Journal of Machine Learning Research 5 (2004) 575-604 Submitted 1/04; Published 6/04

# A Universal Well-Calibrated Algorithm for On-line Classification

**Vladimir Vovk** VOVK@CS.RHUL.AC.UK
*Computer Learning Research Centre*
*Royal Holloway, University of London*
*Egham, Surrey TW20 0EX, UK*

**Editors:** Kristin Bennett and Nicolò Cesa-Bianchi

## Abstract

We study the problem of on-line classification in which the prediction algorithm, for each "significance level" $\delta$, is required to output as its prediction a range of labels (intuitively, those labels deemed compatible with the available data at the level $\delta$) rather than just one label; as usual, the examples are assumed to be generated independently from the same probability distribution $P$. The prediction algorithm is said to be "well-calibrated" for $P$ and $\delta$ if the long-run relative frequency of errors does not exceed $\delta$ almost surely w.r. to $P$. For well-calibrated algorithms we take the number of "uncertain" predictions (i.e., those containing more than one label) as the principal measure of predictive performance. The main result of this paper is the construction of a prediction algorithm which, for any (unknown) $P$ and any $\delta$: (a) makes errors independently and with probability $\delta$ at every trial (in particular, is well-calibrated for $P$ and $\delta$); (b) makes in the long run no more uncertain predictions than any other prediction algorithm that is well-calibrated for $P$ and $\delta$; (c) processes example $n$ in time $O(\log n)$.

**Keywords:** Transductive Confidence Machine, on-line prediction

## 1. Introduction

Typical machine learning algorithms output a point prediction for the label of an unknown object. This paper continues study of an algorithm called the Transductive Confidence Machine (TCM), introduced by Saunders et al. (1999) and Vovk et al. (1999), that complements its predictions with some measures of confidence. There are different ways of presenting TCM's output; in this paper (as in the related Vovk, 2002a,b) we use TCM as a "region predictor", in the sense that it outputs a nested family of prediction regions (indexed by the significance level $\delta$) rather than a point prediction.

Any TCM is well-calibrated when used in the on-line mode: for any significance level $\delta$ the long-run relative frequency of erroneous predictions does not exceed $\delta$. What makes this feature of TCM especially appealing is that it is far from being just an asymptotic phenomenon: a slight modification of TCM called randomized[1] TCM (rTCM) makes errors independently at different trials and with probability $\delta$ at each trial. The property of being well-calibrated then immediately follows by the Borel strong law of large numbers. Figure 1 shows the cumulative numbers of errors at the significance levels 1%–5% made on the well-known USPS data set of hand-written digits (randomly permuted); as expected, these are straight lines with the slope approximately equal to the significance level. For proofs and further information, see Vovk (2002a).

[1] Randomization is needed to break ties and deal efficiently with borderline cases.

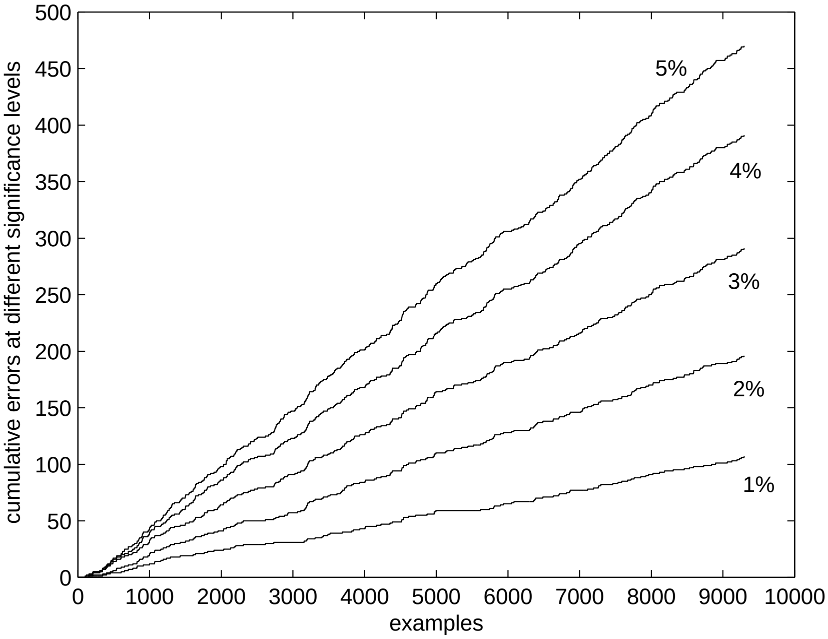


Figure 1: TCM's cumulative errors at the significance levels 1%–5% on the USPS data set

The justification of the study of TCM given by Vovk (2002a) was its good performance on real-world and standard benchmark data sets. For example, Figure 2 shows that for the significance levels between 1% and 5% most examples in the USPS data set can be predicted categorically (by a simple 1-Nearest Neighbour TCM, used in all experiments reported in this paper): the prediction region contains only one label.

This paper presents theoretical results about TCM's performance in the problem of classification, where the number of possible labels is finite; we show that there exists a *universal* rTCM, which, for any significance level $\delta$ and without knowing the true distribution $P$ generating the examples:

- produces, asymptotically, no more uncertain predictions than any other prediction algorithm that is well-calibrated for $P$ and $\delta$;
- produces, asymptotically, at least as many empty predictions as any other prediction algorithm that is well-calibrated for $P$ and $\delta$ and whose percentage of uncertain predictions is optimal (in the sense of the previous item).

The importance of the first item is obvious: we want to minimize the number of uncertain predictions. This asymptotic criterion ceases to work, however, when the number of uncertain predictions stabilizes, as in Figure 2 for significance levels 3%–5%. In such cases the number of empty predictions becomes important: empty predictions (automatically leading to an error) provide a warning that the object is atypical (looks very different from the previous objects), and one would like to be warned as often as possible, taking into account that the relative frequency of errors (including empty predictions) is guaranteed not to exceed $\delta$ in the long run. Remember that TCM outputs a whole family of prediction regions, so the fact that at some significance level the prediction region

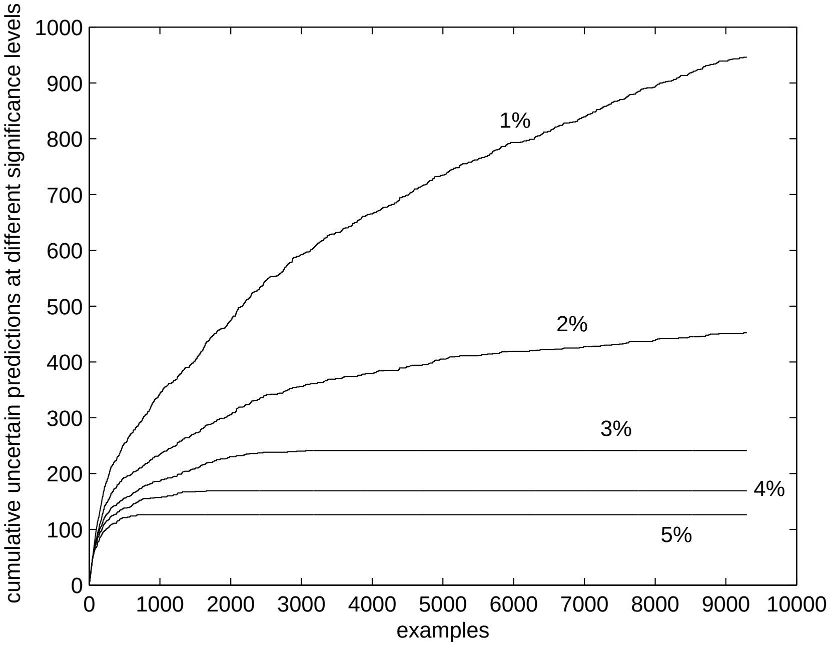


Figure 2: Cumulative number of "uncertain" predictions (i.e., prediction regions containing more than one label) made by the 1-Nearest Neighbour TCM at the significance levels 1%–5% on the USPS data set

becomes empty does not mean that all potential labels for a new object become equally likely: we should just shift our attention to other significance levels. Figure 3 shows the cumulative numbers of empty predictions for the USPS data set.

The full prediction output by a TCM is a complicated mathematical object: for each significance level $\delta$ we have a prediction region. In practice, a good starting point might be first to look at the prediction regions corresponding to two or three conventional significance levels, such as 1% and 5% (afterwards, of course, the prediction regions at other significance levels should be looked at). For example, denoting $\Gamma^\delta$ the prediction region at significance level $\delta$, we could say that the prediction is "highly certain" if $|\Gamma^{1\%}| \leq 1$ and "certain" if $|\Gamma^{5\%}| \leq 1$; similarly, we could say that the new object (whose label is being predicted) is "highly atypical" if $|\Gamma^{1\%}| = 0$ and "atypical" if $|\Gamma^{5\%}| = 0$. In the case of classification, the family of prediction regions $\Gamma^\delta$ can be summarized by reporting the *confidence*

$$\sup\{1-\delta : |\Gamma^\delta| \leq 1\},$$

the *credibility*

$$\inf\{\delta : |\Gamma^\delta| = 0\},$$

and the *prediction* $\Gamma^\delta$, where $1-\delta$ is the confidence (in the case of TCM, $|\Gamma^\delta| \leq 1$ and usually $|\Gamma^\delta| = 1$ when $1-\delta$ is the confidence). Reporting the prediction, confidence, and credibility, as in Saunders et al. (1999) and Vovk et al. (1999), is analogous to reporting the observed level of significance (Cox and Hinkley, 1974, p. 66) in statistics.

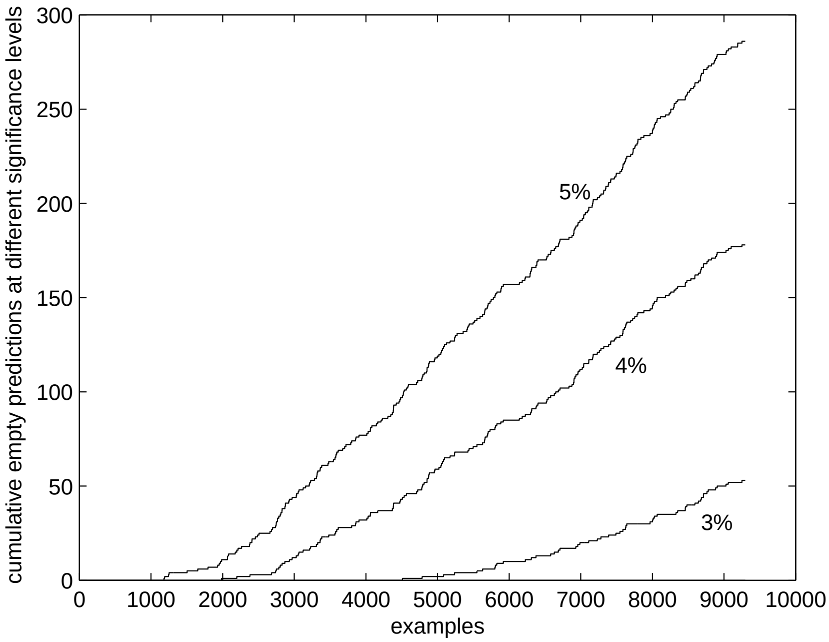


Figure 3: Cumulative number of empty predictions made by the 1-Nearest Neighbour TCM at the significance levels 1%–5% on the USPS data set (there are no empty predictions for 1% and 2%)

This paper's result elaborates on Vovk (2002b), where it was shown that an optimal randomized TCM exists when the distribution $P$ generating the examples is known. In the rest of this paper we consider only randomized TCM, so we drop the adjective "randomized".

The two areas of mainstream machine learning that are most closely connected with this paper are PAC learning theory and Bayesian learning theory. Whereas we often use the rich arsenal of mathematical tools developed in these fields, they do not provide the same kind of guarantees (the right probability of error at each significance level, with errors at different trials independent) under unknown $P$; for more details, see Vovk (2002a) and references therein. Several papers (such as Rivest and Sloan, 1988; Freund et al., 2004) extend the standard PAC framework by allowing the prediction algorithm to abstain from making a prediction at some trials. Our results show that for any significance level $\delta$ there exists a prediction algorithm that: (a) makes a wrong prediction with relative frequency at most $\delta$; (b) has an optimal frequency of abstentions among the prediction algorithms that satisfy property (a) (for details, see Remark 2 on p. 580). The paper by Freund et al. (2004) is especially close to the approach of this paper, defining a very natural TCM in the situation where a hypothesis class is given (the "empirical log ratio" of Freund et al. (2004), taken with appropriate sign, can be used as an "individual strangeness measure", as defined in §3).

