice that the overall mean of f is 0 and that the function f is increasing. This proves the first statement, and the second statement is now obvious. □

The following corollary makes (7) and (9) in Theorem 1 less precise but more explicit using Hoeffding's inequality.

Corollary 1. *Let $\epsilon, \delta, E \in (0,1)$.*

(a) If Γ is an inductive conformal predictor, the set predictor Γ^ϵ is (E, δ)-valid provided

$$E \ge \epsilon + \sqrt{\frac{-\ln \delta}{2n}}. \tag{11}$$

(b) If Γ is a smoothed inductive conformal predictor based on an inductive conformity measure continuous under Q, the set predictor Γ^ϵ is (E, δ)-valid with respect to Q provided

$$E \ge \left(1 + \frac{1}{n}\right) \epsilon + \sqrt{\frac{-\ln \delta}{2n}}. \tag{12}$$

This corollary gives the following recipe for constructing (ϵ, δ)-valid set predictors. The recipe only works if the training set is sufficiently large; in particular, its size l should significantly exceed $N := (-\ln \delta)/(2\epsilon^2)$. Choose an ICP Γ with the size n of the calibration set exceeding N. Then the set predictor $\Gamma^{\epsilon - \sqrt{(-\ln \delta)/(2n)}}$ will be (ϵ, δ)-valid.

Proof of Corollary 1. Suppose $E > \epsilon$. Combining (7) with Hoeffding's inequality (see, e.g., Vovk et al. 2005, p. 287), we can see that the probability of error $Q(\mathbf{Z} \setminus \Gamma^\epsilon(Z_1, \dots, Z_l, \Omega))$ for an ICP will exceed E with probability at most

$$\mathbb{P}(B_n \le \lfloor \epsilon(n+1) - 1 \rfloor) \le \mathbb{P}(B_n \le \epsilon n) \le e^{-2(E-\epsilon)^2 n},$$

where $B_n \sim \mathrm{bin}_{n,E}$ and ϵ is the significance level. Solving $e^{-2(E-\epsilon)^2 n} = \delta$ we obtain that Γ^ϵ is (E, δ)-valid whenever (11) is satisfied.

Analogously, in the case of a smoothed ICP and (9) we have

$$\mathbb{P}(B_n \le \lfloor \epsilon(n+1) \rfloor) \le \mathbb{P}(B_n \le (1 + 1/n)\epsilon n) \le e^{-2(E-(1+1/n)\epsilon)^2 n},$$

and solving $e^{-2(E-(1+1/n)\epsilon)^2 n} = \delta$ leads to (12). □

Remark. The training conditional guarantees discussed in this section are very similar to those for the hold-out estimate of the probability of error of a classifier: compare, e.g., Theorem 1(a) above and Theorem 3.3 in Langford (2005). The former says that Γ^ϵ is (E, δ)-valid for

$$E := \overline{\mathrm{bin}}_{n,\delta}\left(\lfloor \epsilon(n+1) - 1 \rfloor\right) \le \overline{\mathrm{bin}}_{n,\delta}\left(\epsilon n\right) \tag{13}$$

where $\overline{\mathrm{bin}}$ is the inverse function to bin:

$$\overline{\mathrm{bin}}_{n,\delta}(k) := \max\{p \mid \mathrm{bin}_{n,p}(k) \ge \delta\} \tag{14}$$

(unless $k = n$, we can also say that $\overline{\mathrm{bin}}_{n,\delta}(k)$ is the only value of p such that $\mathrm{bin}_{n,p}(k) = \delta$: cf. Lemma 1 above). And the latter says that a point predictor's error probability (over the test example) does not exceed

$$\overline{\mathrm{bin}}_{n,\delta}\left(k\right) \tag{15}$$

with probability at least $1 - \delta$ (over the training set), where k is the number of errors on a held-out set of size n. The main difference between (13) and (15) is that whereas one inequality contains the approximate expected number of errors ϵn for n new examples the other contains the actual number of errors k on n examples. Several researchers have found that the hold-out estimate is surprisingly difficult to beat; however, like the ICP of this section, it is not example conditional at all.

In conclusion of this section we give a statement intermediate between Theorem 1 and Corollary 1.

Corollary 2. *Let $\epsilon, \delta, E \in (0, 1)$.*

(a) If Γ is an inductive conformal predictor, the set predictor Γ^ϵ is (E, δ)-valid provided

$$E \ge \epsilon + \sqrt{\frac{-2\epsilon \ln \delta}{n}} - \frac{2 \ln \delta}{n}.$$

(b) *If* Γ *is a smoothed inductive conformal predictor based on an inductive conformity measure continuous under* Q*, the set predictor* Γ^ϵ *is* (E,δ)*-valid with respect to* Q *provided*

$$E \geq (1+1/n)\epsilon + \sqrt{\frac{-2(1+1/n)\epsilon \ln \delta}{n}} - \frac{2\ln\delta}{n}.$$

Proof. Inequality (7) can be rewritten as

$$E \geq \overline{\mathrm{bin}}_{n,\delta}\left(\lfloor \epsilon(n+1) - 1 \rfloor\right)$$

(using the notation (14)). In combination with inequality 2. in Langford (2005), p. 278, this leads to the first statement. The second statement follows by replacing ϵ with $(1+1/n)\epsilon$. □

4 Conditional inductive conformal predictors

The motivation behind conditional inductive conformal predictors is that ICPs do not always achieve the required probability ϵ of error $Y_{l+1} \notin \Gamma^\epsilon(Z_1,\dots,Z_l,X_{l+1},\Omega)$ conditional on $(X_{l+1},Y_{l+1}) \in E$ for important sets $E \subseteq \mathbf{Z}$. This is often undesirable. If, e.g., our set predictor is valid at the significance level 5% but makes an error with probability 10% for men and 0% for women, both men and women can be unhappy with calling 5% the probability of error. Moreover, in many problems we might want different significance levels for different regions of the example space: e.g., in the problem of spam detection (considered in Sections 6 and 7) classifying spam as email usually makes much less harm than classifying email as spam.

An *inductive* m*-taxonomy* is a measurable function $K : \mathbf{Z}^m \times \mathbf{Z} \to \mathbf{K}$, where $\mathbf{K}$ is a measurable space. Usually the *category* $K((z_1,\dots,z_m),z)$ of an example z is a kind of classification of z, which may depend on the proper training set $(z_1,\dots,z_m)$.

The *conditional inductive conformal predictor* (conditional ICP) corresponding to K and an inductive conformity measure A is defined as the set predictor (2), where the p-values p^y are now defined by

$$p^y := \frac{|\{i = m+1,\dots,l \mid \kappa_i = \kappa^y \mathbin{\&} \alpha_i \leq \alpha^y\}| + 1}{|\{i = m+1,\dots,l \mid \kappa_i = \kappa^y\}| + 1}, \tag{16}$$

the categories κ are defined by

$$\kappa_i := K((z_1,\dots,z_m),z_i), \quad i = m+1,\dots,l, \qquad \kappa^y := K((z_1,\dots,z_m),(x,y)),$$

and the conformity scores α are defined as before by (4). A *label conditional ICP* is a conditional ICP with the inductive m-taxonomy $K(\cdot,(x,y)) := y$; this notion is useful only in classification problems.

The following proposition is the conditional analogue of the first part of Proposition 1; in particular, it shows that in classification problems label conditional ICPs achieve label conditional validity.

Proposition 2. *If random examples $Z_{m+1}, \dots, Z_l, Z_{l+1} = (X_{l+1}, Y_{l+1})$ are exchangeable, the probability of error $Y_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l, X_{l+1}, \Omega)$ given the category $K((Z_1, \dots, Z_m), Z_{l+1})$ of Z_{l+1} does not exceed ϵ for any ϵ and any conditional inductive conformal predictor Γ corresponding to K.*

We refrain from giving the definition of smoothed conditional ICPs, which is straightforward. The categories can also be made dependent on $\omega \in \bar{\Omega}$.

5 Object conditional validity

In this section we prove a negative result (a version of Lemma 1 in Lei and Wasserman 2013) which says that the requirement of precise object conditional validity cannot be satisfied in a non-trivial way for rich object spaces (such as $\mathbb{R}$). If Q is a probability distribution on $\mathbf{Z}$, we let $Q_{\mathbf{X}}$ stand for its marginal distribution on $\mathbf{X}$: $Q_{\mathbf{X}}(A) := Q(A \times \mathbf{Y})$. In this section we consider only set predictors that do not depend on θ, but the case of set predictors depending on θ (such as smoothed ICPs) is also covered by redefining $\omega := (\omega, \theta)$.

Let us say that a set predictor Γ *has* $1 - \epsilon$ *object conditional validity*, where $\epsilon \in (0, 1)$, if, for all probability distributions Q on $\mathbf{Z}$ and $Q_{\mathbf{X}}$-almost all $x \in \mathbf{X}$,

$$(Q^{l+1} \times R)\left(Y_{l+1} \in \Gamma(Z_1, \dots, Z_l, X_{l+1}, \Omega) \mid X_{l+1} = x\right) \geq 1 - \epsilon.$$

If P is a probability distribution on $\mathbf{X}$, we say that a property F of elements of $\mathbf{X}$ holds for *P-almost all* elements of a measurable set $E \subseteq \mathbf{X}$ if $P(E \setminus F) = 0$; a *P-non-atom* is an element $x \in \mathbf{X}$ such that $P(\{x\}) = 0$. The Lebesgue measure on $\mathbb{R}$ will be denoted Λ, and the convex hull of $E \subseteq \mathbb{R}$ will be denoted $\operatorname{co} E$.

