Towards a Universal Well-Calibrated Algorithm for On-line Classification

Vladimir Vovk

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England
vovk@cs.rhul.ac.uk

Technical Report

CLRC-TR-02-03

30th August 2002

Department of Computer Science
Egham, Surrey TW20 0EX, England

Abstract

This paper studies the problem of on-line classification in which the prediction algorithm is required to output a range of labels, rather than just one label, as its prediction; as usual, the examples are assumed to be generated independently from the same probability distribution P. It has been shown recently that there exist non-trivial prediction algorithms in this setting that are "well-calibrated" in the following sense: at any confidence level $1-\delta$, the long-run relative frequency of errors is guaranteed not to exceed δ. The number of "uncertain" predictive regions (i.e., those containing more than one label) is a natural measure of predictive performance for well-calibrated predictive algorithms. In this paper we construct, for any $\epsilon > 0$, an "ϵ-universal" predictive algorithm, which is within ϵ of being well-calibrated and makes an asymptotically smallest, to within ϵ, percentage of uncertain predictions almost surely under any unknown distribution P.

1 Introduction

The vast majority of classification algorithms output a point prediction for the label of an unknown object. This paper considers the problem of *region prediction*, where the algorithm's goal is to output as its prediction a range of possible labels for each unknown object. There are two natural desiderata for region predictors:

- ⋄ They should be *well calibrated*, in the sense that the long-run relative frequency of mistaken predictions should not exceed δ, where $1 - \delta$ is a prespecified confidence level given to the prediction algorithm as input.
- ⋄ They should be *successful*, in the sense that they rarely output *uncertain* predictive regions, i.e., predictive regions containing more than one label.

Paper [6] was the first to prove that a non-trivial on-line classification algorithm is well-calibrated. This paper makes the next step constructing a well-calibrated on-line classification algorithm that is "universal", in the sense of coming arbitrarily closely to having the asymptotically optimal rate of uncertain predictions among all well-calibrated region predictors.

Remark 1 This paper elaborates on [5], where it was shown that an optimal TCM exists when the distribution P generating the examples is known.

Remark 2 One important terminological difference between this paper and the previous ones [6, 5] on this topic is that here we never consider deterministic region predictors, TCM, etc., and therefore drop the word "randomised". The main use of random numbers in this theory is to deal efficiently with borderline cases; in practice deterministic region predictors often perform as well as randomised ones, but in theory the latter are more convenient. In one aspect our treatment of randomisation goes slightly beyond terminology: in our definition of individual strangeness measures we also allow randomisation (because of this Proposition 1 is formally stronger than the corresponding proposition in [6], although the proof of the former is a trivial modification of the proof of the latter).

2 Main result

In our learning protocol, Nature outputs pairs $(x_1, y_1), (x_2, y_2), \ldots$ 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 σ-algebra of all subsets). The protocol includes variables Err_n (the total number of errors made up to and including trial n) and err_n (the binary variable showing if an error is made at trial n); it also includes analogous variables Unc_n and unc_n for uncertain predictions:

$$
\begin{array}{l}
\mathrm{Err}_0 := 0 \\
\mathrm{Unc}_0 := 0 \\
\text{FOR } n = 1, 2, \ldots: \\
\qquad \text{Nature outputs } x_n \in \mathbf{X} \\
\qquad \text{Predictor outputs } \Gamma_n \subseteq \mathbf{Y} \\
\qquad \text{Nature outputs } y_n \in \mathbf{Y} \\
\qquad \mathrm{err}_n := \begin{cases} 1 & \text{if } y_n \notin \Gamma_n \\ 0 & \text{otherwise} \end{cases} \\
\qquad \mathrm{Err}_n := \mathrm{Err}_{n-1} + \mathrm{err}_n \\
\qquad \mathrm{unc}_n := \begin{cases} 1 & \text{if } |\Gamma_n| > 1 \\ 0 & \text{otherwise} \end{cases} \\
\qquad \mathrm{Unc}_n := \mathrm{Unc}_{n-1} + \mathrm{unc}_n \\
\text{END FOR.}
\end{array}
$$

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

We will be assuming 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^∞). This is Nature's randomised strategy.