## 2. Main Result

In our learning protocol, Reality outputs pairs $(x_1,y_1),(x_2,y_2),\dots$ called *examples*. Each example $(x_i,y_i)$ consists of an *object* $x_i$ and its *label* $y_i$. The objects are chosen from a measurable space $\mathbf{X}$

called the *object space* and the labels are elements of a measurable space $\mathbf{Y}$ called the *label space*. In this paper we assume that $\mathbf{Y}$ is finite (and endowed with the $\sigma$-algebra of all subsets). The protocol includes variables $\mathrm{Err}_n^\delta$ (the total number of errors made up to and including trial $n$ at significance level $\delta$) and $\mathrm{err}_n^\delta$ (the binary variable showing whether an error is made at trial $n$). It also includes analogous variables $\mathrm{Unc}_n^\delta$, $\mathrm{unc}_n^\delta$, $\mathrm{Emp}_n^\delta$, $\mathrm{emp}_n^\delta$ for uncertain and empty predictions:

```
$\mathrm{Err}_0^\delta := 0$, $\mathrm{Unc}_0^\delta := 0$, $\mathrm{Emp}_0^\delta := 0$ for all $\delta \in (0,1)$;
FOR $n = 1,2,\ldots$:
    Reality outputs $x_n \in \mathbf{X}$;
    Predictor outputs $\Gamma_n^\delta \subseteq \mathbf{Y}$ for all $\delta \in (0,1)$;
    Reality outputs $y_n \in \mathbf{Y}$;
    $\mathrm{err}_n^\delta := \begin{cases} 1 & \text{if } y_n \notin \Gamma_n^\delta \\ 0 & \text{otherwise} \end{cases}$, $\mathrm{Err}_n^\delta := \mathrm{Err}_{n-1}^\delta + \mathrm{err}_n^\delta$ for all $\delta \in (0,1)$;
    $\mathrm{unc}_n^\delta := \begin{cases} 1 & \text{if } |\Gamma_n^\delta| > 1 \\ 0 & \text{otherwise} \end{cases}$, $\mathrm{Unc}_n^\delta := \mathrm{Unc}_{n-1}^\delta + \mathrm{unc}_n^\delta$ for all $\delta \in (0,1)$;
    $\mathrm{emp}_n^\delta := \begin{cases} 1 & \text{if } |\Gamma_n^\delta| = 0 \\ 0 & \text{otherwise} \end{cases}$, $\mathrm{Emp}_n^\delta := \mathrm{Emp}_{n-1}^\delta + \mathrm{emp}_n^\delta$ for all $\delta \in (0,1)$
END FOR.
```

We will use the notation $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ for the *example space*; $\Gamma_n^\delta$ will be called the *prediction region* (or just *prediction*).

We will assume that each example $z_n = (x_n, y_n)$, $n = 1,2,\ldots$, is output according to a probability distribution $P$ in $\mathbf{Z}$ and the examples are independent of each other (so the sequence $z_1 z_2 \ldots$ is output by the power distribution $P^\infty$). This is Reality's randomized strategy.

A *region predictor* is a measurable function

$$\Gamma^\delta(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n), \tag{1}$$

where $\delta \in (0,1)$, $n = 1,2,\ldots$, the $(x_i, y_i) \in \mathbf{Z}$, $i = 1,\ldots,n-1$, are examples, $x_n \in \mathbf{X}$ is an object, and $\tau_i \in [0,1]$ ($i = 1,\ldots,n$), which satisfies

$$\Gamma^{\delta_1}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) \subseteq \Gamma^{\delta_2}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n)$$

whenever $\delta_1 \geq \delta_2$. The measurability of (1) means that for each $n$ the set

$$\left\{ (\delta, x_1, \tau_1, y_1, \ldots, x_n, \tau_n, y_n) : y_n \in \Gamma^\delta(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) \right\} \\ \subseteq (0,1) \times (\mathbf{X} \times [0,1] \times \mathbf{Y})^n$$

is measurable.

Since we are interested in prediction with confidence, the region predictor (1) is given an extra input $\delta \in (0,1)$, which we call the *significance level* (typically it is close to 0, standard values being 1% and 5%); the complementary value $1 - \delta$ is called the *confidence level*. We will always assume that $\tau_n$ are independent random variables uniformly distributed in $[0,1]$. This makes a region predictor a family (indexed by $\delta \in (0,1)$) of Predictor's randomized strategies.

We will often use the notation $\mathrm{err}_n^\delta$, $\mathrm{unc}_n^\delta$, etc., in the case where Reality and Predictor are using given randomized strategies. For example, $\mathrm{err}_n^\delta(P^\infty, \Gamma)$ is the random variable equal to 0 if Predictor is right at trial $n$ and at significance level $\delta$ and equal to 1 otherwise. It is always assumed that the random numbers $\tau_n$ used by $\Gamma$ and the random examples $z_n$ chosen by Reality are independent.

We say that a region predictor $\Gamma$ is (conservatively) *well-calibrated* for a probability distribution $P$ in $\mathbf{Z}$ and a significance level $\delta \in (0,1)$ if

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n^\delta(P^\infty,\Gamma)}{n} \le \delta \quad \text{a.s.}$$

We say (as in Vovk, 2002b) that $\Gamma$ is *optimal* for $P$ and $\delta$ if, for any region predictor $\Gamma^\dagger$ which is well-calibrated for $P$ and $\delta$,

$$\limsup_{n\to\infty} \frac{\mathrm{Unc}_n^\delta(P^\infty,\Gamma)}{n} \le \liminf_{n\to\infty} \frac{\mathrm{Unc}_n^\delta(P^\infty,\Gamma^\dagger)}{n} \quad \text{a.s.} \tag{2}$$

(It is natural to assume in this and other similar definitions that the random numbers used by $\Gamma$ and $\Gamma^\dagger$ are independent, but this assumption is not needed for our mathematical results and we do not make it.) Of course, the definition of optimality is natural only for well-calibrated $\Gamma$.

A region predictor $\Gamma$ is *universal well-calibrated* if:

- it is well-calibrated for any $P$ and $\delta$;
- it is optimal for any $P$ and $\delta$;
- for any $P$, any $\delta$, and any region predictor $\Gamma^\dagger$ which is well-calibrated and optimal for $P$ and $\delta$,
$$\liminf_{n\to\infty} \frac{\mathrm{Emp}_n^\delta(P^\infty,\Gamma)}{n} \ge \limsup_{n\to\infty} \frac{\mathrm{Emp}_n^\delta(P^\infty,\Gamma^\dagger)}{n} \quad \text{a.s.}$$

Recall that a measurable space $\mathbf{X}$ is *Borel* if it is isomorphic to a measurable subset of the interval $[0,1]$. The class of Borel spaces is very rich; for example, all Polish spaces (such as finite-dimensional Euclidean spaces $\mathbb{R}^n$, $\mathbb{R}^\infty$, functional spaces $C$ and $D$) are Borel.

**Theorem 1** *Suppose the object space* $\mathbf{X}$ *is Borel. There exists a universal well-calibrated region predictor.*

This is the main result of the paper. In §3 we construct a universal well-calibrated region predictor (processing example $n$ in time $O(\log n)$) and in §4 outline the idea of the proof that it indeed satisfies the required properties. Technical details will be given in §5.

**Remark** The protocol of Rivest and Sloan (1988) and Freund et al. (2004) is in fact a restriction of our protocol, in which Predictor is only allowed to output a one-element set or the whole of $\mathbf{Y}$; the latter is interpreted as abstention. (And in the situation where the numbers of errors and uncertain predictions are of primary interest, as in this paper, the difference between these two protocols is not significant.) The universal well-calibrated region predictor can be adapted to the restricted protocol by replacing an uncertain prediction with $\mathbf{Y}$ and replacing an empty prediction with a randomly chosen label. In this way we obtain a prediction algorithm in the restricted protocol which is well-calibrated and has an optimal frequency of abstentions, in the sense of (2), among the well-calibrated algorithms.

## 3. Construction of a Universal Well-Calibrated Region Predictor

In this section we first define the general notion of Transductive Confidence Machine, and then we specialize it using a nearest neighbours procedure to obtain a universal well-calibrated region predictor.

### 3.1 Preliminaries

If $\tau$ is a number in $[0,1]$, we split it into two numbers $\tau',\tau'' \in [0,1]$ as follows: if the binary expansion of $\tau$ is $0.a_1a_2\ldots$ (redefine the binary expansion of 1 to be $0.11\ldots$), set $\tau' := 0.a_1a_3a_5\ldots$ and $\tau'' := 0.a_2a_4a_6\ldots$. If $\tau$ is distributed uniformly in $[0,1]$, then both $\tau'$ and $\tau''$ are, and they are independent of each other.

We will often apply our procedures (e.g., the "individual strangeness measure" in §3.2, the Nearest Neighbours rule in §3.3) not to the original objects $x \in \mathbf{X}$ but to *extended objects* $(x,\sigma) \in \tilde{\mathbf{X}} := \mathbf{X} \times [0,1]$, where $x$ is complemented by a random number $\sigma$ (to be extracted from one of the $\tau_n$). In other words, along with examples $(x,y)$ we will also consider *extended examples* $(x,\sigma,y) \in \tilde{\mathbf{Z}} := \mathbf{X} \times [0,1] \times \mathbf{Y}$.

Let us set $\mathbf{X} := [0,1]$; we can do this without loss of generality since $\mathbf{X}$ is Borel. This makes the extended object space $\tilde{\mathbf{X}} = [0,1]^2$ a linearly ordered set with the *lexicographic order*: $(x_1,\sigma_1) < (x_2,\sigma_2)$ means that either $x_1 = x_2$ and $\sigma_1 < \sigma_2$ or $x_1 < x_2$. We say that $(x_1,\sigma_1)$ is *nearer* to $(x_3,\sigma_3)$ than $(x_2,\sigma_2)$ is if

$$|x_1 - x_3, \sigma_1 - \sigma_3| < |x_2 - x_3, \sigma_2 - \sigma_3|, \tag{3}$$

where

$$|x,\sigma| := \begin{cases} (x,\sigma) & \text{if } (x,\sigma) \geq (0,0) \\ (-x,-\sigma) & \text{otherwise.} \end{cases} \tag{4}$$

The value $|x_1 - x_2, \sigma_1 - \sigma_2|$ plays the role of the distance between extended objects $(x_1,\sigma_1)$ and $(x_2,\sigma_2)$. Despite such distances being two-dimensional, they are still always comparable using the lexicographic order.

Our construction will be based on the Nearest Neighbours algorithm, which is known to be strongly universally consistent in the traditional theory of pattern recognition (see, e.g., Devroye et al., 1996, Chapter 11); the random components $\sigma$ are needed for tie-breaking.

### 3.2 Transductive Confidence Machines

Transductive Confidence Machine, or TCM, is a way of transition from what we call an "individual strangeness measure" to a region predictor. A family of measurable functions $\{A_n : n = 1,2,\ldots\}$, where $A_n : \tilde{\mathbf{Z}}^n \to \mathbb{R}^n$ for all $n$, is called an *individual strangeness measure* if, for any $n = 1,2,\ldots$, each $\alpha_i$ in

$$A_n : (w_1,\ldots,w_n) \mapsto (\alpha_1,\ldots,\alpha_n) \tag{5}$$

is determined by $w_i$ and the multiset $\wr w_1,\ldots,w_n \int$. (The difference between a multiset $\wr w_1,\ldots,w_n \int$ and a set $\{w_1,\ldots,w_n\}$ is that the former can contain several copies of the same element.)

The *TCM associated with an individual strangeness measure* $A_n$ is the following region predictor $\Gamma^\delta(x_1,\tau_1,y_1,\ldots,x_{n-1},\tau_{n-1},y_{n-1},x_n,\tau_n)$: at any trial $n$ and for any label $y \in \mathbf{Y}$, define

$$(\alpha_1,\ldots,\alpha_n) := A_n((x_1,\tau'_1,y_1),\ldots,(x_{n-1},\tau'_{n-1},y_{n-1}),(x_n,\tau'_n,y)),$$

and include $y$ in $\Gamma^\delta$ if and only if

$$\tau''_n < \frac{\#\{i=1,\dots,n:\alpha_i \ge \alpha_n\} - n\delta}{\#\{i=1,\dots,n:\alpha_i = \alpha_n\}} \tag{6}$$

(in particular, include $y$ in $\Gamma^\delta$ if $\#\{i=1,\dots,n:\alpha_i > \alpha_n\}/n > \delta$ and do not include $y$ in $\Gamma^\delta$ if $\#\{i=1,\dots,n:\alpha_i \ge \alpha_n\}/n \le \delta$).

A *TCM* is the TCM associated with some individual strangeness measure. It was shown in Vovk (2002a) that

**Proposition 2** *Every TCM is well-calibrated for every $P$ and $\delta$.*

The definition of TCM can be illustrated by the following simple example of an individual strangeness measure, the one used in producing Figures 1–3: mapping (5) can be defined, in the spirit of the 1-Nearest Neighbour Algorithm, as (assuming the objects are vectors in a Euclidean space)

$$\alpha_i := \frac{\min_{j\ne i: y_j = y_i} d(x_i,x_j)}{\min_{j\ne i: y_j \ne y_i} d(x_i,x_j)},$$

where $d$ is the Euclidean distance (i.e., an object is considered strange if it is in the middle of objects labelled in a different way and is far from the objects labelled in the same way).