Theorem 2. *Suppose $\mathbf{X}$ is a separable metric space equipped with the Borel σ-algebra. Let $\epsilon \in (0, 1)$. Suppose that a set predictor Γ has $1 - \epsilon$ object conditional validity. In the case of regression, we have, for all probability distributions Q on $\mathbf{Z}$ and for $Q_{\mathbf{X}}$-almost all $Q_{\mathbf{X}}$-non-atoms $x \in \mathbf{X}$,*

$$(Q^{l+1} \times R)\left(\Lambda(\Gamma(Z_1, \dots, Z_l, x, \Omega)) = \infty\right) \geq 1 - \epsilon \tag{17}$$

and

$$(Q^{l+1} \times R)\left(\operatorname{co}\Gamma(Z_1, \dots, Z_l, x, \Omega) = \mathbb{R}\right) \geq 1 - 2\epsilon. \tag{18}$$

In the case of classification, we have, for all Q, all $y \in \mathbf{Y}$, and $Q_{\mathbf{X}}$-almost all $Q_{\mathbf{X}}$-non-atoms x,

$$(Q^{l+1} \times R)\left(y \in \Gamma(Z_1, \dots, Z_l, x, \Omega)\right) \geq 1 - \epsilon. \tag{19}$$

The constant ϵ in each of (17), (18), and (19) is optimal, in the sense that it cannot be replaced by a smaller constant.

We are mainly interested in the case of a small ϵ (corresponding to high confidence), and in this case (17) implies that, in the case of regression, the prediction interval (i.e., the convex hull of the prediction set) can be expected

to be infinitely long unless the test object is an atom. Even an infinitely long prediction interval can be somewhat informative providing a one-sided bound on the label of the test example; (18) says that, with probability at least $1-2\epsilon$, the prediction interval is completely uninformative unless the test object is an atom. In the case of classification, (19) says that each particular $y \in \mathbf{Y}$ is likely to be included in the prediction set, and so the prediction set is likely to be large. In particular, (19) implies that the expected size of the prediction set is a least $(1-\epsilon)|\mathbf{Y}|$.

Of course, the condition that the test object x be a non-atom is essential: if $Q_{\mathbf{X}}(\{x\}) > 0$, an inductive conformal predictor that ignores all examples with objects different from the current test object can have $1-\epsilon$ object conditional validity and still produce a small prediction set for a test object x if the training set is big enough to contain many examples with x as their object.

Remark. Nontrivial set predictors having $1-\epsilon$ object conditional validity are constructed by McCullagh et al. (2009) assuming the Gauss linear model.

Proof of Theorem 2. The proof will be based on the ideas of Lei and Wasserman (2013, the proof of Lemma 1).

We start from showing that the ϵ in (17), (18), and (19) cannot be replaced by a smaller constant. For (17) and (19) this follows from the fact that the trivial set predictor predicting $\mathbf{Y}$ with probability $1-\epsilon$ and $\emptyset$ with probability ϵ has $1-\epsilon$ object conditional validity. In the case of (18) the bound $1-2\epsilon$ is attained by the set predictor predicting $\mathbb{R}$ with probability $1-2\epsilon$, $[0,\infty)$ with probability ϵ, and $(-\infty,0]$ with probability ϵ (this assumes $\epsilon < 1/2$; the case $\epsilon \geq 1/2$ is trivial). This predictor's conditional probability of error given all $l+1$ examples is at most ϵ (0 if $y_{l+1} = 0$ and ϵ otherwise); therefore, the conditional probability of error will be at most ϵ given the test object.

Next we prove the first statement about regression. Suppose (17) does not hold on a measurable set E of $Q_{\mathbf{X}}$-non-atoms $x \in \mathbf{X}$ such that $Q_{\mathbf{X}}(E) > 0$. Shrink E in such a way that $Q_{\mathbf{X}}(E) > 0$ still holds but there exist $\delta > 0$ and $C > 0$ such that, for each $x \in E$,

$$(Q^{l+1} \times R)\left(\Lambda(\Gamma(Z_1,\dots,Z_l,x,\Omega)) \leq C\right) \geq \epsilon + \delta. \tag{20}$$

Let V be the total variation distance between probability measures, $V(P,Q) := \sup_A |P(A) - Q(A)|$; we then have

$$V(P^l, Q^l) \leq \sqrt{2}\sqrt{1-(1-V(P,Q))^l}$$

(this follows from the connection of V with the Hellinger distance: see, e.g., Tsybakov 2010, Section 2.4). Shrink E further so that $Q_{\mathbf{X}}(E) > 0$ still holds but

$$\sqrt{2}\sqrt{1-(1-Q_{\mathbf{X}}(E))^l} \leq \delta/2. \tag{21}$$

(This can be done under our assumption that $\mathbf{X}$ is a separable metric space: see Lemma 2 below.) Define another probability distribution P on $\mathbf{Z}$ by the

requirements that $P(A \times B) = Q(A \times B)$ for all measurable $A \subseteq (\mathbf{X} \setminus E)$, $B \subseteq \mathbb{R}$ and that $P(A \times B) = Q_{\mathbf{X}}(A) \times U(B)$ for all measurable $A \subseteq E$, $B \subseteq \mathbb{R}$, where U is the uniform probability distribution on the interval $[-DC, DC]$ and $D > 0$ will be chosen below. Since $V(P, Q) \le Q_{\mathbf{X}}(E)$, we have $V(P^l, Q^l) \le \delta/2$, which implies $V(P^l \times R, Q^l \times R) \le \delta/2$; therefore, by (20),

$$(P^{l+1} \times R)\,(\Lambda(\Gamma(Z_1, \dots, Z_l, x, \Omega)) \le C) \ge \epsilon + \delta/2$$

for each $x \in E$. The last inequality implies, by Fubini's theorem,

$$(P^{l+1} \times R)\,(\Lambda(\Gamma(Z_1, \dots, Z_l, X_{l+1}, \Omega)) \le C \;\&\; X_{l+1} \in E) \ge (\epsilon + \delta/2)\, P_{\mathbf{X}}(E),$$

where $P_{\mathbf{X}}(E) = Q_{\mathbf{X}}(E) > 0$ is the marginal P-probability of E. When D (depending on $\delta P_{\mathbf{X}}(E)$) is sufficiently large this in turn implies

$$(P^{l+1} \times R)\,(Y_{l+1} \notin \Gamma(Z_1, \dots, Z_l, X_{l+1}, \Omega) \;\&\; X_{l+1} \in E) \ge (\epsilon + \delta/4)\, P_{\mathbf{X}}(E).$$

However, the last inequality contradicts

$$\frac{(P^{l+1} \times R)\,(Y_{l+1} \notin \Gamma(Z_1, \dots, Z_l, X_{l+1}, \Omega) \;\&\; X_{l+1} \in E)}{P_{\mathbf{X}}(E)} \le \epsilon, \tag{22}$$

which follows from Γ having $1 - \epsilon$ object conditional validity and the definition of conditional probability.

For the second statement about regression, suppose (18) does not hold on a measurable set E of $Q_{\mathbf{X}}$-non-atoms $x \in \mathbf{X}$ such that $Q_{\mathbf{X}}(E) > 0$. In other words, for all $x \in E$,

$$(Q^{l+1} \times R)\,(\sup \Gamma(Z_1, \dots, Z_l, x, \Omega) < \infty \text{ or } \inf \Gamma(Z_1, \dots, Z_l, x, \Omega) > -\infty) > 2\epsilon.$$

For each $x \in E$ we have either

$$(Q^{l+1} \times R)\,(\sup \Gamma(Z_1, \dots, Z_l, x, \Omega) < \infty) > \epsilon \tag{23}$$

or

$$(Q^{l+1} \times R)\,(\inf \Gamma(Z_1, \dots, Z_l, x, \Omega) > -\infty) > \epsilon. \tag{24}$$

Therefore, either (23) or (24) holds on a subset of E of a positive $Q_{\mathbf{X}}$-probability. Suppose, for concreteness, that (23) does. Shrink E in such a way that $Q_{\mathbf{X}}(E) > 0$ still holds and (23) holds for all $x \in E$. Shrink E further in such a way that $Q_{\mathbf{X}}(E) > 0$ still holds but there exist $\delta > 0$ and $C > 0$ such that, for each $x \in E$,

$$(Q^{l+1} \times R)\,(\sup \Gamma(Z_1, \dots, Z_l, x, \Omega) \le C) \ge \epsilon + \delta. \tag{25}$$

Shrink E further so that both $Q_{\mathbf{X}}(E) > 0$ and (21) hold. Define a probability distribution P on $\mathbf{Z}$ by the requirements that $P(A \times B) = Q(A \times B)$ for all measurable $A \subseteq (\mathbf{X} \setminus E)$ and $B \subseteq \mathbb{R}$ and that $P(A \times \{C + 1\}) = Q_{\mathbf{X}}(A)$ for all measurable $A \subseteq E$ (i.e., modify Q setting the conditional distribution of Y given

$X \in E$ to the unit mass concentrated at $C+1$). Since $V(P^l \times R, Q^l \times R) \le \delta/2$, (25) implies

$$(P^{l+1} \times R)\,(\sup \Gamma(Z_1, \dots, Z_l, x, \Omega) \le C) \ge \epsilon + \delta/2$$

for all $x \in E$, which in turn implies

$$(P^{l+1} \times R)\,(\sup \Gamma(Z_1, \dots, Z_l, X_{l+1}, \Omega) \le C \;\&\; X_{l+1} \in E) \ge (\epsilon + \delta/2) P_{\mathbf{X}}(E),$$

which in turn implies

$$(P^{l+1} \times R)\,(Y_{l+1} \notin \Gamma(Z_1, \dots, Z_l, X_{l+1}, \Omega) \;\&\; X_{l+1} \in E) \ge (\epsilon + \delta/2) P_{\mathbf{X}}(E),$$

which contradicts (22).