A *region predictor* is a measurable function

$$\Gamma_\gamma(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n),$$

where $n = 1, 2, \ldots$, 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

$$
\begin{aligned}
&\Gamma_{\gamma_1}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) \\
&\subseteq \Gamma_{\gamma_2}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n)
\end{aligned}
$$

whenever $\gamma_1 \le \gamma_2$. Since we are interested in prediction with confidence, the region predictor is given an extra input $\gamma = 1 - \delta \in (0, 1)$, which we call the *confidence level* (typically it is close to 1, standard values being 95% and 99%); the complementary value δ is called the *significance level.* We will always assume that τ_n are independent random variables uniformly distributed in $[0, 1]$; this makes a region predictor a family (indexed by $\gamma \in (0, 1)$) of Predictor's randomised strategies.

We will often use the notation err_n, unc_n, etc., in the case where Predictor and Nature are using given randomised strategies: for example, $\mathrm{err}_n(P^\infty, \Gamma_{1-\delta})$ is the random variable equal to 1 if Predictor is right at trial n and 0 otherwise. It is always assumed that the random numbers τ_n used by Γ and the random examples z_n chosen by Nature are independent.

We say that a region predictor Γ is (conservatively) *well calibrated* if, for any

$\delta \in (0,1)$,

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \delta \quad \text{a.s.},$$

and we say that it is *ϵ-calibrated*, where ϵ is a positive constant, if the weaker property

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \delta + \epsilon \quad \text{a.s.}$$

holds for every $\delta \in (0,1)$. A region predictor Γ is *ϵ-universal*, where ϵ is a positive constant, if it is ϵ-calibrated and, for any well-calibrated region predictor $\Gamma^\dagger$,

$$\limsup_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \liminf_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma^\dagger_{1-\delta})}{n} + \epsilon \quad \text{a.s.}$$

(the symbol "a.s." assumes that Γ's and $\Gamma^\dagger$'s random numbers generators are independent).

Recall that a measurable space $\mathbf{X}$ is *Borel* if it is isomorphic to a measurable subset of the real line $\mathbb{R}$. 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 (see [3], B.3.2).

Theorem 1 *Suppose the object space $\mathbf{X}$ is Borel. For any $\epsilon > 0$, there exists an ϵ-universal region predictor.*

This is the main result of the paper; in §3 we construct an ϵ-universal region predictor and in §4 prove that it is indeed ϵ-universal.

3 Construction of an ϵ-universal region predictor

3.1 Random numbers

As we already explained, our prediction algorithm is given a sequence of random numbers τ_n, $n = 1, 2, \ldots$, independent and distributed uniformly in $[0,1]$. These random numbers will have several uses, and first of all we define a way of "splitting" a random number. If τ is a number in $[0,1]$, we split it into two numbers $\tau^{(i)} \in [0,1]$, $i = 1, 2$, as follows: if the binary expansion of τ is $0.a_1a_2\ldots$ (redefine the binary expansion of 1 to be $0.11\ldots$), set

$$\tau^{(i)} := 0.a_i a_{i+2} a_{i+4} \ldots, \quad i = 1, 2.$$

If τ is distributed uniformly in $[0,1]$, then both $\tau^{(i)}$ are, and they are independent of each other. We will let $\tau^{(i,j)}$ stand for $\left(\tau^{(i)}\right)^{(j)}$.

We will often apply our procedures (e.g., the "individual strangeness measure" in §3.2, the Nearest Neighbours rule in §3.4) 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 σ (to be extracted from one of the τ_n). In other words, along with examples (x, y) we will also consider *extended examples* $(x, \sigma, y) \in \tilde{\mathbf{Z}} := \mathbf{X} \times [0,1] \times \mathbf{Y}$.