### 3.3 Universal TCM

Fix a monotonically non-decreasing sequence of integer numbers $K_n$, $n=1,2,\dots$, such that

$$K_n \to \infty,\ K_n = o\left(\sqrt{n/\ln n}\right) \tag{7}$$

as $n\to\infty$. The *Nearest Neighbours TCM* is defined as follows. Let $w_1,\dots,w_n$ be a sequence of extended examples $w_i = (x_i,\sigma_i,y_i)$. To define the corresponding $\alpha$s , as seen in (5), we first define Nearest Neighbours approximations $P_n^{\ne}(y\,|\,x_i,\sigma_i)$ to the true (but unknown) conditional probabilities $P(y\,|\,x_i)$: for every extended example $(x_i,\sigma_i,y_i)$ in the sequence,

$$P_n^{\ne}(y\,|\,x_i,\sigma_i) := N^{\ne}(x_i,\sigma_i,y)/K_n, \tag{8}$$

where $N^{\ne}(x_i,\sigma_i,y)$ is the number of $j=1,\dots,n$ such that $y_j=y$ and $(x_j,\sigma_j)$ is one of the $K_n$ nearest neighbours, in the sense of (3), of $(x_i,\sigma_i)$ in the sequence

$$((x_1,\sigma_1),\dots,(x_{i-1},\sigma_{i-1}),(x_{i+1},\sigma_{i+1}),\dots,(x_n,\sigma_n)).$$

(The upper index $^{\ne}$ reminds us of the fact that $(x_i,\sigma_i)$ is not counted as one of its own nearest neighbours in this definition.) If $K_n \ge n$ or $K_n \le 0$, this definition does not work, so set, e.g., $P_n^{\ne}(y\,|\,x_i,\sigma_i) := 1/|\mathbf{Y}|$ for all $y$ and $i$ (this particular convention is not essential since, by (7), $0 < K_n < n$ from some $n$ on). If the expression "$K_n$ nearest neighbours" is not defined because of distance ties, we again set $P_n^{\ne}(y\,|\,x_i,\sigma_i) := 1/|\mathbf{Y}|$ for all $y$ and $i$ (this convention is not essential since distance ties happen with probability zero).

Define the "empirical predictability function" $f_n^{\ne}$ by

$$f_n^{\ne}(x_i,\sigma_i) := \max_{y\in\mathbf{Y}} P_n^{\ne}(y\,|\,x_i,\sigma_i). \tag{9}$$

For each $(x_i,\sigma_i)$ fix some

$$\hat{y}_n(x_i,\sigma_i) \in \arg\max_y P_n^{\neq}(y\,|\,x_i,\sigma_i) \tag{10}$$

(e.g., take the first element of $\arg\max_y P_n^{\neq}(y\,|\,x_i,\sigma_i)$ in a fixed ordering of $\mathbf{Y}$) and define the mapping (5) (where $w_i = (x_i,\sigma_i,y_i)$, $i = 1,\dots,n$) setting

$$\alpha_i := \begin{cases} -f_n^{\neq}(x_i,\sigma_i) & \text{if } y_i = \hat{y}_n(x_i,\sigma_i) \\ f_n^{\neq}(x_i,\sigma_i) & \text{otherwise.} \end{cases} \tag{11}$$

This completes the definition of the Nearest Neighbours TCM, which will later be shown to be universal.

**Proposition 3** *Let $\Delta \subseteq (0,1)$ be finite. If $\mathbf{X} = [0,1]$ and $K_n \to \infty$ sufficiently slowly, the Nearest Neighbours TCM can be implemented for significance levels $\delta \in \Delta$ so that the computations at trial $n$ are performed in time $O(\log n)$.*

Proposition 3 assumes a computational model that allows operations (such as comparison) with real numbers. If $\mathbf{X}$ is an arbitrary Borel space, for this proposition to be applicable $\mathbf{X}$ should be embedded in $[0,1]$ first; e.g., if $\mathbf{X} \subseteq [0,1]^n$, an $x = (x_1,\dots,x_n) \in \mathbf{X}$ can be represented as

$$(x_{1,1},x_{2,1},\dots,x_{n,1},x_{1,2},x_{2,2},\dots,x_{n,2},\dots) \in [0,1],$$

where $0.x_{i,1}x_{i,2}\dots$ is the binary expansion of $x_i$. We use the expression “can be implemented” in a wide sense, only requiring that the implementation should give the correct results almost surely.

## 4. Fine Details of Region Prediction

In this section we make first steps towards the proof of Theorem 1. Let $P$ be the true distribution in $\mathbf{Z}$ generating the examples. We denote by $P_{\mathbf{X}}$ the marginal distribution of $P$ in $\mathbf{X}$ (i.e., $P_{\mathbf{X}}(E) := P(E \times \mathbf{Y})$) and by $P_{\mathbf{Y}\,|\,\mathbf{X}}(y\,|\,x)$ the conditional probability that, for a random example $(X,Y)$ chosen from $P$, $Y = y$ provided $X = x$ (we fix arbitrarily a regular version of this conditional probability). We will often omit lower indices $_{\mathbf{X}}$ and $_{\mathbf{Y}\,|\,\mathbf{X}}$ and $P$ itself from our notation.

The *predictability* of an object $x \in \mathbf{X}$ is

$$f(x) := \max_{y\in\mathbf{Y}} P(y\,|\,x)$$

and the *predictability distribution function* is the function $F : [0,1] \to [0,1]$ defined by

$$F(\beta) := P\{x : f(x) \le \beta\}.$$

An example of such a function $F$ is given in Figure 4 (left), where the graph of $F$ is the thick line.

The *success curve* $\mathbf{S}_P$ of $P$ is defined by the equality

$$\mathbf{S}_P(\delta) = \inf\left\{B \in [0,1] : \int_0^1 (F(\beta) - B)^+ d\beta \le \delta\right\}, \tag{12}$$

where $t^+$ stands for $\max(t,0)$; the function $\mathbf{S}_P$ is also of the type $[0,1] \to [0,1]$. Geometrically, $\mathbf{S}_P(\delta)$ is defined from the graph of $F$ as follows (see Figure 4, left; we often drop the lower index $_P$):

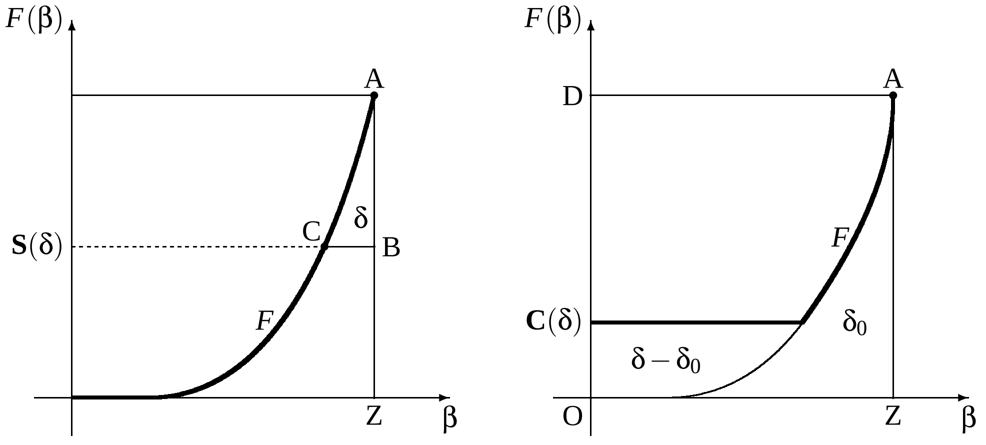


Figure 4: The predictability distribution function $F$ and the success curve $\mathbf{S}(\delta)$ (left); the complementary success curve $\mathbf{C}(\delta)$ (right)

move the point B from A to Z until the area of the curvilinear triangle ABC becomes $\delta$ or B reaches Z; the ordinate of B is then $\mathbf{S}(\delta)$.

The *complementary success curve* $\mathbf{C}_P$ of $P$ is defined by

$$\mathbf{C}_P(\delta) = \sup\left\{B \in [0,1] : B + \int_0^1 (F(\beta) - B)^+ d\beta \le \delta\right\}, \tag{13}$$

where $\sup\emptyset$ is interpreted as 0. Similarly to the case of $\mathbf{S}(\delta)$, $\mathbf{C}(\delta)$ is defined as the value such that the area of the part of the box AZOD below the thick line in Figure 4 (right) is $\delta$ ($\mathbf{C}(\delta) = 0$ if such a value does not exist).

Define the *critical significance level* $\delta_0$ as

$$\delta_0 := \int_0^1 F(\beta) d\beta. \tag{14}$$

It is clear that

$$\delta \le \delta_0 \Longrightarrow \int_0^1 (F(\beta) - \mathbf{S}(\delta))^+ d\beta = \delta \;\&\; \mathbf{C}(\delta) = 0$$

$$\delta \ge \delta_0 \Longrightarrow \mathbf{S}(\delta) = 0 \;\&\; \mathbf{C}(\delta) + \int_0^1 (F(\beta) - \mathbf{C}(\delta))^+ d\beta = \delta.$$

The following result is proved in Vovk (2002b).

**Proposition 4** *Let $P$ be a probability distribution in $\mathbf{Z}$ and $\delta \in (0,1)$ be a significance level. If a region predictor $\Gamma$ is well-calibrated for $P$ and $\delta$, then*

$$\liminf_{n\to\infty} \frac{\mathrm{Unc}_n^{\delta}(P^\infty, \Gamma)}{n} \ge \mathbf{S}_P(\delta) \quad \textit{a.s.} \tag{15}$$

In this paper we complement Proposition 4 with

**Proposition 5** *Let $P$ be a probability distribution in $\mathbf{Z}$ and $\delta \in (0,1)$ be a significance level. If a region predictor $\Gamma$ is well-calibrated for $P$ and $\delta$ and satisfies*

$$\limsup_{n\to\infty} \frac{\mathrm{Unc}_n^\delta(P^\infty,\Gamma)}{n} \le \mathbf{S}_P(\delta) \quad a.s., \tag{16}$$

*then*

$$\limsup_{n\to\infty} \frac{\mathrm{Emp}_n^\delta(P^\infty,\Gamma)}{n} \le \mathbf{C}_P(\delta) \quad a.s.$$

Theorem 1 immediately follows from Propositions 2, 4, 5 and the following proposition.

**Proposition 6** *Suppose $\mathbf{X}$ is Borel. The Nearest Neighbours TCM constructed in §3.3 satisfies, for any $P$ and any significance level $\delta$,*

$$\limsup_{n\to\infty} \frac{\mathrm{Unc}_n^\delta(P^\infty,\Gamma)}{n} \le \mathbf{S}_P(\delta) \quad a.s. \tag{17}$$

*and*

$$\liminf_{n\to\infty} \frac{\mathrm{Emp}_n^\delta(P^\infty,\Gamma)}{n} \ge \mathbf{C}_P(\delta) \quad a.s. \tag{18}$$

## 5. Proofs

In this section we will assume that all extended objects $(x_i, \tau_i') \in [0,1]^2$, where $x_i$ are output by Reality and $\tau_i$ are the random numbers used, are different and that all pairwise distances between them are also different (this is true with probability one, since $\tau_i'$ are independent random numbers uniformly distributed in $[0,1]$).

### 5.1 Proof Sketch of Proposition 3

Without loss of generality we assume that $\Delta$ contains only one significance level $\delta$, which will be omitted from our notation. Our computational model has an operation of splitting $\tau \in [0,1]$ into $\tau'$ and $\tau''$ (or is allowed to generate both $\tau_n'$ and $\tau_n''$ at every trial $n$).

We will use two main data structures in our implementation of the Nearest Neighbours TCM:

- a red-black binary *search tree*;[2]
- a growing *array* of nonnegative integers indexed by $k \in \{-K_n, -K_n+1, \dots, K_n\}$ (where $n$ is the ordinal number of the example being processed).

Immediately after processing the $n$th extended example $(x_n, \tau_n, y_n)$ the contents of these data structures are as follows:

- The search tree contains $n$ vertices, corresponding to the extended examples $(x_i, \tau_i, y_i)$ seen so far. The key of vertex $i$ is the extended object $(x_i, \tau_i') \in [0,1]^2$; the linear order on the keys is the lexicographic order. The other information contained in vertex $i$ is the random number $\tau_i''$,

[2] See, e.g., Cormen et al. (2001), Chapters 12–14. The only two operations on red-black trees we need in this paper are the query SEARCH and the modifying operation INSERT.

the label $y_i$, the set $\{P_n^{\neq}(y\,|\,x_i,\tau'_i):y\in\mathbf{Y}\}$ of conditional probability estimates (8), the pointer to the following vertex (i.e., the vertex that has the smallest key greater than $(x_i,\tau'_i)$; if there is no greater key, the pointer is NIL), and the pointer to the previous vertex (i.e., the vertex that has the greatest key smaller than $(x_i,\tau'_i)$; if $(x_i,\tau'_i)$ is the smallest key, the pointer is NIL).