It remains to prove the statement about classification. Suppose (19) does not hold on a measurable set E of $Q_{\mathbf{X}}$-non-atoms $x \in \mathbf{X}$ such that $Q_{\mathbf{X}}(E) > 0$. Shrink E in such a way that $Q_{\mathbf{X}}(E) > 0$ still holds but there exists $\delta > 0$ such that, for each $x \in E$,

$$(Q^{l+1} \times R)\,(y \in \Gamma(Z_1, \dots, Z_l, x, \Omega)) \le 1 - \epsilon - \delta.$$

Without loss of generality we further assume that (21) also holds. Define a probability distribution P on $\mathbf{Z}$ by the requirements that $P(A \times B) = Q(A \times B)$ for all measurable $A \subseteq (\mathbf{X} \setminus E)$ and all $B \subseteq \mathbf{Y}$ and that $P(A \times \{y\}) = Q_{\mathbf{X}}(A)$ for all measurable $A \subseteq E$ (i.e., modify Q setting the conditional distribution of Y given $X \in E$ to the unit mass concentrated at y). Then for each $x \in E$ we have

$$(P^{l+1} \times R)\,(y \in \Gamma(Z_1, \dots, Z_l, x, \Omega)) \le 1 - \epsilon - \delta/2,$$

which implies

$$(P^{l+1} \times R)\,(Y_{l+1} \in \Gamma(Z_1, \dots, Z_l, X_{l+1}, \Omega) \;\&\; X_{l+1} \in E) \le (1 - \epsilon - \delta/2)\, P_{\mathbf{X}}(E).$$

The last inequality contradicts Γ having $1 - \epsilon$ object conditional validity. □

In the proof of Theorem 2 we used the following lemma.

Lemma 2. *If Q is a probability measure on $\mathbf{X}$, which is assumed to be a separable metric space, E is a set of Q-non-atoms such that $Q(E) > 0$, and $\delta > 0$ is an arbitrarily small number, then there is $E' \subseteq E$ such that $0 < Q(E') < \delta$.*

Proof. We can take as E' the intersection of E and an open ball centred at any element of $\mathbf{X}$ for which all such intersections have a positive Q-probability. Let us prove that such elements exist. Suppose they do not.

Fix a countable dense subset A_1 of $\mathbf{X}$. Let A_2 be the union of all open balls B with rational radii centred at points in A_1 such that $Q(B \cap E) = 0$. On one hand, the σ-additivity of measures implies $Q(A_2 \cap E) = 0$. On the other hand, $A_2 = \mathbf{X}$: indeed, for each $x \in \mathbf{X}$ there is an open ball B of some radius $\epsilon > 0$ centred at x that satisfies $Q(B \cap E) = 0$; since x belongs to the radius $\epsilon/2$ open ball centred at a point in A_1 at a distance of less than $\epsilon/2$ from x, we have $x \in A_2$. This contradicts $Q(E) > 0$. □

Theorem 2 demonstrates an interesting all-or-nothing phenomenon for set predictors having $1-\epsilon$ object conditional validity: each such predictor produces hopelessly large prediction sets with probability at least $1-\epsilon$; on the other hand, already a trivial predictor of this kind (mentioned in the proof) produces the smallest possible prediction sets with probability ϵ.

The theorem does not prevent the existence of efficient set predictors that are object conditionally valid in an asymptotic sense; indeed, the paper by Lei and Wasserman (2013) is devoted to constructing asymptotically efficient and asymptotically object conditionally valid set predictors in the case of regression.

6 Experiments

This section describes some simple experiments on the well-known `Spambase` data set contributed by George Forman to the UCI Machine Learning Repository (Frank and Asuncion, 2010). Its overall size is 4601 examples and it contains examples of two classes: `email` (also written as 0) and `spam` (also written as 1). Hastie et al. (2009) report results of several machine-learning algorithms on this data set split randomly into a training set of size 3065 and test set of size 1536. The best result is achieved by MART (multiple additive regression tree; 4.5% error rate according to the second edition of Hastie et al. 2009).

All our experiments are for (unsmoothed) ICPs. We randomly permute the data set and divide it into 2602 examples for the proper training set, 999 for the calibration set, and 1000 for the test set. Our split between the proper training, calibration, and test sets, approximately 2:1:1, is inspired by the standard recommendation for the allocation of data into training, validation, and test sets (see, e.g., Hastie et al. 2009, Section 7.2). We consider the ICP whose conformity measure is defined by (1) where f is output by MART and

$$\Delta(y, f(x)) := \begin{cases} f(x) & \text{if } y = 1 \\ -f(x) & \text{if } y = 0. \end{cases} \tag{26}$$

MART's output $f(x)$ models the log-odds of `spam` vs `email`,

$$f(x) = \log \frac{P(1 \mid x)}{P(0 \mid x)},$$

which makes the interpretation of (26) as conformity score very natural.

The R programs used in the experiments described in this and next sections for producing the tables and figures in the conference version of this paper (Vovk, 2012) are available from the web site `http://alrw.net`; the programs use the `gbm` package with virtually all parameters set to the default values (given in the description provided in response to `help("gbm")`).

The upper left plot in Figure 2 is the scatter plot of the pairs $(p^{\text{email}}, p^{\text{spam}})$ produced by the ICP for all examples in the test set. Email is shown as (blue) noughts and spam as (red) crosses (and when the figure is viewed in colour, it is noticeable that the noughts were drawn after the crosses). The other two plots

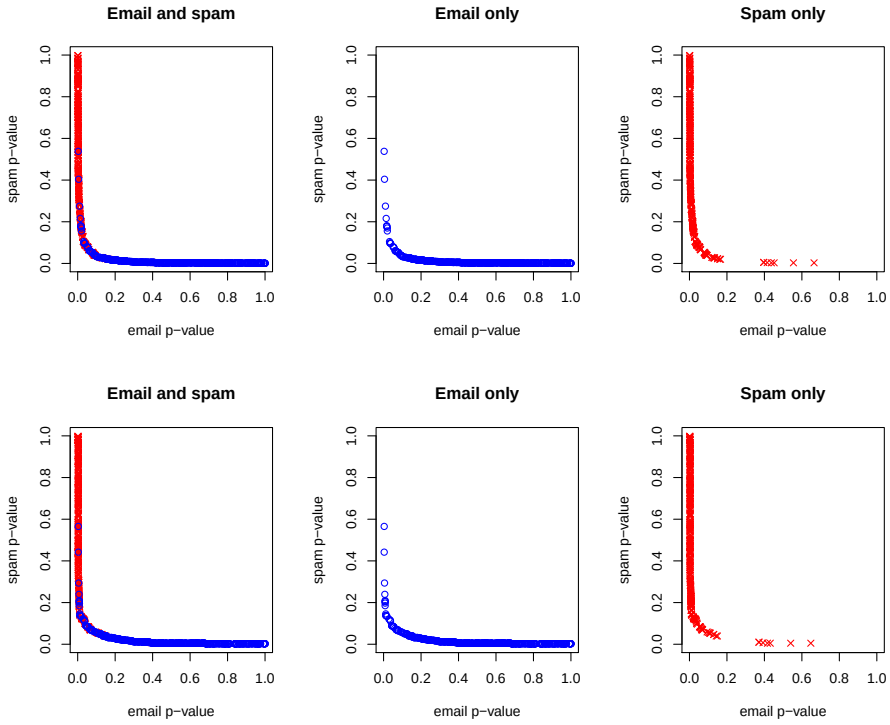

Figure 2: Scatter plots of the pairs $(p^{\text{email}}, p^{\text{spam}})$ for all examples in the test set (left plots), for email only (middle), and for spam only (right). Email is shown as (blue) noughts and spam as (red) crosses. The three upper plots are for the ICP and the three lower ones are for the label conditional ICP.

in the upper row are for email and spam separately. Ideally, email should be close to the horizontal axis and spam to the vertical axis; we can see that this is often true, with a few exceptions. The picture for the label conditional ICP looks almost identical: see the lower row of Figure 2. However, on the log scale the difference becomes more noticeable: see Figure 3.