3.2 Transductive Confidence Machines

Transductive Confidence Machine (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 and $\mathbb{R}$ is the set of all real numbers (equipped with the Borel σ-algebra), is called an *individual strangeness measure* if, for any $n = 1, 2, \ldots$, each α_i in

$$A_n : (w_1, \ldots, w_n) \mapsto (\alpha_1, \ldots, \alpha_n) \tag{1}$$

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

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

$$(\alpha_1, \ldots, \alpha_n) := A_n((x_1, \tau_1^{(1)}, y_1), \ldots, (x_{n-1}, \tau_{n-1}^{(1)}, y_{n-1}), (x_n, \tau_n^{(1)}, y)),$$

and

1. include the label y in $\Gamma_{1-\delta}$ if
$$\#\{i = 1, \ldots, n : \alpha_i > \alpha_n\}/n > \delta$$
2. do not include y in $\Gamma_{1-\delta}$ if
$$\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\}/n \leq \delta$$
3. otherwise, include y in $\Gamma_{1-\delta}$ if and only if
$$\tau_n^{(2)} < \frac{\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\} - n\delta}{\#\{i = 1, \ldots, n : \alpha_i = \alpha_n\}}. \tag{2}$$

A *TCM* is the TCM associated with some individual strangeness measure. Following [6], it is easy to show that

Proposition 1 *Every TCM is well-calibrated under every P.*

3.3 Universal TCM

Suppose $\mathbf{X}$ is a metric space (this will not restrict generality; see the next subsection). 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., [2], Chapter 11).

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

$$\forall n : K_n \le n \ \text{ and } \ K_n \to \infty,\ K_n/n \to 0\ (n \to \infty). \tag{3}$$

The *Nearest Neighbours TCM* is defined as follows. Let $w_1, \ldots, w_n$ be a sequence of extended examples $w_i = (x_i, \sigma_i, y_i)$. To define the corresponding αs (see (1)), we first define the Nearest Neighbours approximations $P_n(y \,|\, x)$ to the true (but unknown) conditional probabilities $P(y \,|\, x)$: for any example $(x, y) \in \mathbf{Z}$,

$$P_n(y \,|\, x) := N(x, y)/K_n,$$

where $N(x, y)$ is the number of $i = 1, \ldots, n$ such that x_i is among the K_n nearest neighbours of x and $y_i = y$. (To give a precise meaning to the expression "the kth nearest neighbour" in a sequence $x_1, \ldots, x_n$ to x, we use the convention that objects with smaller indices are considered to be nearer.)

Define the "empirical predictability function" f_n by

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

Now we can define the mapping (1); it will depend on a parameter B (large positive constant). Suppose we are given $w_i = (x_i, \sigma_i, y_i)$, $i = 1, \ldots, n$. For each $i = 1, \ldots, n$ choose a random label $\hat{y}_{n,i} \in \mathbf{Y}$ setting $\hat{y}_{n,i} := y$ with probability

$$\frac{B^{P_n(y \,|\, x_i)}}{\sum_y B^{P_n(y \,|\, x_i)}};$$

formally, for a fixed ordering $y^{(1)}, \ldots, y^{(|\mathbf{Y}|)}$ of the labels $y \in \mathbf{Y}$, we set $\hat{y}_{n,i} := y^{(M)}$ if

$$\sum_{m=1}^{M-1} \frac{B^{P_n(y^{(m)} \,|\, x_i)}}{\sum_y B^{P_n(y \,|\, x_i)}} \le \sigma_i \le \sum_{m=1}^{M} \frac{B^{P_n(y^{(m)} \,|\, x_i)}}{\sum_y B^{P_n(y \,|\, x_i)}}$$

(notice that the $\hat{y}_{n,i}$, $i = 1, \ldots, n$, are chosen independently, provided σ_i are independent). Finally, setting

$$\alpha_i := \begin{cases} 0 & \text{if } y_i = \hat{y}_{n,i} \\ f_n(x_i) & \text{otherwise} \end{cases} \tag{4}$$

completes the definition of the Nearest Neighbours TCM. In the next section we will show that this TCM is ϵ-universal when applied in a special way (see the next subsection) and provided B is large enough.

Remark 3 The way we defined the αs in [5] for the optimal TCM was much simpler, since the distribution P generating the examples was known precisely (and not "probably approximately"). For any $x \in \mathbf{X}$ we were able to fix some

$$\hat{y}(x) \in \arg\max_{y} P(y \mid x), \tag{5}$$

and then we could set

$$\alpha_i := \begin{cases} 0 & \text{if } y_i = \hat{y}(x_i) \\ P(\hat{y}_n(x_i) \mid x_i) & \text{otherwise.} \end{cases}$$

It remains an open question whether this construction, with P replaced by P_n, can be used for constructing a universal TCM; if so, one would only need deterministic individual strangeness measures, as in [6, 5], and one would be able to prove 0-universality instead of merely ϵ-universality. The problem with this definition is that unpleasant borderline effects appear for poorly predictable (viz., of predictability ≤ 0.5) examples. The definition used in this paper behaves in a much more regular way, replacing the "hard maximum" (5) with a softer version (B is large, so choosing y with probabilities proportional to $B^{P_n(y \mid x_i)}$ means, more or less, choosing one of the ys with the largest $P_n(y \mid x_i)$).