- The array contains the numbers

$$N(k) := \#\{i=1,\dots,n:\alpha_i = k/K_n\}$$

($\alpha_i$ are defined by (11) with $\sigma_i := \tau'_i$).

Notice that the information contained in vertex $i$ of the search tree is sufficient to find $\hat{y}_n(x_i,\tau'_i)$ and $\alpha_i$ in time $O(1)$.

We will say that an extended object $(x_j,\tau'_j)$ is in the *vicinity* of an extended object $(x_i,\tau'_i)$, $i\neq j$, if there are less than $K_n$ extended objects $(x_k,\tau'_k)$ (strictly) between $(x_i,\tau'_i)$ and $(x_j,\tau'_j)$.

When a new object $x_n$ becomes known, the algorithm does the following:

- Generates $\tau'_n$ and $\tau''_n$.
- Locates the successor and predecessor of $(x_n,\tau'_n)$ in the search tree (using the query SEARCH and the pointers to the following and previous vertices); this requires time $O(\log n)$.
- Computes the estimated conditional probabilities $\{P_n^{\neq}(y\,|\,x_n,\tau'_n):y\in\mathbf{Y}\}$; this also gives $\hat{y}_n(x_n,\tau'_n)$. This involves scanning the vicinity of $(x_n,\tau'_n)$ for the $K_n$ nearest neighbours of $(x_n,\tau'_n)$, which can be done in time $O(K_n)$: the $K_n$ nearest neighbours can be extracted from the vicinity of $(x_n,\tau'_n)$ sorted in the order of increasing distances from $(x_n,\tau'_n)$; since initially the vicinity consists of two sorted lists (to the left and to the right of $(x_n,\tau'_n)$), the procedure MERGE used in the merge sort algorithm (see, e.g., Cormen et al. 2001, §2.3.1) will sort the whole vicinity in time $O(K_n)$. Therefore, the required time is $O(K_n)=O(\log n)$.
- For each $y\in\mathbf{Y}$ looks at what happens if the $n$th example is $(x_n,\tau_n,y_n)=(x_n,\tau_n,y)$: computes $\alpha_n$ and updates (if necessary) $\alpha_i$ for $(x_i,\tau'_i)$ in the vicinity of $(x_n,\tau'_n)$; using the array and $\tau''_n$, finds whether $y\in\Gamma_n$. This requires time $O(K_n^2)=O(\log n)$, since there are $O(K_n)$ $\alpha_i$'s in the vicinity of $(x_n,\tau'_n)$ and each of them can be computed in time $O(K_n)$.
- Outputs the prediction region $\Gamma_n$ (time $O(1)$).

When the label $y_n$ arrives, the algorithm:

- Inserts the new vertex $(x_n,\tau'_n,\tau''_n,y_n,\{P_n^{\neq}(y\,|\,x_n,\tau'_n):y\in\mathbf{Y}\})$ in the search tree, repairs the pointers to the following and previous elements for $(x_n,\tau'_n)$'s left and right neighbours, initializes the pointers to the following and previous elements for $(x_n,\tau'_n)$ itself, and rebalances the tree (time $O(\log n)$).
- Updates (if necessary) the conditional probabilities

$$\{P_{n-1}^{\neq}(y\,|\,x_i,\tau'_i):y\in\mathbf{Y}\}\mapsto\{P_n^{\neq}(y\,|\,x_i,\tau'_i):y\in\mathbf{Y}\}$$

for the $2K_n$ existing vertices $(x_i,\tau'_i)$ in the vicinity of $(x_n,\tau'_n)$; this requires time $O(K_n^2)=O(\log n)$. The conditional probabilities for other $(x_i,\tau'_i)$, $i=1,\dots,n-1$, do not change.

- Updates the array, changing $N(K_n\alpha_i)$ for the $(x_i,\tau'_i) \neq (x_n,\tau'_n)$ in the vicinity of $(x_n,\tau'_n)$ and for both old and new values of $\alpha_i$ and changing $N(K_n\alpha_n)$ (time $O(K_n) = O(\log n)$).

In conclusion we discuss how to do the updates required when $K_n$ changes. At the critical trials $n$ when $K_n$ changes the array and the estimated conditional probabilities $P_n^{\neq}(y\,|\,x_i,\tau'_i)$ have to be recomputed, which, if done naively, would require time $\Theta(nK_n)$.

The assumption we have made about $K_n$ so far is that $K_n = O(\sqrt{\log n})$. We now also assume that $K_n$ is monotonic non-decreasing and

$$\#\{n\!:\!K_n < c\} = O\left(\#\{n\!:\!K_n = c\}\right) \tag{19}$$

as $c \to \infty$. This is the full explication of the "$K_n \to \infty$ sufficiently slowly" in the statement of the lemma, as used in this proof.

An *epoch* is defined to be a maximal sequence of $n$s with the same $K_n$. Since the changes that need to be done when a new epoch starts are substantial, they will be spread over the whole preceding epoch; we will only discuss updating the estimated conditional probabilities $P_n^{\neq}(y\,|\,x_i,\tau'_i)$: the array is treated similarly. An epoch is *odd* if the corresponding $K_n$ is odd and *even* if $K_n$ is even. At every step in an epoch we prepare the ground for the next epoch. By the end of epoch $n = A+1, A+2, \ldots, B$ we need to change $B$ sets $\{P_n^{\neq}(y\,|\,x_i,\tau'_i)\!:\!y \in \mathbf{Y}\}$ in $B-A$ steps (the duration of the epoch). Therefore, each vertex of the search tree should contain not only $\{P_n^{\neq}(y\,|\,x_i,\tau'_i)\}$ for the current epoch but also $\{P_n^{\neq}(y\,|\,x_i,\tau'_i)\}$ for the next epoch (two structures for holding $\{P_n^{\neq}(y\,|\,x_i,\tau'_i)\}$ will suffice, one for even epochs and one for odd epochs). Our assumptions of the slow growth of $K_n$, as seen in 19), imply that $B = O(B-A)$. This means that at each step $O(1)$ sets $\{P_n^{\neq}(y\,|\,x_i,\tau'_i)\}$ for the next epoch should be added. This will take time $O(K_n) = O(\log n)$. As soon as a set $\{P_n^{\neq}(y\,|\,x_i,\tau'_i)\!:\!y \in \mathbf{Y}\}$ for the next epoch is added at some trial, both sets (for the current and next epoch) will have to be updated for each new example.

### 5.2 Proof Sketch of Proposition 5

The proof of Proposition 5 is similar to (but more complicated than) the proof of Theorems 1 and 1r in Vovk (2002b); this proof sketch can be made rigorous using the Neyman–Pearson lemma, as in Vovk (2002b).

We will use the notations $g'_{\text{left}}$ and $g'_{\text{right}}$ for the left and right derivatives, respectively, of a function $g$. The following lemma parallels Lemma 2 in Vovk (2002b), which deals with $\mathbf{S}(\delta)$.

**Lemma 7** *The complementary success curve $\mathbf{C} : [0,1] \to [0,1]$ always satisfies these properties:*

1. *There is a point $\delta_0 \in [0,1]$ (namely, the critical significance level) such that $\mathbf{C}(\delta) = 0$ for $\delta \leq \delta_0$ and $\mathbf{C}(\delta)$ is concave for $\delta \geq \delta_0$.*

2. *$\mathbf{C}'_{\text{right}}(\delta_0) < \infty$ and $\mathbf{C}'_{\text{left}}(1) \geq 1$; therefore, for $\delta \in (\delta_0, 1)$, $1 \leq \mathbf{C}'_{\text{right}}(\delta) \leq \mathbf{C}'_{\text{left}}(\delta) < \infty$ and the function $\mathbf{C}(\delta)$ is increasing.*

3. *$\mathbf{C}(\delta)$ is continuous at $\delta = \delta_0$; therefore, it is continuous everywhere in $[0,1]$.*

*If a function $\mathbf{C} : [0,1] \to [0,1]$ satisfies these properties, there exist a measurable space $\mathbf{X}$, a finite set $\mathbf{Y}$, and a probability distribution $P$ in $\mathbf{X} \times \mathbf{Y}$ for which $\mathbf{C}$ is the complementary success curve.*

**Proof sketch** The statement of the lemma follows from the fact that the complementary success curve **C** can be obtained from the predictability distribution function $F$ using these steps (labelling the horizontal and vertical axes as $x$ and $y$ respectively):

1. Invert $F$: $F_1 := F^{-1}$.
2. Integrate $F_1$: $F_2(x) := \int_0^x F_1(t)dt$.
3. Increase $F_2$: $F_3(x) := F_2(x) + \delta_0$, where $\delta_0 := \int_0^1 F(x)dx$.
4. Invert $F_3$: $F_4 := F_3^{-1}$.

It can be shown that $\mathbf{C} = F_4$, if we define $g^{-1}(y) := \sup\{x : g(x) \le y\}$ for non-decreasing $g$ (so that $g^{-1}$ is continuous on the right). ■

Complement the protocol of §2 in which Reality plays $P^\infty$ and Predictor plays $\Gamma$ with the following variables:

$$\begin{aligned}
\overline{\mathrm{err}}_n &:= (P \times \mathbf{U})\bigl\{(x,y,\tau) : y \notin \Gamma^\delta(x_1,\tau_1,y_1,\ldots,x_{n-1},\tau_{n-1},y_{n-1},x,\tau)\bigr\},\\
\overline{\mathrm{unc}}_n &:= (P_{\mathbf{X}} \times \mathbf{U})\bigl\{(x,\tau) : |\Gamma^\delta(x_1,\tau_1,y_1,\ldots,x_{n-1},\tau_{n-1},y_{n-1},x,\tau)| > 1\bigr\},\\
\overline{\mathrm{emp}}_n &:= (P_{\mathbf{X}} \times \mathbf{U})\bigl\{(x,\tau) : |\Gamma^\delta(x_1,\tau_1,y_1,\ldots,x_{n-1},\tau_{n-1},y_{n-1},x,\tau)| = 0\bigr\},
\end{aligned}$$

$\delta$ being fixed and $\mathbf{U}$ standing for the uniform distribution in $[0,1]$, and

$$\overline{\mathrm{Err}}_n := \sum_{i=1}^n \overline{\mathrm{err}}_i, \quad \overline{\mathrm{Unc}}_n := \sum_{i=1}^n \overline{\mathrm{unc}}_i, \quad \overline{\mathrm{Emp}}_n := \sum_{i=1}^n \overline{\mathrm{emp}}_i .$$

By the martingale strong law of large numbers, to prove the proposition it suffices to consider only these "predictable" versions of $\mathrm{Err}_n$, $\mathrm{Unc}_n$, and $\mathrm{Emp}_n$: indeed, since $\mathrm{Err}_n - \overline{\mathrm{Err}}_n$, $\mathrm{Unc}_n - \overline{\mathrm{Unc}}_n$, and $\mathrm{Emp}_n - \overline{\mathrm{Emp}}_n$ are martingales (with increments bounded by 1 in absolute value) with respect to the filtration $\mathcal{F}_n$, $n = 0,1,\ldots$, where each $\mathcal{F}_n$ is generated by $(x_1,\tau_1,y_1),\ldots,(x_n,\tau_n,y_n)$, we have

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n - \overline{\mathrm{Err}}_n}{n} = 0 \qquad \text{a.s.,}$$

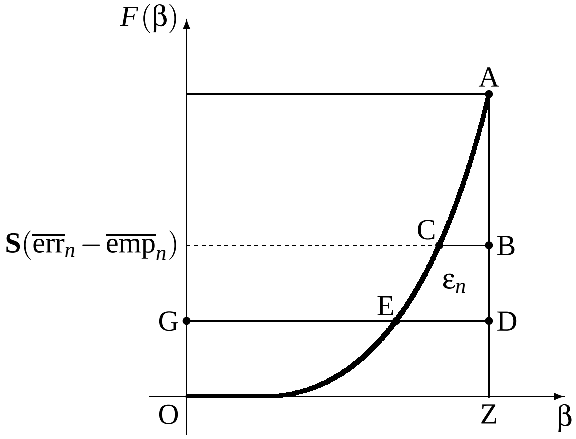
$$\lim_{n\to\infty} \frac{\mathrm{Unc}_n - \overline{\mathrm{Unc}}_n}{n} = 0 \qquad \text{a.s.,}$$

and

$$\lim_{n\to\infty} \frac{\mathrm{Emp}_n - \overline{\mathrm{Emp}}_n}{n} = 0 \qquad \text{a.s.}$$

(See, e.g., Shiryaev, 1996, Theorem VII.5.4.)