Table 1 gives some statistics for the numbers of errors, multiple set predictions $\{0,1\}$, and empty set predictions $\emptyset$ in the case of the (unconditional) ICP $\Gamma^{5\%}$ at significance level 5% (we obtain different numbers not only because of different splits but also because MART is randomized; the columns of the table correspond to the random number generator seeds 0, 1, 2, etc.). The table demonstrates the validity, (lack of) conditional validity, and efficiency of the algorithm (the latter is of course inherited from the efficiency of MART). We give two kinds of conditional figures: the percentages of errors, multiple, and empty predictions for different labels and for two different kinds of objects. The

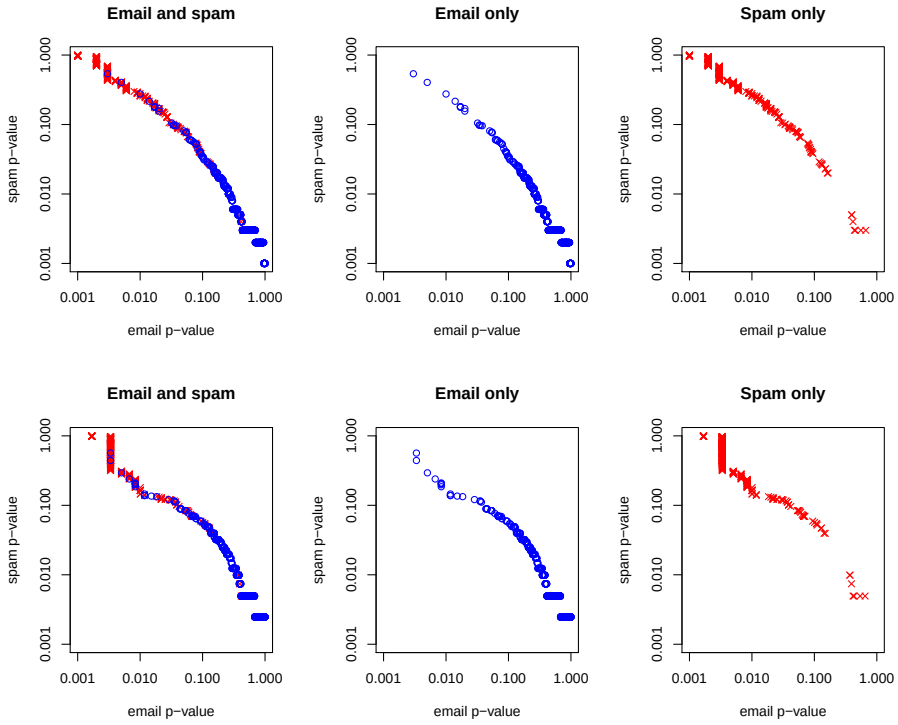

Figure 3: The analogue of Figure 2 on the log scale.

two kinds of objects are obtained by splitting the object space $\mathbf{X}$ by the value of an attribute that we denote \$: it shows the percentage of the character \$ in the text of the message. The condition \$ $< 5.55\%$ was the root of the decision tree chosen both by Hastie et al. (2009, Section 9.2.5), who use all attributes in their analysis, and by Maindonald and Braun (2007, Chapter 11), who use 6 attributes chosen by them manually. (Both books use the `rpart` R package for decision trees.)

Notice that the numbers of errors, multiple predictions, and empty predictions tend to be greater for spam than for email. Somewhat counter-intuitively, they also tend to be greater for "email-like" objects containing few \$ characters than for "spam-like" objects. The percentage of multiple and empty predictions is relatively small since the error rate of the underlying predictor happens to be close to our significance level of 5%.

In practice, using a fixed significance level (such as the standard 5%) is not a good idea; we should at least pay attention to what happens at several significance levels. However, experimenting with prediction sets at a fixed significance level facilitates a comparison with theoretical results.

Table 1: Percentages of errors, multiple predictions, and empty predictions at significance level 5% on the full test set and separately on email and spam and on two kinds of objects. The results are given for the first 100 seeds for the R (pseudo)random number generator (RNG); column "Average" gives the average percentages for all 100 seeds 0–99, and column "St. dev." gives usual estimates of the standard deviations (namely, the square roots of the standard unbiased estimates of the variances) of the percentages for the 100 seeds.

RNG seed	0	1	2	...	99	Average	St. dev.
errors overall	4.1%	6.9%	4.6%	...	4.2%	5.08%	1.00%
for email	2.44%	4.61%	2.26%	...	2.82%	3.35%	0.92%
for spam	6.77%	10.43%	8.42%	...	6.30%	7.74%	1.64%
for \$ < 5.55%	4.36%	7.91%	5.15%	...	4.34%	5.76%	1.24%
for \$ > 5.55%	3.29%	4.12%	2.69%	...	3.75%	2.96%	1.02%
multiple overall	2.7%	0%	0.1%	...	1.2%	0.86%	0.98%
for email	2.11%	0%	0.16%	...	0.83%	0.60%	0.68%
for spam	3.65%	0%	0%	...	1.76%	1.26%	1.52%
for \$ < 5.55%	3.04%	0%	0.13%	...	1.18%	0.98%	1.15%
for \$ > 5.55%	1.65%	0%	0%	...	1.25%	0.49%	0.68%
empty overall	0%	2.7%	0%	...	0%	0.31%	0.63%
for email	0%	1.48%	0%	...	0%	0.24%	0.47%
for spam	0%	4.58%	0%	...	0%	0.42%	0.96%
for \$ < 5.55%	0%	3.14%	0%	...	0%	0.36%	0.73%
for \$ > 5.55%	0%	1.50%	0%	...	0%	0.14%	0.40%

Table 2 gives similar statistics in the case of the label conditional ICP. The error rates are now about equal for email and spam, as expected. We refrain from giving similar predictable results for "object conditional" ICP with \$ < 5.55% and \$ > 5.55% as categories.

We define the calibration plot of an ICP Γ on a test set as the percentage of errors made by Γ^ϵ plotted against $\epsilon \in (0, 1)$. Figure 4 gives three calibration plots for the ICP: for the full test set and for email and spam separately. It shows approximate validity even for email and spam separately, except for the all-important lower-left corners. The latter are shown separately in Figure 5, where the lack of conditional validity becomes evident; cf. Figure 6 for the label conditional ICP.

From the numbers in the "errors overall" row of Table 1 (both given and hidden in the ... part) we can extract the corresponding confidence intervals for the probability of error conditional on the training set and MART's internal coin tosses; these are shown in Figure 7. It can be seen that training conditional validity is not grossly violated. (Notice that the 100 training sets used for producing this figure are not completely independent. Besides, the assumption of randomness might not be completely satisfied: permuting the data set ensures exchangeability but not necessarily randomness.) It is instructive to compare

Table 2: The analogue of a subset of Table 1 in the case of the label conditional ICP.

RNG seed	0	1	2	...	99	Average	St. dev.
errors overall	3.4%	6.0%	3.8%	...	3.6%	4.92%	0.91%
for email	3.73%	6.92%	3.87%	...	3.48%	4.97%	1.15%
for spam	2.86%	4.58%	3.68%	...	3.78%	4.82%	1.33%
multiple overall	4.2%	0%	4.0%	...	2.6%	1.68%	1.54%
for email	3.90%	0%	5.48%	...	2.49%	1.94%	1.86%
for spam	4.69%	0%	1.58%	...	2.77%	1.28%	1.26%
empty overall	0%	1.0%	0%	...	0%	0.15%	0.45%
for email	0%	1.48%	0%	...	0%	0.15%	0.47%
for spam	0%	0.25%	0%	...	0%	0.15%	0.47%

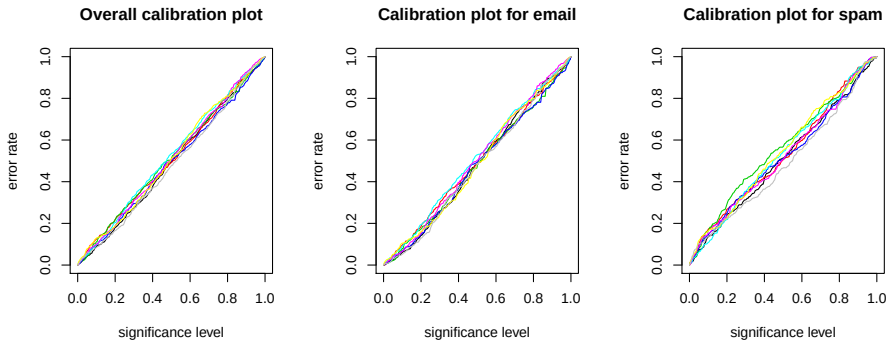

Figure 4: The calibration plot for the test set overall, the email in the test set, and the spam in the test set (for the first 8 seeds, 0–7).

Figure 7 with the "theoretical" Figure 8 obtained from Theorem 1(a) (the thick black line), Corollary 1(a) (the thin solid line, which may be shown in red), and Corollary 2(a) (the thin dashed line, which may be shown in blue). The dotted black line corresponds to the significance level 5%. There is no obvious discrepancy between Figures 7 and 8.