3.4 Regularisation

There is one complication to this already fairly complicated picture. Remember that B is a (large) positive constant. The above construction will be applied not to the original classification problem with the example space $\mathbf{X} \times \mathbf{Y}$ and the sequence $(\tau_1, \tau_2, \ldots) \in [0,1]^\infty$ of random numbers but to a slightly different *regularised* classification problem in which:

- the label space $\mathbf{Y}$ is extended by adding a new label y^*;
- each object x_n output by Nature is replaced by the extended object $(x_n, \tau_n^{(1,1)})$;
- each label y_n is changed to y^* with probability $\tau_n^{(1,1)}/B$ (formally, it is changed if $\tau_n^{(1,2)} < \tau_n^{(1,1)}/B$);
- the random numbers τ_n used by the prediction algorithm are replaced by $\tau_n^{(2)}$ (so that Nature's and Predictor's randomisations remain independent).

The new problem will be "close" to the original one when B is large, but more "regular" in two respects: the ties in the Nearest Neighbours procedure will occur with probability zero; the "predictability distribution function" will satisfy a Lipschitz condition. The predictability distribution function will be defined in the next section, and in the rest of this section we discuss the application of the Nearest Neighbours rule in the regularised classification problem, whose objects are extended objects (x, σ) of the original problem.

The last step in our construction is to set $\mathbf{X} := [0,1]$; we do this without loss of generality since $\mathbf{X}$ is Borel. Therefore, we need to apply the Nearest Neighbours rule to the extended objects (x_n, σ_n) that are elements of the set $\mathbf{X} \times [0,1] = [0,1]^2$. The distance between two such extended objects is defined as the usual Euclidean distance in the plane $\mathbb{R}^2$.

Remark 4 If typical distances between objects are small, the Nearest Neighbours procedure applied to extended objects will be very different from the Nearest Neighbours procedure applied to original objects. For a more natural extension of the Nearest Neighbours algorithm, see [2], §11.2, and [1].

4 Proof of Theorem 1

4.1 Fine details of region prediction

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 P and subindices $\mathbf{X}$ and $\mathbf{Y}\,|\,\mathbf{X}$ 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

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

(essentially, it is the distribution function of the image $P \circ f^{-1}$ of the probability distribution P under the mapping f). An example of such a function F is given in Figure 1; the graph of F is the thick line, and the unit box is also shown. It is clear that the predictability distribution function F satisfies these properties:

- $F(\epsilon) = 0$ for some $\epsilon > 0$ and $F(1) = 1$;
- F is non-decreasing;
- F is continuous on the right.

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

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

where t^+ stands for $\max(t, 0)$; the function $\mathbf{S}$ is also of the type $[0,1] \to [0,1]$. Geometrically, $\mathbf{S}$ is defined from the graph of F as follows (see Figure 1): move