Without loss of generality we can assume that Predictor's move $\Gamma_n$ at trial $n$ is $\{\hat{y}(x_n)\}$ (where $x \mapsto \hat{y}(x) \in \arg\max_y P(y \mid x)$ is a fixed "choice function") or the empty set $\emptyset$ or the whole label space $\mathbf{Y}$. Furthermore, we can assume that, at every trial, the predictions are certain for the new objects



$F(\beta)$
A
$\mathbf{S}(\overline{\text{err}}_n - \overline{\text{emp}}_n)$
C
B
$\varepsilon_n$
E
G
D
O
Z
$\beta$

Figure 5: An admissible region predictor. The thick line is the predictability distribution function $F$; the area of the curvilinear triangle ABC is $\overline{\text{err}}_n - \overline{\text{emp}}_n$; the area of the rectangle DZOG is $\overline{\text{emp}}_n$; the (non-negative) area of the curvilinear quadrangle BDEC is denoted $\varepsilon_n$

above the straight line BC in Figure 5,[3] and that the predictions are empty for the objects below the straight line DG in Figure 5.[4] It is clear that for the region predictor to satisfy (16) it must hold that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i \wedge \overline{\text{emp}}_i) = 0$$

(otherwise $\overline{\text{Unc}}_n$ can be decreased substantially, which contradicts (15); $\varepsilon_i$ are defined in the caption of Figure 5), and so we can assume, without loss of generality, that either $\varepsilon_n = 0$ or $\overline{\text{emp}}_n = 0$ at every trial $n$, i.e., that

$$\overline{\text{unc}}_n = \mathbf{S}(\overline{\text{err}}_n), \quad \overline{\text{emp}}_n = \mathbf{C}(\overline{\text{err}}_n)$$

at every trial.

Let us check that to achieve (16) the region predictor must satisfy

$$\delta < \delta_0 \Longrightarrow \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\overline{\text{err}}_i - \delta_0)^+ = 0 \tag{20}$$

$$\delta \geq \delta_0 \Longrightarrow \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\delta_0 - \overline{\text{err}}_i)^+ = 0, \tag{21}$$

where the convergence is, as usual, almost certain. It was shown in Vovk (2002b) (Lemma 2) that the success curve $\mathbf{S}$ is convex, non-increasing, continuous, and has slope at most $-1$ before it hits

[3] More formally, predictions are certain for new extended objects $(x,\tau)$ satisfying

$$F(x,\tau) := F(f(x)-) + \tau(F(f(x)+) - F(f(x)-)) \geq \mathbf{S}(\overline{\text{err}}_n - \overline{\text{emp}}_n).$$

Intuitively, considering extended objects makes the vertical axis "infinitely divisible".

[4] Indeed, predictions of this kind are admissible in the sense that we cannot improve $\overline{\text{unc}}_n$ and $\overline{\text{emp}}_n$ simultaneously, and all admissible predictions are equivalent to predictions of this kind. A formal argument for the case where $\text{emp}_n$ are omitted is given in Vovk (2002b).

the $x$ axis at $\delta = \delta_0$. The second implication, (21), now immediately follows from the fact that, under $\delta \geq \delta_0$ and (16),

$$0 = \limsup_{n\to\infty} \frac{\overline{\mathrm{Unc}}_n}{n} = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{S}(\overline{\mathrm{err}}_i) \geq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (\delta_0 - \overline{\mathrm{err}}_i)^+ .$$

The first implication, (20), can be extracted from the chain

$$\frac{\overline{\mathrm{Unc}}_n}{n} = \frac{1}{n}\sum_{i=1}^{n} \overline{\mathrm{unc}}_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{S}(\overline{\mathrm{err}}_i) \geq \mathbf{S}\left(\frac{1}{n}\sum_{i=1}^{n} \overline{\mathrm{err}}_i\right) = \mathbf{S}\left(\frac{\overline{\mathrm{Err}}_n}{n}\right) \geq \mathbf{S}(\delta) - \varepsilon \qquad (22)$$

(with the last inequality holding almost surely for an arbitrary $\varepsilon > 0$ from some $n$ on) used by Vovk (2002b, in the proof of Theorems 1 and 1r). Indeed, it can be seen from (22) that, assuming the predictor is well-calibrated and optimal and $\delta < \delta_0$,

$$\overline{\mathrm{Err}}_n / n \to \delta \quad \text{a.s.}$$

and, therefore,

$$\begin{aligned}
\mathbf{S}(\delta) \geq \limsup_{n\to\infty} \frac{\overline{\mathrm{Unc}}_n}{n} &= \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{S}(\overline{\mathrm{err}}_i) = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{S}(\overline{\mathrm{err}}_i \wedge \delta_0) \\
&\geq \limsup_{n\to\infty} \mathbf{S}\left(\frac{1}{n}\sum_{i=1}^{n} (\overline{\mathrm{err}}_i \wedge \delta_0)\right) = \limsup_{n\to\infty} \mathbf{S}\left(\frac{\overline{\mathrm{Err}}_n}{n} - \frac{1}{n}\sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \delta_0)^+\right) \\
&= \limsup_{n\to\infty} \mathbf{S}\left(\delta - \frac{1}{n}\sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \delta_0)^+\right) = \mathbf{S}\left(\delta - \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \delta_0)^+\right)
\end{aligned}$$

almost surely. This proves (20).

Using (20), (21), and the fact that the complementary success curve $\mathbf{C}$ is concave, increasing, and (uniformly) continuous for $\delta \geq \delta_0$ (see Lemma 7), we obtain: if $\delta < \delta_0$,

$$\begin{aligned}
\frac{\overline{\mathrm{Emp}}_n}{n} &= \frac{1}{n}\sum_{i=1}^{n} \overline{\mathrm{emp}}_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}(\overline{\mathrm{err}}_i) \\
&\leq \frac{1}{n} \mathbf{C}'_{\mathrm{right}}(\delta_0) \sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \delta_0)^+ \to 0 \quad (n \to \infty);
\end{aligned}$$

if $\delta \geq \delta_0$,

$$\begin{aligned}
\frac{\overline{\mathrm{Emp}}_n}{n} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}(\overline{\mathrm{err}}_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}(\overline{\mathrm{err}}_i \vee \delta_0) \\
&\leq \mathbf{C}\left(\frac{1}{n}\sum_{i=1}^{n} (\overline{\mathrm{err}}_i \vee \delta_0)\right) = \mathbf{C}\left(\frac{1}{n}\sum_{i=1}^{n} \overline{\mathrm{err}}_i + \frac{1}{n}\sum_{i=1}^{n} (\delta_0 - \overline{\mathrm{err}}_i)^+\right) \\
&\leq \mathbf{C}\left(\frac{1}{n}\sum_{i=1}^{n} \overline{\mathrm{err}}_i\right) + o(1) \leq \mathbf{C}(\delta) + \varepsilon,
\end{aligned}$$

the last inequality holding almost surely for an arbitrary $\varepsilon > 0$ from some $n$ on and $\delta$ being the significance level used.

### 5.3 Proof Sketch of Proposition 6

Let us first modify and extend the notation $P_n^{\neq}(y\,|\,x_i,\sigma_i)$ introduced in (8). Consider the sequence of extended examples $w_i = (x_i,\tau'_i,y_i)$, $i = 1,\ldots,n$ ($(x_i,y_i)$ are the first $n$ examples chosen by Reality and $\tau_i$ are the random numbers used by Predictor). We define the Nearest Neighbours approximations $P_n(y\,|\,x,\sigma)$ to the conditional probabilities $P(y\,|\,x)$ as follows: for every $(x,\sigma,y) \in \tilde{\mathbf{Z}}$,

$$P_n(y\,|\,x,\sigma) := N(x,\sigma,y)/K_n, \tag{23}$$

where $N(x,\sigma,y)$ is the number of $i = 1,\ldots,n$ such that $(x_i,\tau'_i)$ is among the $K_n$ nearest neighbours of $(x,\sigma)$ and $y_i = y$ (this time $(x_i,\tau'_i)$ is not prevented from being counted as one of the $K_n$ nearest neighbours of $(x,\sigma)$ if $(x_i,\tau'_i) = (x,\sigma)$). We define the empirical predictability function $f_n$ by

$$f_n(x,\sigma) := \max_{y\in\mathbf{Y}} P_n(y\,|\,x,\sigma). \tag{24}$$

The proof will be based on the following version of a well-known fundamental result.

**Lemma 8** *Suppose $K_n \to \infty$, $K_n = o(n)$, and $\mathbf{Y} = \{0,1\}$. For any $\epsilon > 0$ and large enough $n$,*

$$\mathbb{P}\left\{\int |P(1\,|\,x) - P_n(1\,|\,x,\sigma)|\,P_{\mathbf{X}}(dx)\mathbf{U}(d\sigma) > \epsilon\right\} \le e^{-n\epsilon^2/40},$$

*where the outermost probability distribution $\mathbb{P}$ (essentially $(P\times\mathbf{U})^\infty$) generates the extended examples $(x_i,\tau_i,y_i)$, which determine the empirical distributions $P_n$.*

**Proof** This is almost a special case of Devroye et al.'s (1994) Theorem 1. There is, however, an important difference between the way we break distance ties and the way Devroye et al. (1994) do this. In that work, instead of our (3),

$$(|x_1 - x_3|, |\sigma_1 - \sigma_3|) < (|x_2 - x_3|, |\sigma_2 - \sigma_3|)$$

is used. (Our way of breaking ties better agrees with the lexicographic order on $[0,1]^2$, which is useful in the proof of Proposition 3 and, less importantly, in the proof of Lemma 10.) It is easy to check that the proof given by Devroye et al. (1994) also works (and becomes simpler) for our way of breaking distance ties. ■

**Lemma 9** *Suppose $K_n \to \infty$ and $K_n = o(n)$. For any $\epsilon > 0$ there exists an $\epsilon^* > 0$ such that, for large enough $n$,*

$$\mathbb{P}\left\{(P_{\mathbf{X}}\times\mathbf{U})\left\{(x,\sigma) : \max_{y\in\mathbf{Y}} |P_n(y\,|\,x,\sigma) - P(y\,|\,x)| > \epsilon\right\} > \epsilon\right\} \le e^{-\epsilon^* n};$$

*in particular,*

$$\mathbb{P}\{(P_{\mathbf{X}}\times\mathbf{U})\{(x,\sigma) : |f_n(x,\sigma) - f(x)| > \epsilon\} > \epsilon\} \le e^{-\epsilon^* n}.$$

**Proof** We apply Lemma 8 to the binary classification problem obtained from our classification problem by replacing label $y \in \mathbf{Y}$ with 1 and replacing all other labels with 0:

$$\mathbb{P}\left\{\int |P(y\,|\,x) - P_n(y\,|\,x,\sigma)|\,P_{\mathbf{X}}(dx)\mathbf{U}(d\sigma) > \epsilon\right\} \le e^{-n\epsilon^2/40}.$$

By Markov's inequality this implies

$$\mathbb{P}\left\{(P_{\mathbf{X}} \times \mathbf{U})\{|P(y\,|\,x) - P_n(y\,|\,x,\sigma)| > \sqrt{\varepsilon}\} > \sqrt{\varepsilon}\right\} \le e^{-n\varepsilon^2/40},$$

which, in turn, implies

$$\mathbb{P}\left\{(P_{\mathbf{X}} \times \mathbf{U})\left\{\max_{y\in\mathbf{Y}} |P(y\,|\,x) - P_n(y\,|\,x,\sigma)| > \sqrt{\varepsilon}\right\} > |\mathbf{Y}|\sqrt{\varepsilon}\right\} \le e^{-n\varepsilon^2/40}.$$

This completes the proof, since we can take the ε in the last equation arbitrarily small as compared to the ε in the statement of the lemma. ■

We will use the shorthand "$\forall^\infty n$" for "from some $n$ on".

**Lemma 10** *Suppose $K_n \to \infty$ and $K_n = o(n)$. For any $\varepsilon > 0$ there exists an $\varepsilon^* > 0$ such that, for large enough $n$,*

$$\mathbb{P}\left\{\frac{\#\left\{i : \max_y \left|P(y\,|\,x_i) - P_n^{\neq}(y\,|\,x_i,\tau_i')\right| > \varepsilon\right\}}{n} > \varepsilon\right\} \le e^{-\varepsilon^* n}.$$

*In particular,*

$$\forall^\infty n : \mathbb{P}\left\{\frac{\#\left\{i : \left|f(x_i) - f_n^{\neq}(x_i,\tau_i')\right| > \varepsilon\right\}}{n} > \varepsilon\right\} \le e^{-\varepsilon^* n}.$$

**Proof** Since

$$\left|P_n^{\neq}(y\,|\,x_i,\tau_i') - P_n(y\,|\,x_i,\tau_i')\right| \le \frac{1}{K_n} = o(1),$$

we can, and will, ignore the upper indices $\neq$ in the statement of the lemma.