Figure 8 gives bounds on the training conditional error probability as a function of δ for a fixed size $n = 999$ of the calibration set. Figure 9, on the other hand, gives bounds on the training conditional error probability as a function of the size n of the calibration set for a fixed δ, namely for $\delta = 1\%$.

Figure 10 is the analogue of Figure 8 for significance level $\epsilon = 1\%$. Notice that the thin solid line (corresponding to Corollary 1(a) and perhaps shown in red) simply shifts down by 4%. However, the quality of the thick black line (corresponding to Theorem 1(a)) and the thin dashed line (corresponding to

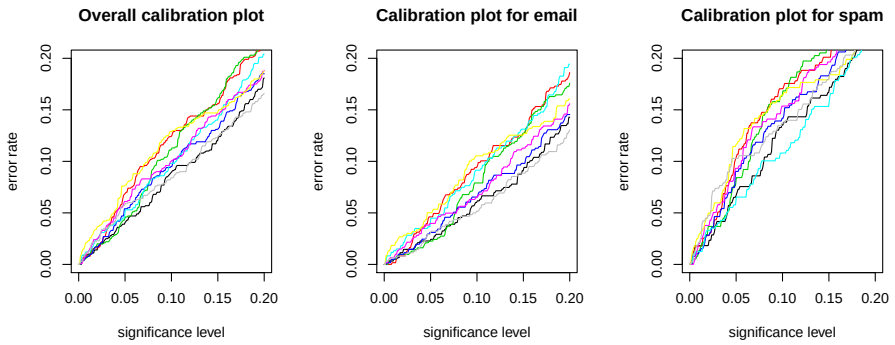

Figure 5: The lower left corners of the plots in Figure 4.

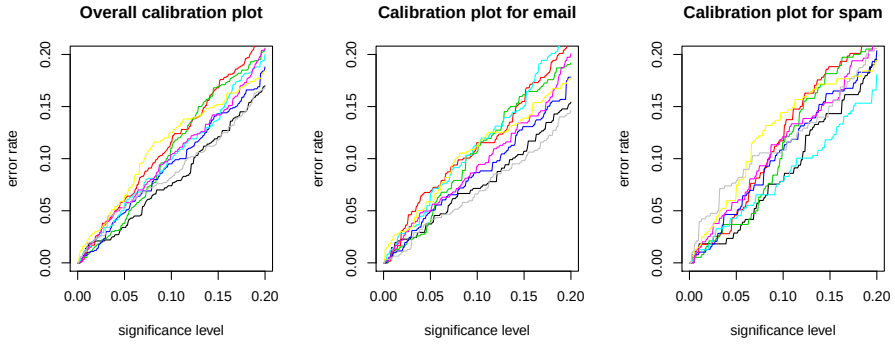

Figure 6: The analogue of Figure 5 for the label conditional ICP.

Corollary 2(a) and perhaps shown in blue) becomes significantly better that that.

7 ICPs and ROC curves

This section discusses a close connection between an important class of ICPs ("scoring-type" label conditional ICPs) and ROC curves. (For a previous study of connections between conformal prediction and ROC curves, see Vanderlooy and Sprinkhuizen-Kuyper 2007.) Let us say that an ICP or a label conditional ICP is *scoring-type* if its inductive conformity measure is defined by (1) where f takes values in $\mathbb{R}$ and Δ is defined by (26).

The reader might have noticed that the two leftmost plots in Figure 2 look similar to a ROC curve. The following proposition will show that this is not

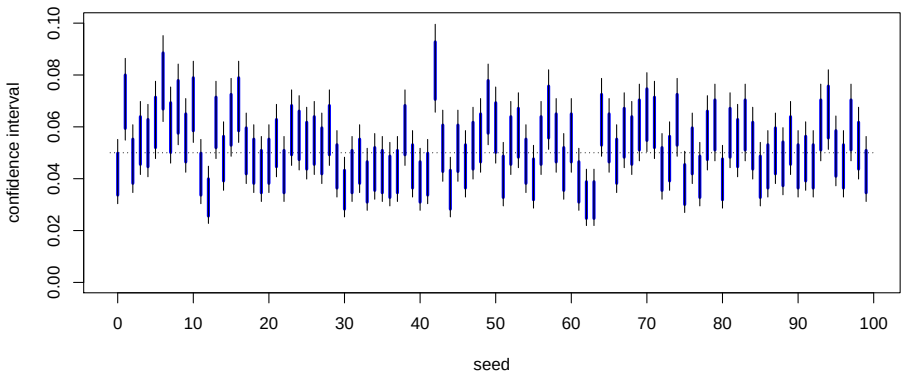

Figure 7: Confidence intervals for training conditional error probabilities: 95% shown as thin lines (in black) and 80% shown as thick lines (perhaps in blue). The 5% significance level is shown as the horizontal dotted black line.

coincidental in the case of the lower left one. However, before we state it, we need a few definitions. We will now consider a general binary classification problem and will denote the labels as 0 and 1. For a threshold $c \in \mathbb{R}$, the *type I error on the calibration set* is

$$\alpha(c) := \frac{|\{i = m+1, \dots, l \mid f(x_i) \geq c \,\&\, y_i = 0\}|}{|\{i = m+1, \dots, l \mid y_i = 0\}|} \tag{27}$$

and the *type II error on the calibration set* is

$$\beta(c) := \frac{|\{i = m+1, \dots, l \mid f(x_i) \leq c \,\&\, y_i = 1\}|}{|\{i = m+1, \dots, l \mid y_i = 1\}|} \tag{28}$$

(with 0/0 set, e.g., to 1/2). Intuitively, these are the error rates for the classifier that predicts 1 when $f(x) > c$ and predicts 0 when $f(x) < c$ (our definition is conservative in that it counts the prediction as error whenever $f(x) = c$); namely, $\alpha(c)$ is the false positive rate and $\beta(c)$ is the false negative rate. The *empirical ROC curve* is the parametric curve

$$\{(\alpha(c), \beta(c)) \mid c \in \mathbb{R}\} \subseteq [0,1]^2. \tag{29}$$

(Our version of ROC curves is the original version reflected in the line $y = 1/2$; in deviating from the original version we follow Hastie et al. 2009, whose version is the original one reflected in the line $x = 1/2$, and many other books and papers; see, e.g., Bengio et al. 2005, Figure 1.) Since $\alpha(c)$ and $\beta(c)$ take only finitely many values, the empirical ROC curve (along with its modifications introduced below) is not continuous but consists of discrete points.

Proposition 3. *In the case of a scoring-type label conditional ICP, for any object $x \in \mathbf{X}$, the distance between the pair (p^0, p^1) (see (16)) and the empirical*

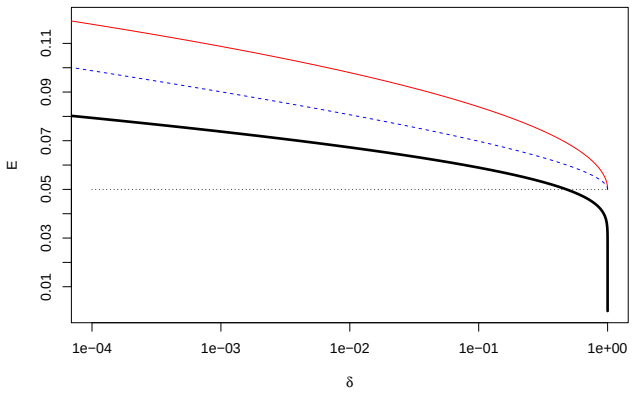

Figure 8: The upper bounds on the training conditional probability of error vs δ given by Theorem 1(a) (the thick black line), Corollary 1(a) (the thin solid line, perhaps shown in red), and Corollary 2(a) (the thin dashed line, perhaps shown in blue), where $\epsilon = 5\%$ and $n = 999$.

ROC curve is at most

$$\sqrt{\frac{1}{(n^0+1)^2} + \frac{1}{(n^1+1)^2}}, \tag{30}$$

where n^y is the number of examples in the calibration set labelled as y.