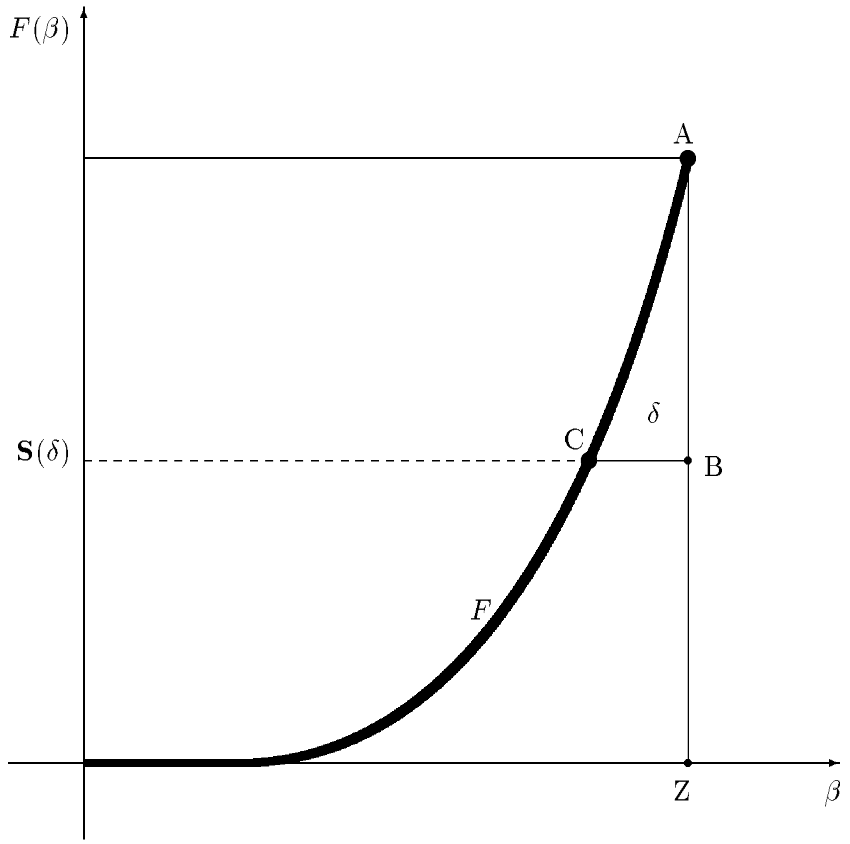

Figure 1 The predictability distribution function F. The function F is non-decreasing, continuous on the right, and $F(1/|\mathbf{Y}|) = 0$.

the point B from A to Z until the area of the curvilinear triangle ABC becomes δ (assuming this area does become δ eventually, i.e., δ is not too large); the ordinate of B is then $\mathbf{S}(\delta)$. Notice that the value $\mathbf{S}(\delta)$ in fact satisfies the equality

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

provided δ does not exceed the *critical significance level*

$$\delta_0 := \int_0^1 F(\beta) d\beta. \tag{6}$$

It is easy to see that the predictability distribution function F in the regularised classification problem is Lipschitz with coefficient B:

$$F(\beta_2) - F(\beta_1) = \mathbb{P}\{\beta_1 < f^* \le \beta_2\} \le B\,(\beta_2 - \beta_1) \tag{7}$$

for all $\beta_1, \beta_2 \in [0,1]$ such that $\beta_1 < \beta_2$ (we have used the fact that the predictability of an object $(x_i, \tau_i^{(1,1)})$ in the regularised classification problem equals the predictability of x_i in the original classification problem plus $\tau_i^{(1,1)}/B$, assuming B is large enough for y^* not to enter the set of the likeliest labels; to see that (7) holds for all β_1 and β_2, notice first that it holds for β_1 and β_2 that are close to each other).

If a region predictor Γ designed for the regularised classification problem is used in the original classification problem, its error rate

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n}{n}$$

will not increase by more than $1/B$ and its uncertainty rate

$$\limsup_{n\to\infty} \frac{\mathrm{Unc}_n}{n}$$

will not change at all.

The following result is proved in [5].

Proposition 2 *Let P be a probability distribution in $\mathbf{Z}$ with the success curve $\mathbf{S}$. If a region predictor Γ is well-calibrated under P, then for any significance level δ,*

$$\liminf_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \ge \mathbf{S}(\delta) \quad a.s. \tag{8}$$

The "a.s." in this proposition refers to the probability distribution $(P \times \mathbf{U})^\infty$ generating the sequence $x_1, \tau_1, y_1, x_2, \tau_2, y_2, \ldots$ of examples and random numbers, with $\mathbf{U}$ standing for the uniform distribution in $[0,1]$.

Theorem 1 will immediately follow from Proposition 1, Proposition 2 and the following proposition.

Proposition 3 *Suppose in the regularised classification problem the example-generating probability distribution in* $\mathbf{Z}$ *is* P *and the success curve is* $\mathbf{S}$*; let* $\epsilon > 0$*. If* B *is large enough, the Nearest Neighbours TCM constructed in §3.3 (for the regularised classification problem) satisfies, for any significance level* δ*,*

$$\limsup_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \mathbf{S}(\delta) + \epsilon \quad a.s. \tag{9}$$

Indeed, Proposition 1 implies that the constructed TCM is well calibrated in the regularised classification problem, and so is ϵ-calibrated in the original problem. It is clear that (9) will also be true in the original problem. To prove Theorem 1, it suffices to apply Proposition 2 to the original problem.