Define

$$I_n(x,\sigma) := \begin{cases} 0 & \text{if } \max_y |P(y\,|\,x) - P_n(y\,|\,x,\sigma)| \le \varepsilon \\ 1 & \text{if } \max_y |P(y\,|\,x) - P_n(y\,|\,x,\sigma)| \ge 2\varepsilon \\ (\max_y |P(y\,|\,x) - P_n(y\,|\,x,\sigma)| - \varepsilon)/\varepsilon & \text{otherwise} \end{cases}$$

(intuitively, $I_n(x,\sigma)$ is a "soft version" of $\mathbb{I}_{\{\max_y |P(y\,|\,x) - P_n(y\,|\,x,\sigma)| > \varepsilon\}}$).

The main tool in this proof (and several other proofs in this section) will be McDiarmid's theorem (see, e.g., Devroye et al., 1996, Theorem 9.2). First we check the possibility of its application. If we replace an extended object $(x_j,\tau_j')$ by another extended object $(x_j^*,\tau_j^*)$, the expression

$$\sum_{i=1}^n I_n(x_i,\tau_i')$$

will change as follows:

- the addend $I_n(x_i,\tau_i')$ for $i = j$ changes by 1 at most;
- the addends $I_n(x_i,\tau_i')$ for $i \neq j$ such that neither $(x_j,\tau_j')$ nor $(x_j^*,\tau_j^*)$ are among the $K_n$ nearest neighbours of $(x_i,\tau_i')$ do not change at all;

- the sum over the at most $4K_n$ (see below) addends $I_n(x_i,\tau'_i)$ for $i \neq j$ such that either $(x_j,\tau'_j)$ or $(x^*_j,\tau^*_j)$ (or both) are among the $K_n$ nearest neighbours of $(x_i,\tau'_i)$ can change by at most

$$4K_n \frac{1}{\varepsilon}\frac{1}{K_n} = \frac{4}{\varepsilon}. \tag{25}$$

The left-hand side of (25) reflects the following facts: the change in $P_n(y\,|\,x_i,\tau'_i)$ for $i \neq j$ is at most $1/K_n$; the number of $i \neq j$ such that $(x_j,\tau'_j)$ is among the $K_n$ nearest neighbours of $(x_i,\tau'_i)$ does not exceed $2K_n$ (since the extended objects are linearly ordered and (3) is used for breaking distance ties); analogously, the number of $i \neq j$ such that $(x^*_j,\tau^*_j)$ is among the $K_n$ nearest neighbours of $(x_i,\tau'_i)$ does not exceed $2K_n$.

Therefore, by McDiarmid's theorem,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n I_n(x_i,\tau'_i) - \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n I_n(x_i,\tau'_i)\right) > \varepsilon\right\} \\ \leq \exp\left(-2\varepsilon^2 n/\left(1+4/\varepsilon\right)^2\right) = \exp\left(-\frac{2\varepsilon^4}{(4+\varepsilon)^2}n\right). \tag{26}$$

Next we find:

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n I_n(x_i,\tau'_i)\right) = \mathbb{E}\left(I_n(x_n,\tau'_n)\right) \leq \mathbb{E}\left(I_{n-1}(x_n,\tau'_n)\right) + o(1) \\
&\leq \mathbb{E}(P_{\mathbf{X}} \times \mathbf{U})\{(x,\sigma) : \max_y |P(y\,|\,x) - P_{n-1}(y\,|\,x,\sigma)| > \varepsilon\} + o(1) \\
&\leq e^{-\varepsilon^* n} + \varepsilon + o(1) \leq 2\varepsilon
\end{aligned}$$

(the penultimate inequality follows from Lemma 9) from some $n$ on. In combination with (26) this implies

$$\forall^\infty n : \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n I_n(x_i,\tau'_i) > 3\varepsilon\right\} \leq \exp\left(-\frac{2\varepsilon^4}{(4+\varepsilon)^2}n\right),$$

in particular

$$\mathbb{P}\left\{\frac{\#\{i : \max_y |P(y\,|\,x_i) - P_n(y\,|\,x_i,\tau'_i)| \geq 2\varepsilon\}}{n} > 3\varepsilon\right\} \leq \exp\left(-\frac{2\varepsilon^4}{(4+\varepsilon)^2}n\right).$$

Replacing $3\varepsilon$ by $\varepsilon$, we obtain that, from some $n$ on,

$$\mathbb{P}\left\{\frac{\#\{i : \max_y |P(y\,|\,x_i) - P_n(y\,|\,x_i,\tau'_i)| > \varepsilon\}}{n} > \varepsilon\right\} \leq \exp\left(-\frac{2(\varepsilon/3)^4}{(4+\varepsilon/3)^2}n\right),$$

which completes the proof. ■

We say that an extended example $(x_i,\tau_i,y_i)$, $i = 1,\ldots,n$, is *n-strange* if $y_i \neq \hat{y}_n(x_i,\tau'_i)$; otherwise, $(x_i,\tau_i,y_i)$ will be called *n-ordinary*. We will assume that $(f^{\neq}_n(x_i,\tau'_i),\tau''_i)$, $i = 1,\ldots,n$, are all different for all $n$; even more than that, we will assume that $\tau''_i$ are all different (we can do so since the probability of this event is one).

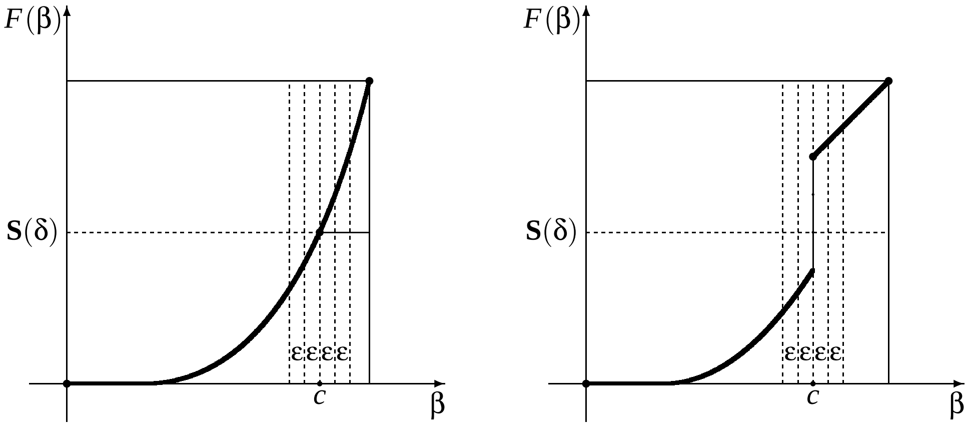


Figure 6: Cases $F(c) = \mathbf{S}(\delta)$ (left) and $F(c) > \mathbf{S}(\delta)$ (right). The vertical bands of width ε determine the division of the first $n$ extended examples into five classes

**Lemma 11** *Suppose (7) is satisfied and $\delta \le \delta_0$. With probability one, the $\lfloor (1-\mathbf{S}(\delta))n \rfloor$ extended examples with the largest (in the sense of the lexicographic order) $(f_n^{\neq}(x_i,\tau_i'),\tau_i'')$ among $(x_1,\tau_1,y_1),\dots,(x_n,\tau_n,y_n)$ contain at most $n\delta + o(n)$ n-strange extended examples as $n \to \infty$.*

**Proof** Define

$$c := \sup\{\beta : F(\beta) \le \mathbf{S}(\delta)\}.$$

It is clear that $0 < c < 1$. Our proof will work both in the case where $F(c) = \mathbf{S}(\delta)$ and in the case where $F(c) > \mathbf{S}(\delta)$, as illustrated in Figure 6.

Let $\varepsilon > 0$ be a small constant (we will let $\varepsilon \to 0$ eventually). Define a “threshold” $(c_n', c_n'') \in [0,1]^2$ requiring that

$$\mathbb{P}\left\{ f(x_n) = c, (f_{n-1}(x_n,\tau_n'),\tau_n'') > (c_n',c_n'') \right\} = F(c) - \mathbf{S}(\delta) - \varepsilon \tag{27}$$

if $F(c) > \mathbf{S}(\delta)$; we assume that ε is small enough for

$$2\varepsilon < F(c) - \mathbf{S}(\delta) \tag{28}$$

to hold . Among other things this will ensure the validity of the definition (27). If $F(c) = \mathbf{S}(\delta)$, we set $(c_n',c_n'') := (c+\varepsilon, 0)$; in any case, we will have

$$\mathbb{P}\left\{ f(x_n) = c, (f_{n-1}(x_n,\tau_n'),\tau_n'') > (c_n',c_n'') \right\} \ge F(c) - \mathbf{S}(\delta) - \varepsilon. \tag{29}$$

Let us say that an extended example $(x_i,\tau_i,y_i)$ is *above the threshold* if

$$(f_n^{\neq}(x_i,\tau_i'),\tau_i'') > (c_n',c_n'');$$

otherwise, we say it is *below the threshold*. Divide the first $n$ extended examples $(x_i,\tau_i,y_i)$, $i = 1,\dots,n$, into five classes:

**Class I:** Those satisfying $f(x_i) \le c - 2\varepsilon$.

**Class II:** Those that satisfy $f(x_i) = c$ and are below the threshold.

**Class III:** Those satisfying $c - 2\varepsilon < f(x_i) \le c + 2\varepsilon$ but not $f(x_i) = c$.

**Class IV:** Those that satisfy $f(x_i) = c$ and are above the threshold.

**Class V:** Those satisfying $f(x_i) > c + 2\varepsilon$.

First we explain the general idea of the proof. The threshold $(c', c'')$ was chosen so that approximately $\lfloor (1 - \mathbf{S}(\delta))n \rfloor$ of the available extended examples will be above the threshold. Because of this, the extended examples above the threshold will essentially be the $\lfloor (1 - \mathbf{S}(\delta))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ referred to in the statement of the lemma. For each of the five classes we will be interested in the following questions:

- How many extended examples are there in the class?
- How many of those are above the threshold?
- How many of those above the threshold are $n$-strange?

If the sum of the answers to the last question does not exceed $n\delta$ by too much, we are done.

With this plan in mind, we start the formal proof. (Of course, we will not be following the plan literally: for example, if a class is very small, we do not need to answer the second and third questions.) The first step is to show that

$$c - \varepsilon \le c_n' \le c + \varepsilon \tag{30}$$

from some $n$ on; this will ensure that the classes are conveniently separated from each other. We only need to consider the case $F(c) > \mathbf{S}(\delta)$. The inequality $c_n' \le c + \varepsilon$ follows from

$$\forall^\infty n : \mathbb{P}\left\{ f(x_n) = c, f_{n-1}(x_n, \tau_n') > c + \varepsilon \right\} < \varepsilon < F(c) - \mathbf{S}(\delta) - \varepsilon$$

Simply combine Lemma 9 with (28). The inequality $c - \varepsilon \le c_n'$ follows in a similar way from

$$\begin{aligned}\forall^\infty n : \quad & \mathbb{P}\left\{ f(x_n) = c, f_{n-1}(x_n, \tau_n') \ge c - \varepsilon \right\} \\ & = \mathbb{P}\{ f(x_n) = c \} - \mathbb{P}\left\{ f(x_n) = c, f_{n-1}(x_n, \tau_n') < c - \varepsilon \right\} \\ & > F(c) - F(c-) - \varepsilon \ge F(c) - \mathbf{S}(\delta) - \varepsilon.\end{aligned}$$

Now we are ready to analyze the composition of our five classes. Among the Class I extended examples at most

$$\varepsilon n \tag{31}$$

will be above the threshold from some $n$ on almost surely (by Lemma 10 and the Borel–Cantelli lemma). None of the Class II extended examples will be above the threshold, by definition. The fraction of Class III extended examples among the first $n$ extended examples will tend to

$$F(c + 2\varepsilon) - F(c) + F(c-) - F(c - 2\varepsilon) \tag{32}$$

as $n \to \infty$ almost surely.

To estimate the number $N_n^{\mathrm{IV}}$ of Class IV extended examples among the first $n$ extended examples, we use McDiarmid's theorem. If one extended example is replaced by another, $N_n^{\mathrm{IV}}$ will change by at most $2K_n+1$ (since this extended example can affect $f_n^{\neq}(x_i,\tau_i')$ for at most $2K_n$ other extended examples $(x_i,\tau_i,y_i)$). Therefore,

$$\mathbb{P}\left\{\left|\frac{1}{n}N_n^{\mathrm{IV}}-\frac{1}{n}\mathbb{E}N_n^{\mathrm{IV}}\right|\geq\varepsilon\right\}\leq 2e^{-2\varepsilon^2 n/(2K_n+1)^2};$$

the assumption $K_n=o\left(\sqrt{n/\ln n}\right)$ and the Borel–Cantelli lemma imply that

$$\left|\frac{1}{n}N_n^{\mathrm{IV}}-\frac{1}{n}\mathbb{E}N_n^{\mathrm{IV}}\right|<\varepsilon$$

from some $n$ on almost surely. Since

$$\frac{1}{n}\mathbb{E}N_n^{\mathrm{IV}}=\mathbb{P}\left\{f(x_n)=c,(f_{n-1}(x_n,\tau_n'),\tau_n'')>(c_n',c_n'')\right\}\geq F(c)-\mathbf{S}(\delta)-\varepsilon,$$

as in (29), we have

$$N_n^{\mathrm{IV}}>(F(c)-\mathbf{S}(\delta)-2\varepsilon)n \tag{33}$$

from some $n$ on almost surely. Of course, all these examples are above the threshold.