Proof. Let $c := f(x)$. Then we have

$$(p^0, p^1) = \left(\frac{n^0_{\geq} + 1}{n^0 + 1}, \frac{n^1_{\leq} + 1}{n^1 + 1} \right) \tag{31}$$

where $n^0_{\geq}$ is the number of examples (x_i, y_i) in the calibration set such that $y_i = 0$ and $f(x_i) \geq c$ and $n^1_{\leq}$ is the number of examples in the calibration set such that $y_i = 1$ and $f(x_i) \leq c$. It remains to notice that the point $\left(n^0_{\geq}/n^0, n^1_{\leq}/n^1\right)$ belongs to the empirical ROC curve: the horizontal (resp. vertical) distance between this point and (31) does not exceed $1/(n^0+1)$ (resp. $1/(n^1+1)$), and the overall Euclidean distance does not exceed (30). □

So far we have discussed the empirical ROC curve: (27) and (28) are the empirical probabilities of errors of the two types on the calibration set. It corresponds to the estimate k/n of the parameter of the binomial distribution based on observing k successes out of n. The minimax estimate is $(k+1/2)/(n+1)$, and the corresponding ROC curve (29) where $\alpha(c)$ and $\beta(c)$ are defined by (27) and (28) with the numerators increased by $\frac{1}{2}$ and the denominators increased by 1 will be called the *minimax ROC curve*. Notice that for the minimax ROC curve we can put a coefficient of $\frac{1}{2}$ in front of (30). Similarly,

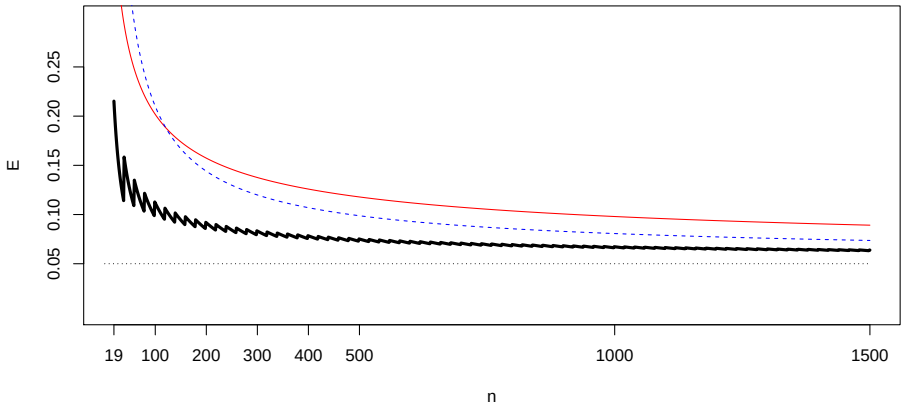

Figure 9: The upper bounds on the training conditional probability of error vs δ in the same format as in Figure 8, except that now δ is fixed at 1% and n ranges between 19 (the smallest value giving non-trivial prediction sets) and 1500; as before, $\epsilon = 5\%$.

when using the Laplace estimate $(k+1)/(n+2)$, we obtain the *Laplace ROC curve*. See the left panel of Figure 11 for the lower left corner of the lower left plot of Figure 2 with different ROC curves added to it.

The non-standard estimate $(k+1)/(n+1)$ of the parameter of the binomial distribution leads to a version of ROC curve that is connected to the label conditional ICP in the most direct way. Let us call this estimate the *upper Venn estimate* and the corresponding ROC curve the *upper Venn ROC curve* (cf. the discussion of the Venn predictor in Vovk et al. 2005, pp. 159–160). (The upper Venn estimate is unusual in that the estimate of the probability of an event plus the estimate of the probability of its complement is different from 1.) Notice that the upper Venn ROC curve lies Northeast of all three ROC curves discussed earlier. In the square $[0, 0.5] \times [0, 0.5]$ the order of the ROC curves from Southwest to Northeast is: empirical, minimax, Laplace, and upper Venn; the last two are very close to each other for large n^0 and n^1 and small ratios $n^0_{\geq}/n^0$ and $n^1_{\leq}/n^1$, as in Figure 11.

The rest of this section is devoted to a discussion of the upper Venn ROC curve. Remember that it is defined as the parametric curve (29), where now

$$\alpha(c) := \frac{|\{i = m+1, \dots, l \mid f(x_i) \geq c \,\&\, y_i = 0\}| + 1}{|\{i = m+1, \dots, l \mid y_i = 0\}| + 1}$$
$$\beta(c) := \frac{|\{i = m+1, \dots, l \mid f(x_i) \leq c \,\&\, y_i = 1\}| + 1}{|\{i = m+1, \dots, l \mid y_i = 1\}| + 1}.$$

The pair (p^0, p^1) of p-values for any test example belongs to the upper Venn ROC curve; therefore, this curve passes through all test examples in Figure 11.

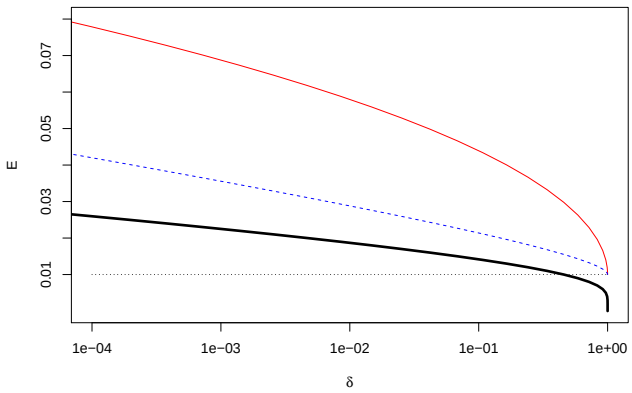

Figure 10: The analogue of Figure 8 for $\epsilon = 1\%$.

The curve can serve as a convenient classification of all possible test objects: each of them corresponds to a point on the curve.

The label conditional ICP can also be conveniently described in terms of the upper Venn ROC curve. An example is given as the right panel of Figure 11. Each test object is represented by a point (p^0, p^1). Let ϵ be the significance level; it is 5% in Figure 11 (but as mentioned earlier, there is no need to have the same significance level for email and spam). If the point (ϵ, ϵ) lies Southwest of the curve, the label conditional ICP can produce multiple predictions but never produces empty predictions. If it lies Northeast of the curve, the predictor can produce empty predictions but never produces multiple predictions. In particular, it is impossible to produce both multiple and empty predictions for the same calibration set, which is demonstrated by columns 0–99 of Table 2. (Lying on the curve is regarded as a special case of lying Northeast of it. Because of the discreteness of the upper Venn ROC curve it is also possible that (ϵ, ϵ) lies neither Northeast nor Southwest of it; in this case predictions are always singletons.)

If the test object is in the Northeast region NE with respect to (ϵ, ϵ) (i.e., $p^0 > \epsilon$ and $p^1 > \epsilon$), the prediction set is multiple, $\{0, 1\}$. If it is in the region SW (i.e., $p^0 \le \epsilon$ and $p^1 \le \epsilon$), the prediction set is empty. Otherwise the prediction set is a singleton: $\{1\}$ if it is in NW ($p^0 \le \epsilon$ and $p^1 > \epsilon$) and $\{0\}$ if it is in SE ($p^0 > \epsilon$ and $p^1 \le \epsilon$). This is shown in the right panel of Figure 11.

However, a one-sided approach may be more appropriate in the case of the `Spambase` data set. There is a clear asymmetry of the two kinds of error in spam detection: classifying email as `spam` is much more harmful than letting occasional spam in. A reasonable approach is to start from a small number $\epsilon > 0$, the maximum tolerable percentage of email classified as `spam`, and then to try to minimize the percentage of spam classified as `email` under this constraint. For example, we can use the "one-sided label conditional ICP" classifying x as

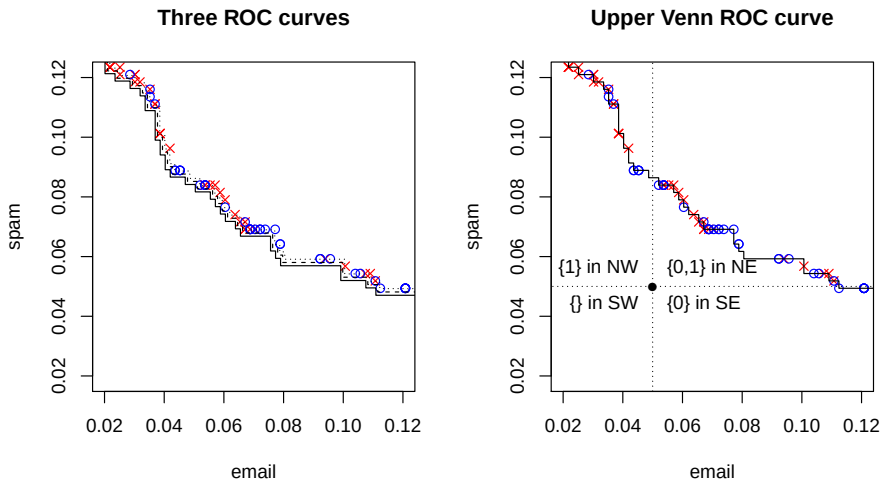

Figure 11: Left panel: the lower left corner of the lower left plot of Figure 2 with the empirical (solid), minimax (dashed), and Laplace (dotted) ROC curves. Right panel: the lower left corner of the lower left plot of Figure 2 with the upper Venn ROC curve and the partition of the plane corresponding to the label conditional ICP with significance level 5%.

`spam` if and only if[1] $p^0 \le \epsilon$ for x; otherwise, x is classified as `email`. In the case of $\epsilon = 5\%$, this means classifying a test object as `spam` if and only if it lands to the left of (or onto) the vertical dotted line in the right panel of Figure 11.

Both our procedures, two-sided and one-sided, look very similar to the standard uses of ROC curves. However, the standard justification of these uses presupposes that we know the true ROC curve. In practice, we only have access to an estimate of the true ROC curve, and the error of estimation is usually very significant. The upper Venn ROC curve is defined in terms of the data rather than the unknown true distribution. Despite this, we still have guarantees of validity. For example, our one-sided procedure guarantees that the (unconditional) probability of mistaking email for spam is at most ϵ (see Proposition 2).