The rest of this proof will be about the regularised problem; $\mathbf{X}$ and $\mathbf{Y}$ will always stand for the object and label space of the regularised rather than original problem. We will be assuming, therefore, that the predictability distribution function is Lipschitz with coefficient B.

The goal of the next subsection is to prove Proposition 3.

4.2 Proof of Proposition 3

The idea of this proof elaborates that of Theorems 2 and 2r in [5].

Let P be a probability distribution in $\mathbf{Z}$, $\epsilon > 0$, and $\delta \in (0, 1)$ be some significance level; our goal is to prove (9).

The proof will be based on the following well-known fundamental result; it is applicable to the regularised problem, but we do not state the exact condition of regularity (which is that the second displayed equation on p. 172 of [2] always have a solution).

Lemma 1 *Suppose (3) is satisfied. For any* $\epsilon > 0$ *there exists an* $\epsilon^* > 0$ *such that, for large enough* n*,*

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

where the outermost probability distribution $\mathbb{P} := P^\infty$ *generates the examples, which determine the empirical distributions* P_n*.*

Proof Devroye et al. [2] (p. 172; see also the discussion of breaking distance ties in §11.2) state that

$$\mathbf{P}\left\{ \int |\eta(x) - \eta_n(x)|\, \mu(dx) > \epsilon \right\} \le 2e^{-n\epsilon^2/72}$$

(in their notation, except that their ϵ is replaced by 2ϵ and their γ_d is replaced by $\gamma_1 = 2$). We apply this bound 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:

$$\mathbf{P}\left\{\int |P(y \mid x) - \eta_n^y(x)|\,\mu(dx) > \epsilon\right\} \leq 2e^{-n\epsilon^2/72}, \tag{10}$$

where $\eta_n^y(x) = P_n(y \mid x)$ is just $\eta_n(x)$ for this binary classification problem. Markov's inequality and (10) imply

$$\mathbf{P}\left\{\mu\{|P(y \mid x) - \eta_n^y(x)| > \sqrt{\epsilon}\} > \sqrt{\epsilon}\right\} \leq 2e^{-n\epsilon^2/72},$$

which, in its turn, implies

$$\mathbf{P}\left\{\mu\left\{\max_{y \in \mathbf{Y}} |P(y \mid x) - \eta_n^y(x)| > \sqrt{\epsilon}\right\} > |\mathbf{Y}|\sqrt{\epsilon}\right\} \leq 2e^{-n\epsilon^2/72}.$$

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". Define $\mathcal{F}_n$, $n = 1, 2, \ldots$, to be the σ-algebra on $\tilde{\mathbf{Z}}^\infty$ generated by the first n extended examples (x_i, τ_i, y_i), $i = 1, \ldots, n$.

Lemma 2 *For any $\epsilon > 0$,*

$$\left(\forall^\infty n : \mathbb{P}\left\{\max_{y \in \mathbf{Y}} |P_n(y \mid x_n) - P(y \mid x_n)| > \epsilon \mid \mathcal{F}_{n-1}\right\} \leq \epsilon\right) \quad a.s. \tag{11}$$

and

$$\forall^\infty n : \mathbb{P}\left\{\max_{y \in \mathbf{Y}} |P_n(y \mid x_n) - P(y \mid x_n)| > \epsilon\right\} \leq \epsilon. \tag{12}$$

Proof Since $K_n \to \infty$,

$$\mathbb{P}\left\{\max_{y \in \mathbf{Y}} |P_n(y \mid x_n) - P(y \mid x_n)| > \epsilon \mid \mathcal{F}_{n-1}\right\}$$

will be less, from some n on, than

$$P^\infty\left\{\max_{y \in \mathbf{Y}} |P_{n-1}(y \mid x_n) - P(y \mid x_n)| > \frac{\epsilon}{2} \mid \mathcal{F}_{n-1}\right\},$$

where P_{n-1} is the Nearest Neighbours estimate based on the first $n-1$ examples (indeed, $|P_n(y \mid x_n) - P_{n-1}(y \mid x_n)| \leq 1/K_n$). To prove (11), it remains to apply the basic lemma.