Now we estimate the number $N_n^{\mathrm{IV,str}}$ of $n$-strange extended examples of Class IV. Again McDiarmid's theorem implies that

$$\left|\frac{1}{n}N_n^{\mathrm{IV,str}}-\frac{1}{n}\mathbb{E}N_n^{\mathrm{IV,str}}\right|<\varepsilon$$

from some $n$ on almost surely. Now, from some $n$ on,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}N_n^{\mathrm{IV,str}}&=\mathbb{P}\left\{f(x_n)=c,(f_{n-1}(x_n,\tau_n'),\tau_n'')>(c_n',c_n''),\hat{y}_n(x_n,\tau_n')\neq y_n\right\}\\
&=\mathbb{E}\left(\left(1-P_{\mathbf{Y}\mid\mathbf{X}}\left(\hat{y}_n(x_n,\tau_n')\mid x_n\right)\right)\mathbb{I}_{\{f(x_n)=c,(f_{n-1}(x_n,\tau_n'),\tau_n'')>(c_n',c_n'')\}}\right)\\
&\leq e^{-\varepsilon^* n}+\varepsilon+(1-c+2\varepsilon)\\
&\quad\times\mathbb{P}\{f(x_n)=c,(f_{n-1}(x_n,\tau_n'),\tau_n'')>(c_n',c_n'')\}\\
&=e^{-\varepsilon^* n}+\varepsilon+(1-c+2\varepsilon)(F(c)-\mathbf{S}(\delta)-\varepsilon) &(34)\\
&\leq(F(c)-\mathbf{S}(\delta))(1-c)+4\varepsilon &(35)
\end{aligned}$$

in the case $F(c)>\mathbf{S}(\delta)$; the first inequality in this chain follows from Lemma 9: indeed, this lemma implies that, unless an event of the small probability $e^{-\varepsilon^* n}+\varepsilon$ happens,

$$P\left(\hat{y}_n(x_n,\tau_n')\mid x_n\right)\geq P_{n-1}\left(\hat{y}_n(x_n,\tau_n')\mid x_n,\tau_n'\right)-\varepsilon=f_{n-1}\left(x_n,\tau_n'\right)-\varepsilon\geq f(x_n)-2\varepsilon. \tag{36}$$

If $F(c)=\mathbf{S}(\delta)$, the lines (34) and (35) of that chain have to be changed to

$$\begin{aligned}
&\leq e^{-\varepsilon^* n}+\varepsilon+(1-c+2\varepsilon)\mathbb{P}\{f(x_n)=c,\,f_{n-1}(x_n,\tau_n')\geq c+\varepsilon\}\\
&\leq e^{-\varepsilon^* n}+\varepsilon+(1-c+2\varepsilon)\left(e^{-\varepsilon^* n}+\varepsilon\right)<4\varepsilon
\end{aligned}$$

(where the obvious modification of Lemma 9 with all "$> \epsilon$" changed to "$\geq \epsilon$" is used), but the inequality between the extreme terms of the chain still holds. Therefore, the number of $n$-strange Class IV extended examples does not exceed

$$((F(c) - \mathbf{S}(\delta))(1-c) + 5\epsilon)\, n \tag{37}$$

from some $n$ on almost surely.

By the Borel strong law of large numbers, the fraction of Class V extended examples among the first $n$ extended examples will tend to

$$1 - F(c + 2\epsilon) \tag{38}$$

as $n \to \infty$ almost surely. By Lemma 10, the Borel–Cantelli lemma, and (30), almost surely from some $n$ on at least

$$(1 - F(c+2\epsilon) - 2\epsilon)n \tag{39}$$

extended examples in Class V will be above the threshold.

Finally, we estimate the number $N_n^{\mathrm{V,str}}$ of $n$-strange extended examples of Class V among the first $n$ extended examples. By McDiarmid's theorem,

$$\left| \frac{1}{n} N_n^{\mathrm{V,str}} - \frac{1}{n} \mathbb{E} N_n^{\mathrm{V,str}} \right| < \epsilon$$

from some $n$ on almost surely. Now

$$\begin{aligned}
\frac{1}{n} \mathbb{E} N_n^{\mathrm{V,str}} &= \mathbb{P}\left\{ f(x_n) > c + 2\epsilon, \hat{y}_n(x_n, \tau_n') \neq y_n \right\} \\
&= \mathbb{E}\left( \left(1 - P_{\mathbf{Y}\,|\,\mathbf{X}}\left( \hat{y}_n(x_n, \tau_n') \,|\, x_n \right)\right) \mathbb{I}_{\{f(x_n) > c+2\epsilon\}} \right) \\
&\leq e^{-\epsilon^* n} + \epsilon + \mathbb{E}\left( (1 - f(x_n) + 2\epsilon)\, \mathbb{I}_{\{f(x_n) > c+2\epsilon\}} \right) \\
&\leq e^{-\epsilon^* n} + 3\epsilon + \mathbb{E}\left( (1 - f(x_n))\, \mathbb{I}_{\{f(x_n) > c+2\epsilon\}} \right) \\
&= e^{-\epsilon^* n} + 3\epsilon + \int_0^1 (F(\beta) - F(c+2\epsilon))^+ d\beta \\
&< \int_0^1 (F(\beta) - F(c))^+ d\beta + 4\epsilon
\end{aligned}$$

from some $n$ on. The first inequality follows from Lemma 9, as in (36). Therefore,

$$\frac{1}{n} N_n^{\mathrm{V,str}} < \int_0^1 (F(\beta) - F(c))^+ d\beta + 5\epsilon \tag{40}$$

from some $n$ on almost surely.

Summarizing, we can see that the total number of extended examples above the threshold among the first $n$ extended examples will be at least

$$(F(c) - \mathbf{S}(\delta) - 2\epsilon + 1 - F(c+2\epsilon) - 2\epsilon)\, n = (1 - \mathbf{S}(\delta) + F(c) - F(c+2\epsilon) - 4\epsilon)\, n \tag{41}$$

(see (33) and (39)) from some $n$ on almost surely. The number of $n$-strange extended examples among them will not exceed

$$\begin{aligned}\Big(&\epsilon + F(c+2\epsilon) - F(c) + F(c-) - F(c-2\epsilon) + \epsilon \\ &\quad + (F(c) - \mathbf{S}(\delta))(1-c) + 5\epsilon + \int_0^1 (F(\beta) - F(c))^+ d\beta + 5\epsilon\Big) n \\ &= \Big(F(c+2\epsilon) - F(c) + F(c-) - F(c-2\epsilon) \\ &\quad + (F(c) - \mathbf{S}(\delta))(1-c) + \int_0^1 (F(\beta) - F(c))^+ d\beta + 12\epsilon\Big) n \qquad (42)\end{aligned}$$

(see (31), (32), (37), and (40)) from some $n$ on almost surely. Combining (41) and (42), we can see that the number of $n$-strange extended examples among the $\lfloor (1-\mathbf{S}(\delta))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ does not exceed

$$\begin{aligned}\Big(&F(c+2\epsilon) - F(c) + F(c-) - F(c-2\epsilon) + (F(c) - \mathbf{S}(\delta))(1-c) \\ &\quad + \int_0^1 (F(\beta) - F(c))^+ d\beta + 12\epsilon\Big) n + (F(c+2\epsilon) - F(c) + 4\epsilon)\, n \\ &= \Big(2(F(c+2\epsilon) - F(c)) + (F(c-) - F(c-2\epsilon)) + (F(c) - \mathbf{S}(\delta))(1-c) \\ &\quad + \int_0^1 (F(\beta) - F(c))^+ d\beta + 16\epsilon\Big) n\end{aligned}$$

from some $n$ on almost surely. Since $\epsilon$ can be arbitrarily small, the coefficient in front of $n$ in the last expression can be made arbitrarily close to

$$(F(c) - \mathbf{S}(\delta))(1-c) + \int_0^1 (F(\beta) - F(c))^+ d\beta = \int_0^1 (F(\beta) - \mathbf{S}(\delta))^+ d\beta = \delta,$$

which completes the proof. ■

**Lemma 12** *Suppose (7) is satisfied. The fraction of $n$-strange extended examples among the first $n$ extended examples $(x_i, \tau_i, y_i)$ approaches $\delta_0$ asymptotically with probability one.*

**Proof sketch** The lemma is not difficult to prove using McDiarmid's theorem and the fact that, by Lemma 10, $P(\hat{y}_n(x_i, \tau_i') \,|\, x_i)$ will typically differ little from $f(x_i)$. Notice, however, that the part that we really need in this paper (that the fraction of $n$-strange extended examples does not exceed $\delta_0 + o(1)$ as $n \to \infty$ with probability one) is just a special case of Lemma 11, corresponding to $\delta = \delta_0$. ■

**Lemma 13** *Suppose (7) is satisfied and $\delta > \delta_0$. The fraction of $n$-ordinary extended examples among the $\lfloor \mathbf{C}(\delta) n \rfloor$ extended examples $(x_i, \tau_i, y_i)$, $i = 1, \dots, n$, with the lowest $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ does not exceed $\delta - \delta_0 + o(1)$ as $n \to \infty$ with probability one.*

Lemma 13 can be proved analogously to Lemma 11.

**Lemma 14** *Let $\mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \cdots$ be a decreasing sequence of σ-algebras and $\xi_1, \xi_2 \dots$ be a bounded adapted (in the sense that $\xi_n$ is $\mathcal{F}_n$-measurable for all n) sequence of random variables such that*

$$\limsup_{n\to\infty} \mathbb{E}(\xi_n \mid \mathcal{F}_{n+1}) \le 0 \quad \textit{a.s.}$$

*Then*

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \xi_i \le 0 \quad \textit{a.s.}$$

**Proof** Replacing, if necessary, $\xi_n$ by $\xi_n - \mathbb{E}(\xi_n \mid \mathcal{F}_{n+1})$, we reduce our task to the following special case (a reverse Borel strong law of large numbers): if $\xi_1, \xi_2, \dots$ is a bounded *reverse martingale difference*, in the sense of being adapted and satisfying $\forall n : \mathbb{E}(\xi_n \mid \mathcal{F}_{n+1}) = 0$, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \xi_i = 0 \quad \text{a.s.} \tag{43}$$

Fix a bounded reverse martingale difference $\xi_1, \xi_2, \dots$; our goal is to prove (43). By the martingale version of Hoeffding's inequality (Devroye et al., 1996, Theorem 9.1) applied to the martingale difference $(\xi_i, \mathcal{F}_i)$, $i = n, \dots, 1$,

$$\mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} \xi_i \right| \ge \varepsilon \right\} \le 2e^{-2\varepsilon^2 n/(2C)^2}, \tag{44}$$

where $C$ is an upper bound on $\sup_n |\xi_n|$. Combined with the Borel–Cantelli–Lévy lemma, (44) implies (43). ■

Now we can sketch the proof of Proposition 6. Define $\mathcal{F}_n$, $n = 1, 2, \dots$, to be the σ-algebra on $\tilde{\mathbf{Z}}^\infty$ generated by the multiset of the first $n-1$ extended examples $(x_i, \tau_i, y_i)$, $i = 1, \dots, n-1$, and the sequence of extended examples $(x_i, \tau_i, y_i)$, $i = n, n+1, \dots$ (starting from the $n$th extended example).

Suppose first that $\delta < \delta_0$. Consider the $\lfloor (1 - \mathbf{S}(\delta - \varepsilon))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ among $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$, where $\varepsilon \in (0, \delta)$ is a small constant. Let us show that each of these examples will be predicted with certainty from the other extended examples in the sequence $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$, from some $n$ on. We will be assuming $n$ large enough.