This section of the paper raises a large number of questions. Not all inductive conformity measures are scoring-type; can other types be analyzed using the notion of ROC curves? Can other kinds of conditional ICPs be analyzed this way? What about smoothed ICPs? And even in the case of scoring-type label conditional ICPs we have not proved their property of training conditional validity (i.e., the version of Theorem 1 for label conditional ICPs).

[1] In practice, we might want to improve the predictor by adding another step and changing the classification from `spam` to `email` if p^1 is also small, in which case x looks neither like spam nor email. This step can usually be disregarded for scoring-type ICPs unless ϵ is very lax.

8 Conclusion

The goal of this paper has been to explore various versions of the requirement of conditional validity. With a small training set, we have to content ourselves with unconditional validity (or abandon any formal requirement of validity altogether). For bigger training sets training conditional validity will be approached by ICPs automatically, and we can approach example conditional validity by using conditional ICPs but making sure that the size of a typical category does not become too small (say, less than 100). In problems of binary classification, we can control false positive and false negative rates by using label conditional ICPs.

The known property of validity of inductive conformal predictors (Proposition 1) can be stated in the traditional statistical language (see, e.g., Fraser 1957 and Guttman 1970) by saying that they are $1-\epsilon$ expectation tolerance regions, where ϵ is the significance level. In classical statistics, however, there are two kinds of tolerance regions: $1-\epsilon$ expectation tolerance regions and PAC-type $1-\delta$ tolerance regions for a proportion $1-\epsilon$, in the terminology of Fraser (1957). We have seen (Theorem 1) that inductive conformal predictors are tolerance regions in the second sense as well (cf. Appendix A).

A disadvantage of inductive conformal predictors is their potential predictive inefficiency: indeed, the calibration set is wasted as far as the development of the prediction rule f in (1) is concerned, and the proper training set is wasted as far as the calibration (3) of conformity scores into p-values is concerned. Conformal predictors use the full training set for both purposes, and so can be expected to be significantly more efficient. (There have been reports of comparable and even better predictive efficiency of ICPs as compared to conformal predictors but they may be unusual artefacts of the methods used and particular data sets.) It is an open question whether we can guarantee training conditional validity under (11) or a similar condition for conformal predictors different from classical tolerance regions. Perhaps no universal results of this kind exist, and different families of conformal predictors will require different methods. See Appendix B for an empirical study of a simple conformal predictor.

Acknowledgments

I am grateful to Bob Williamson for a useful discussion. Many thanks to the reviewers the conference and journal versions of this paper for their suggestions, which led, in particular, to Appendix B and Figures 9 and 10. The empirical studies described in this paper used the R system, the `gbm` package for R written by Greg Ridgeway (based on the work of Freund and Schapire 1997 and Friedman 2001, 2002), MATLAB, and the C program for computing tangent distance written by Daniel Keysers and adapted to MATLAB by Aditi Krishn. This work was partially supported by the Cyprus Research Promotion Foundation.

References

Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. Elsevier, Waltham, MA, 2013. To appear.

Samy Bengio, Johnny Mariéthoz, and Mikaela Keller. The expected performance curve. In *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, 2005. URL `http://users.dsic.upv.es/~flip/ROCML2005/`.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

Donald A. S. Fraser. *Nonparametric Methods in Statistics*. Wiley, New York, 1957.

Donald A. S. Fraser and R. Wormleighton. Nonparametric estimation IV. *Annals of Mathematical Statistics*, 22:294–298, 1951.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 2002.

Irwin Guttman. *Statistical Tolerance Regions: Classical and Bayesian*. Griffin, London, 1970.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition, 2009.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

Jing Lei and Larry Wasserman. Distribution free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society B*, 2013. (to appear), preliminary version published as Technical Report arXiv:1203.5422 [stat.ME].

Jing Lei, James Robins, and Larry Wasserman. Distribution free prediction sets. *Journal of the American Statistical Association*, 108:278–287, 2013. Preliminary version published as Technical Report arXiv:1111.1418 [math.ST].

Jon Maindonald and John Braun. *Data Analysis and Graphics Using R: An Example-Based Approach*. Cambridge University Press, Cambridge, second edition, 2007.

Peter McCullagh, Vladimir Vovk, Ilia Nouretdinov, Dmitry Devetyarov, and Alex Gammerman. Conditional prediction intervals for linear regression. In *Proceedings of the Eighth International Conference on Machine Learning and Applications (December 13–15, Miami, FL)*, pages 131–138, 2009. Available from `http://www.stat.uchicago.edu/~pmcc/reports/predict.pdf`.

National Institute of Standards and Technology. Digital library of mathematical functions. 23 March 2012. URL `http://dlmf.nist.gov/`.

Ilia R. Nouretdinov. Offline Nearest Neighbour transductive Confidence Machine. In *Poster and Workshop Proceedings of the Eighth Industrial Conference on Data Mining*, pages 16–24, 2008.

Harris Papadopoulos, Konstantinos Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive Confidence Machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Proceedings of the Thirteenth European Conference on Machine Learning (August 19–23, 2002, Helsinki)*, volume 2430 of *Lecture Notes in Computer Science*, pages 345–356, Berlin, 2002a. Springer.

Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Qualified predictions for large data sets in the case of pattern recognition. In *Proceedings of the First International Conference on Machine Learning and Applications (June 24–27, 2002, Las Vegas, NV)*, pages 159–163, Las Vegas, NV, 2002b. CSREA Press.

Harris Papadopoulos, Alex Gammerman, and Vladimir Vovk, editors. *Special Issue of the* Annals of Mathematics and Artificial Intelligence *on Conformal Prediction and its Applications*. Springer, 2013. to appear.

Craig Saunders, Alex Gammerman, and Vladimir Vovk. Transduction with confidence and credibility. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (July 31 – August 6, 1999, Stockholm)*, volume 2, pages 722–726, San Francisco, CA, 1999. Morgan Kaufmann.

Henry Scheffé and John W. Tukey. Nonparametric estimation I: Validation of order statistics. *Annals of Mathematical Statistics*, 16:187–192, 1945.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2010.

John W. Tukey. Nonparametric estimation II: Statistically equivalent blocks and tolerance regions – the continuous case. *Annals of Mathematical Statistics*, 18:529–539, 1947.

John W. Tukey. Nonparametric estimation III: Statistically equivalent blocks and tolerance regions – the discontinuous case. *Annals of Mathematical Statistics*, 19:30–39, 1948.

Stijn Vanderlooy and Ida G. Sprinkhuizen-Kuyper. A comparison of two approaches to classify with guaranteed performance. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases (September 17–21, 2007, Warsaw)*, volume 4702 of *Lecture Notes in Computer Science*, pages 288–299, Berlin, 2007. Springer.

Stijn Vanderlooy, Laurens van der Maaten, and Ida Sprinkhuizen-Kuyper. Off-line learning with Transductive Confidence Machines: an empirical evaluation. In Petra Perner, editor, *Proceedings of the Fifth International Conference on Machine Learning and Data Mining in Pattern Recognition (July 18–20, 2007, Leipzig, Germany)*, volume 4571 of *Lecture Notes in Artificial Intelligence*, pages 310–323, Berlin, 2007. Springer.

Vladimir Vovk. On-line Confidence Machines are well-calibrated. In *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science (November 16–19, 2002, Vancouver)*, pages 187–196, Los Alamitos, CA, 2002. IEEE Computer Society.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine, editors, *JMLR: Workshop and Conference Proceedings*, volume 25 (Asian Conference on Machine Learning), pages 475–490, 2012.

Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning (June 27–30, 1999, Bled, Slovenia)*, pages 444–453, San Francisco, CA, 1999. Morgan Kaufmann.

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

Samuel S. Wilks. Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12:91–96, 1941.

A Training conditional validity for classical tolerance regions

In this appendix we compare Theorem 1 with the results (see, e.g., Fraser 1957 and Guttman 1970) about classical tolerance regions (which are a special case of conformal predictors, as explained in Vovk et al. 2005, p. 257). It is well known that under appropriate continuity assumptions the classical tolerance regions

that discard $\epsilon(n+1)$ out of the $n+1$ statistically equivalent blocks (in this appendix we always assume that $\epsilon(n+1)$ is an integer number) have coverage probability following the beta distribution with parameters $(1-\epsilon)(n+1)$ and $\epsilon(n+1)$ (see, e.g., Tukey 1947 or Guttman 1970, Theorems 2.2 and 2.3); in particular, their expected coverage probability is $1-\epsilon$. This immediately implies the following corollary: if Γ is a classical tolerance predictor with sample size n and expected coverage probability $1-\epsilon$, it is (E,δ)-valid if and only if

$$\delta \geq \mathrm{Bet}_{(1-\epsilon)(n+1),\epsilon(n+1)}(1-E) = 1 - \mathrm{Bet}_{\epsilon(n+1),(1-\epsilon)(n+1)}(E), \tag{32}$$

where $\mathrm{Bet}_{\alpha,\beta}$ is the cumulative beta distribution function with parameters α and β.

The following lemma shows that in fact (32) coincides with the condition (7) for ICPs (under our assumption $\epsilon(n+1) \in \mathbb{Z}$). Of course, n means different things in (7) and (32): the size of the calibration set in the former and the size of the full training set in the latter.

Lemma 3 (`http://dlmf.nist.gov/8.17.E5`). *For all $n \in \{1,2,\ldots\}$, all $k \in \{1,\ldots,n\}$, and all $E \in (0,1)$,*

$$\mathrm{bin}_{n,E}(k-1) = \mathrm{Bet}_{n+1-k,k}(1-E) = 1 - \mathrm{Bet}_{k,n+1-k}(E). \tag{33}$$

Proof. The equality between the last two terms of (33) is obvious. The last term of (33) is the probability that the kth smallest value in a sample of size n from the uniform probability distribution U on $[0,1]$ exceeds E. This event is equivalent to at most $k-1$ of n independent random variables generated from U belonging to the interval $[0,E]$, and so the probability of this event is given by the first term of (33). □

The assumption of continuity was removed by Tukey (1948) and Fraser and Wormleighton (1951). We will state this result only for the simplest kind of classical tolerance regions, essentially those introduced by Wilks (1941) (this special case was obtained already by Scheffé and Tukey 1945, p. 192). Suppose the object space $\mathbf{X}$ is a one-element set and the label space is $\mathbf{Y} = \mathbb{R}$ (therefore, we consider the problem of predicting real numbers without objects). For two numbers $L \leq U$ in the set $\{0,1,\ldots,n+1\}$ consider the set predictor $[y_{(L)}, y_{(U)}]$, where $y_{(i)}$ is the ith order statistics (the ith smallest value in the training set $(y_1,\ldots,y_n)$, except that $y_{(0)} := -\infty$ and $y_{(n+1)} := \infty$). This set predictor is (E,δ)-valid provided we have (32) with $(1-\epsilon)(n+1)$ replaced by $U-L$ and $\epsilon(n+1)$ replaced by $n+1+L-U$.

It is easy to see that Theorem 1(a) can in fact be deduced from Scheffé and Tukey's result. This follows from the interpretation of inductive conformal predictors as a "conditional" version of Wilks's predictors corresponding to $L := \epsilon(n+1)$ and $U := n+1$. After observing the proper training set we apply Wilks's predictors to the conformity scores α_i of the calibration examples to predict the conformity score of a test example; the set prediction of the conformity score for the test object is then transformed into the prediction set consisting of the labels leading to a score in the predicted range.

B Training conditional validity for conformal predictors

This appendix is a rudimentary empirical study of the training conditional validity of conformal predictors (see Nouretdinov 2008, Theorem 1, for a preliminary theoretical study). The top right figure in Table 1, 1.00%, estimates the standard deviation of the random percentage of errors made by the ICP. This random percentage of errors consists of two components:

- one minus the random coverage probability
- and the random percentage of errors for a given coverage probability.

The variance of the random coverage probability, which is distributed as $\mathrm{Bet}(950, 50)$ according to Appendix A, is

$$\frac{950}{1000} \times \frac{50}{1000} \times \frac{1}{1001}, \tag{34}$$

which corresponds to the standard deviation 0.69%. The conditional variance of the random percentage of errors for a given coverage probability (approximately 95%) is approximately

$$\frac{0.05 \times 0.95}{1000}, \tag{35}$$

which also corresponds to the standard deviation 0.69%. (Notice how similar the expressions (34) and (35) are: the only difference is that (34) has 1001 where (35) has 1000.) Therefore, the total variance of the random percentage of errors will be close to the sum of (34) and (35), which corresponds to the standard deviation $0.69\%\sqrt{2} \approx 0.98\%$. This agrees with Table 1: $1.00\% \approx 0.98\%$.

This ANOVA-type decomposition of the variance of the error rate for ICPs suggests looking at the standard deviation of the error rate for conformal predictors as a measure of their training conditional validity. (An even more natural measure of the training conditional validity would be the square root of the difference between the variance of the error rate and the variance $\epsilon(1-\epsilon)/n$ corresponding to a fixed coverage probability $1-\epsilon$, where ϵ is the significance level and n is the size of the test set; however, these two measures are monotonic functions of each other.) These standard deviations are given in Table 3 (see below for details). They suggest that ICPs and conformal predictors possess training conditional validity to a similar degree.

Table 3 describes experiments performed on the standard USPS data set (available on the Internet) of 9298 hand-written digits. Conformal predictors are defined in, e.g., Vovk et al. (2005). We test the 1-Nearest Neighbour ICP and 1-Nearest Neighbour conformal predictor, both based on the (inductive) conformity measure (6) with d the tangent distance. In the case of the ICP, we randomly choose three disjoint subsets of the USPS data set: a proper training set of size 1000, a calibration set of size 999, and a test set of size 1000. And in the case of the conformal predictor, we randomly choose two disjoint subsets: a

Table 3: Percentages of errors, multiple predictions, and empty predictions at significance levels 5% and 1% for the ICP and conformal predictor (CP) on the USPS data set. The results are given for the first 100 seeds for the MATLAB random number generator in the same format as in Table 1.

RNG seed	0	1	2	...	99	Average	St. dev.
errors for ICP at 5%	6.3%	5.1%	5.7%	...	4.9%	5.18%	0.95%
for CP at 5%	3.0%	4.5%	3.9%	...	3.2%	4.41%	0.96%
for ICP at 1%	1.2%	1.4%	1.7%	...	2.2%	1.06%	0.43%
for CP at 1%	0.7%	0.2%	0.7%	...	0.6%	0.85%	0.42%
multiple for ICP at 5%	0%	0%	0%	...	0%	0%	0%
for CP at 5%	0%	0%	0%	...	0%	0%	0%
for ICP at 1%	7.1%	4.1%	4.4%	...	3.0%	7.40%	2.53%
for CP at 1%	6.0%	11.3%	10.5%	...	8.1%	8.86%	3.30%
empty for ICP at 5%	5.5%	4.1%	4.1%	...	3.1%	3.78%	1.14%
for CP at 5%	1.3%	3.7%	2.4%	...	2.0%	2.82%	1.14%
for ICP at 1%	0%	0%	0%	...	0%	0%	0%
for CP at 1%	0%	0%	0%	...	0%	0%	0%

training set of size 999 and a test set of size 1000. For each choice we compute the percentages of errors, multiple, and empty set predictions. This is repeated 100 times. The experiments are run for two significance levels: 5%, in which case there are no multiple set predictions, and 1%, in which case there are no empty set predictions. Using the significance level 1% instead of 5% in (34) and (35) we obtain the predicted value of 0.44% for the standard deviation of the percentage of errors, which is close to the experimental results both for the ICP (0.43%) and for the conformal predictor (0.42%). For 5% the experimental results (0.95% and 0.96%) are also close to the predicted value (0.98%).

Our discussion so far in this appendix has ignored the fact that the standard deviations in Tables 1 and 3 are only estimates. The following figures give an idea of their sensitivity to the choice of the seeds for the random number generators:

- Using other seeds, instead of the standard deviation 1.00% in Table 1 we obtain: 0.91% (seeds 100–199), 0.87% (seeds 200–299), 1.09% (seeds 300–399), 0.94% (seeds 400–499).
- Instead of the standard deviations 0.95% for the ICP and 0.96% for the conformal predictor at 5% in Table 3 we obtain:
  - 0.93% for the ICP and 1.00% for the conformal predictor (seeds 100–199)
  - 0.93% for the ICP and 0.87% for the conformal predictor (seeds 200–299)

  - 1.04% for the ICP and 0.88% for the conformal predictor (seeds 300–399)
  - 0.81% for the ICP and 0.95% for the conformal predictor (seeds 400–499).

- Instead of the standard deviations 0.43% for the ICP and 0.42% for the conformal predictor at 1% in Table 3 we obtain:
  - 0.37% for the ICP and 0.44% for the conformal predictor (seeds 100–199)
  - 0.38% for the ICP and 0.43% for the conformal predictor (seeds 200–299)
  - 0.45% for the ICP and 0.43% for the conformal predictor (seeds 300–399)
  - 0.42% for the ICP and 0.43% for the conformal predictor (seeds 400–499).

We can see that the variability due to the choice of seeds does not affect our conclusion that the ICP and conformal predictor have comparable variability of coverage probability.

In conclusion, we discuss a theoretical result by Nouretdinov (2008, Theorem 1) about the 1-Nearest Neighbour conformal predictor. Nouretdinov's result is asymptotic, involving the term $o(1)$. A further complication is that it contains an error (Nouretdinov, private communication): the proof of Corollary 2 applies Hoeffding's inequality in a wrong way (the e^{-m}/ϵ in the last line of the proof should be $e^{-2\epsilon^2 m}$). In the case of the 1-Nearest Neighbour conformal predictor, Nouretdinov's corrected result replaces (11) by

$$E \geq \epsilon + 6^{1/3} \frac{\ln^{2/3} n}{(n\delta)^{1/3}}, \tag{36}$$

in our notation and ignoring the $o(1)$ term (i.e., replacing it by 0), where n is the size of the training set. The dependence on δ is much worse in (36) than in (11), and the dependence on n is also somewhat worse. Our empirical results suggest that Nouretdinov's result can be improved.