Let $(x_k, \tau_k, y_k)$ be the extended example with the $(\lfloor (\delta - \varepsilon/2)n \rfloor + 1)$th largest (in the sense of the lexicographic order) $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ among all $n$-strange extended examples $(x_i, \tau_i, y_i)$, $i = 1, \dots, n$. (Remember that all $\tau_i''$ are assumed to be different.) Let $(x_j, \tau_j, y_j)$ be one of the $\lfloor (1 - \mathbf{S}(\delta - \varepsilon))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ and let $y \in \mathbf{Y}$ be a label different from $\hat{y}_n(x_j, \tau_j')$. It suffices to prove that

$$\tau_j'' \ge \frac{\#\{i = 1, \dots, n : \alpha_i^y \ge \alpha_j^y\} - n\delta}{\#\{i = 1, \dots, n : \alpha_i^y = \alpha_j^y\}} \tag{45}$$

(cf. (6) on p. 582), where all $\alpha^y$ are computed as $\alpha$ in (11) from the sequence

$$(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$$

with $y_j$ replaced by $y$. It will be more convenient to write (45) in the form

$$\#\{i:\alpha_i^y > \alpha_j^y\} + (1-\tau_j'')\#\{i:\alpha_i^y = \alpha_j^y\} \le n\delta.$$

Since $\alpha_j^y = f_n^{\neq}(x_j,\tau_j')$ and $\alpha_i^y \neq \alpha_i$ for at most $2K_n+1$ values of $i$ (indeed, changing $y_j$ will affect at most $2K_n+1$ $\alpha$s), it suffices to prove

$$\#\{i:\alpha_i > f_n^{\neq}(x_j,\tau_j')\} + (1-\tau_j'')\#\{i:\alpha_i = f_n^{\neq}(x_j,\tau_j')\} \le n(\delta-\varepsilon^*), \tag{46}$$

where $\varepsilon^* \ll \varepsilon$ is a positive constant.

Since $(f_n^{\neq}(x_j,\tau_j'),\tau_j'') \ge (\alpha_k,\tau_k'')$ (indeed, by Lemma 11, there are less than $(\delta-\varepsilon/2)n$ $n$-strange extended examples among the $\lfloor(1-\mathbf{S}(\delta-\varepsilon))n\rfloor$ extended examples with the largest $(f_n^{\neq}(x_i,\tau_i'),\tau_i'')$), (46) will follow from

$$\#\{i:\alpha_i > \alpha_k\} + (1-\tau_k'')\#\{i:\alpha_i = \alpha_k\} \le n(\delta-\varepsilon^*). \tag{47}$$

If $\#\{i:\alpha_i = \alpha_k\} \le \frac{\varepsilon}{3}n$, the left-hand side of (47) does not exceed

$$\left(\delta-\frac{\varepsilon}{2}\right)n + \frac{\varepsilon}{3}n < n(\delta-\varepsilon^*),$$

so we can, and will, assume without loss of generality that

$$\#\{i:\alpha_i = \alpha_k\} > \frac{\varepsilon}{3}n. \tag{48}$$

Since $\tau_i''$ for the extended examples satisfying $\alpha_i = \alpha_k$ are output according to the uniform distribution $\mathbf{U}$, the expected value of $1-\tau_k''$ is about

$$\frac{(\delta-\varepsilon/2)n - \#\{i:\alpha_i > \alpha_k\}}{\#\{i:\alpha_i = \alpha_k\}},$$

and so by Hoeffding's inequality and the Borel–Cantelli lemma we will have (from some $n$ on)

$$1-\tau_k'' \le \frac{(\delta-\varepsilon/2)n - \#\{i:\alpha_i > \alpha_k\}}{\#\{i:\alpha_i = \alpha_k\}} + \varepsilon^*, \tag{49}$$

remembering (48). Equation (47) will hold because its left-hand side can be transformed using (49) as

$$\begin{aligned}\#\{i:\alpha_i > \alpha_k\} + (1-\tau_k'')\#\{i:\alpha_i = \alpha_k\} &\le (\delta-\varepsilon/2)n + \varepsilon^*\#\{i:\alpha_i = \alpha_k\}\\ &\le (\delta-\varepsilon/2+\varepsilon^*)n \le (\delta-\varepsilon^*)n.\end{aligned}$$

The assertion we have just proved means that, almost surely from some $n$ on,

$$\mathbb{P}(\{\mathrm{unc}_n = 0\} \mid \mathcal{F}_{n+1}) \ge \frac{\lfloor(1-\mathbf{S}(\delta-\varepsilon))n\rfloor}{n} \ge 1-\mathbf{S}(\delta-\varepsilon) - \frac{1}{n}.$$

Since $\varepsilon$ can be arbitrarily small and $\mathbf{S}$ is continuous (Vovk, 2002b, Lemma 2), this implies

$$\limsup_{n\to\infty} \mathbb{E}(\mathrm{unc}_n \mid \mathcal{F}_{n+1}) \le \mathbf{S}(\delta) \quad \text{a.s.}$$

By Lemma 14 this implies, in turn,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{unc}_i \le \mathbf{S}(\delta) \quad \text{a.s.},$$

which coincides with (17).

If $\delta \ge \delta_0$, Lemma 12 implies that

$$\lim_{n\to\infty} \mathbb{E}(\mathrm{unc}_n \mid \mathcal{F}_{n+1}) = 0 \quad \text{a.s.}$$

(and actually $\mathbb{E}(\mathrm{unc}_n \mid \mathcal{F}_{n+1}) = 0$ from some $n$ on if $\delta > \delta_0$); in combination with Lemma 14 this again implies (17).

Inequality (18) is treated in a similar way to (17). Lemmas 12 and 13 imply that

$$\liminf_{n\to\infty} \mathbb{E}(\mathrm{emp}_n \mid \mathcal{F}_{n+1}) \ge \mathbf{C}(\delta) \quad \text{a.s.} \tag{50}$$

(this inequality is vacuously true when $\delta \le \delta_0$). Another application of Lemma 14 gives

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{emp}_i \ge \mathbf{C}(\delta) \quad \text{a.s.},$$

i.e., (18).

**Remark** The derivation of Proposition 6 from Lemmas 11–14 would be very simple if we defined the individual strangeness measure by, say,

$$\alpha_i := \begin{cases} (-f_n^{\neq}(x_i,\sigma_i),\sigma_i) & \text{if } y_i = \hat{y}_n(x_i,\sigma_i) \\ (f_n^{\neq}(x_i,\sigma_i),\sigma_i) & \text{otherwise} \end{cases}$$

(with the lexicographic order on the $\alpha$'s) instead of (11) (in which case the denominator of (6) would be 1 almost surely). Our definition (11), however, is simpler and, most importantly, facilitates the proof of Proposition 3. Another simplification would be to use Lemma 11 (applied to $\delta := \delta - \mathbf{C}(\delta)$) instead of Lemma 13 in the derivation of (50); we preferred a more symmetric picture.

## 6. Conclusion

We have shown that there exist universal well-calibrated region predictors, thus satisfying, to some degree, the desiderata mentioned in §1: well-calibratedness and optimal performance. Notice, however, that the ways in which these two desiderata are satisfied are very different: the well-calibratedness holds in a very specific finitary sense, since the errors have probability $\delta$ and are independent, whereas the optimal performance is achieved only asymptotically.

An important direction of further research is to obtain non-asymptotic results about TCM's optimality. A natural setting is where we have a Bayesian model for Reality's strategy, $\{P_\theta : \theta \in \Theta\}$ with a prior $\mu(d\theta)$ on $\Theta$, and our goal is to minimize $\mathrm{Unc}_n^{\delta}$ under this model. The intuition behind this setting is that we do not really believe that the data is generated from our model and so prefer a predictor that is well-calibrated regardless the correctness of the model; but if the model is correct, we would like to have an optimal performance. A special case of this setting, with $\mu(d\theta)$

concentrated at one point, was considered in Vovk (2002b); however, all results in that paper are asymptotic.

## Acknowledgments

This is a full version of the conference paper (Vovk, 2003); I am grateful to both sets of anonymous referees (for the conference and journal versions), whose comments and suggestions helped to improve the quality of presentation and correct several mistakes. This work was partially supported by EPSRC (grant GR/R46670/01), BBSRC (grant 111/BIO14428), and EU (grant IST-1999-10226).

## Appendix A. Notation

The following table contains, strictly speaking, not only the notation used in this paper but also the preferred use of symbols.

| | |
|---|---|
| $\mathbf{X}$ | object space |
| $\mathbf{Y}$ | label space |
| $\mathbf{Z}$ | example space ($\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$) |
| $P$ | the probability distribution in $\mathbf{Z}$ generating individual examples $z_1 = (x_1, y_1), z_2 = (x_2, y_2), \ldots$ |
| $\delta$ | significance level |
| $\Gamma_n^\delta$ | prediction region |
| $\mathrm{err}_n^\delta$ | indicator of error at trial $n$ |
| $\mathrm{unc}_n^\delta$ | indicator of uncertain prediction at trial $n$ |
| $\mathrm{emp}_n^\delta$ | indicator of empty prediction at trial $n$ |
| $\mathrm{Err}_n^\delta$ | cumulative number of errors up to trial $n$ |
| $\mathrm{Unc}_n^\delta$ | cumulative number of uncertain predictions up to trial $n$ |
| $\mathrm{Emp}_n^\delta$ | cumulative number of empty predictions up to trial $n$ |
| $\tau_n$ | the $n$th random number used by a region predictor |
| $\tau'_n$, $\tau''_n$ | two components of $\tau_n$, as defined in §3.1 |
| $\tilde{\mathbf{X}}$ | the extended object space $\mathbf{X} \times [0,1]$ |
| $\tilde{\mathbf{Z}}$ | the extended example space $\mathbf{X} \times [0,1] \times \mathbf{Y}$ |
| $<, \leq$ | may refer to the lexicographic order on $[0,1]^2$, as defined on p. 581 |
| $\lvert x, \sigma \rvert$ | the absolute value of $(x, \sigma) \in [0,1]^2$, as defined in (4) |
| $A_n$ | individual strangeness measure |
| $\alpha_i$ | values taken by an individual strangeness measure |
| $\#E$ | the size of set $E$ |
| $K_n$ | the number of nearest neighbours taken into account at trial $n$ |
| $P_n^{\neq}(y \mid x_i, \sigma_i)$ | empirical estimate of $P(y \mid x_i)$ without taking $y_i$ into account, as defined in (8) |
| $f_n^{\neq}(x_i, \sigma_i)$ | corresponding empirical predictability function, (9) |
| $\hat{y}_n(x_i, \sigma_i)$ | “choice function”, as defined in (10) |
| $\Delta$ | finite set of significance levels |
| $P_{\mathbf{X}}$ | the marginal distribution of $P$ in $\mathbf{X}$ |

| | |
|---|---|
| $P_{\mathbf{Y}\,\vert\,\mathbf{X}}$ | the regular conditional distribution of $y \in \mathbf{Y}$ given $x \in \mathbf{X}$,<br>where $(x,y)$ is distributed as $P$ |
| $f(x)$ | predictability of object $x$ |
| $F(\beta)$ | predictability distribution function |
| $\mathbf{S}(\delta)$ | success curve, defined in (12) |
| $\mathbf{C}(\delta)$ | complementary success curve, defined in (13) |
| $\delta_0$ | critical significance level, defined in (14) |
| $\overline{\mathrm{err}}$, $\overline{\mathrm{unc}}$, $\overline{\mathrm{emp}}$ | "predictable" versions of err, unc, emp, as defined on p. 588 |
| $\overline{\mathrm{Err}}$, $\overline{\mathrm{Unc}}$, $\overline{\mathrm{Emp}}$ | "predictable" versions of Err, Unc, Emp |
| $F(t-)$ | the limit of $F(u)$ as $u$ approaches $t$ from below |
| $F(t+)$ | the limit of $F(u)$ as $u$ approaches $t$ from above |
| $u \vee v$ | the maximum of $u$ and $v$, also denoted $\max(u,v)$ |
| $u \wedge v$ | the minimum of $u$ and $v$, also denoted $\min(u,v)$ |
| $t^+$ | $t \vee 0$ |
| $t^-$ | $(-t) \vee 0$ |
| $\mathbf{U}$ | the uniform probability distribution in $[0,1]$ |
| $P_n(y\,\vert\,x,\sigma)$ | empirical estimate of $P(y\,\vert\,x)$, defined by (23) |
| $f_n(x,\sigma)$ | corresponding empirical predictability function, defined by (24) |
| $\mathbb{P}$ | probability |
| $\mathbb{E}$ | expectation |
| $\forall^\infty n$ | from some $n$ on |
| $\mathbb{I}_E$ | the indicator function of set $E$ |

## References

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

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

Luc Devroye, László Györfi, Adam Krzyżak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385, 1994.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

Yoav Freund, Yishay Mansour, and Robert E. Schapire. Generalization bounds for averaged classifiers. *Annals of Statistics*, 32(4), 2004.

Ronald L. Rivest and Robert H. Sloan. Learning complicated concepts reliably and usefully. In *Proceedings of the First Annual Conference on Computational Learning Theory*, pages 69–79, San Mateo, CA, 1988. Morgan Kaufmann.

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

Albert N. Shiryaev. *Probability*. Springer, New York, second edition, 1996.

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

Vladimir Vovk. Asymptotic optimality of Transductive Confidence Machine. In *Proceedings of the Thirteenth International Conference on Algorithmic Learning Theory*, volume 2533 of *Lecture Notes in Artificial Intelligence*, pages 336–350, Berlin, 2002b. Springer.

Vladimir Vovk. Universal well-calibrated algorithm for on-line classification. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines: Sixteenth Annual Conference on Learning Theory and Seventh Kernel Workshop*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 358–372, Berlin, 2003. Springer.

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