Equation (12) follows from (11) and the general fact that, for any sequence of adapted random variables ξ_n taking values in $[0, 1]$,

$$(\forall^\infty n : \mathbb{P}\{\xi_n > \epsilon \mid \mathcal{F}_{n-1}\} \leq \epsilon) \quad \text{a.s.}$$

implies

$$\forall^\infty n : \mathbb{P}\{\xi_n > \epsilon\} \leq \epsilon.$$ ∎

Define

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

Since F satisfies the Lipschitz condition, we have $c < 1$ and $F(c) = \mathbf{S}(\delta)$.

By the martingale strong law of large numbers (see, e.g., [4], Theorem VII.5.4), it suffices to prove, for any $\epsilon > 0$,

$$\mathbb{P}(\mathrm{unc}_n \,|\, \mathcal{F}_{n-1}) \le F(c+2\epsilon) + \epsilon \quad \text{a.s.}$$

The prediction for the new object x_n will be certain if (a) $f_n(x_n) \ge c + \epsilon$, and (b) the number of $i = 1, \ldots, n-1$ for which $\hat{y}_{n,i} \ne y_i$ and $f_n(x_i) \ge c + \epsilon$ is less than $n\delta - 10$. Therefore, it suffices to prove that, almost surely,

- the conditional (on $\mathcal{F}_{n-1}$) probability of the event (a) is at least $1 - F(c + 2\epsilon) - \epsilon$ from some n on
- and that the event (b) always happens from some n on.

The first statement is very easy to prove: by Lemma 1, Lemma 2 (formula (11)), and the Borel–Cantelli lemma, we will have from some n on almost surely,

$$\begin{aligned}\mathbb{P}\{f_n(x_n) \ge c + \epsilon \,|\, \mathcal{F}_{n-1}\} &\ge \mathbb{P}\{f(x_n) > c + 2\epsilon \,|\, \mathcal{F}_{n-1}\} - \epsilon \\ &= 1 - F(c+2\epsilon) - \epsilon.\end{aligned}$$

To prove that the event (b) always happens from some n on, we first estimate

$$\mathbb{E}\,\#\{i = 1, \ldots, n-1 : \hat{y}_{n,i} \ne y_i \;\&\; f_n(x_i) \ge c + \epsilon\}, \tag{13}$$

where the expectation is w.r. to the randomisation in the choice of $\tau_n^{(1)}$. (Remember that the τ_n, as everything else in this part of the proof, are those of the regularised rather than original classification problem.) The expectation (13) is bounded above by

$$\sum_{i=1}^{n-1} I(x_i) \left(1 - \frac{B^{P_n(y_i \,|\, x_i)}}{\sum_{y \in \mathbf{Y}} B^{P_n(y_i \,|\, x_i)}}\right), \tag{14}$$

where the function I is defined as follows:

$$I(x) := \begin{cases} 0 & \text{if } f_n(x) \ge c + \epsilon + \frac{1}{B} \\ 1 & \text{if } f_n(x) \le c + \epsilon \\ 1 - B(f_n(x) - c - \epsilon) & \text{otherwise} \end{cases}$$

(in other words, I is a "soft version" of $\mathbb{I}_{f_n \ge c+\epsilon}$, continuously depending on f_n).

The main tool in this part of the proof will be McDiarmid's theorem (see, e.g., [2], Theorem 9.2). We start by checking the possibility of its application. If we replace an example $z_j = (x_j, y_j)$ by another example $z_j^* = (x_j^*, y_j^*)$, the expression (14) will change as follows:

- the term for $i = j$ changes by 1 at most;
- the terms $i \neq j$ outside the *vicinity* of x_j (i.e., the terms corresponding to the examples (x_i, y_i) such that x_j is among the K_n nearest neighbours of x_i) and outside the vicinity of x_j^* do not change at all;
- the sum over the at most $2K_n$ terms $i \neq j$ in the vicinity of x_j and/or in the vicinity of x_j^* can change by at most

$$\begin{aligned}
&\max_{a,b\in[0,1]} \left(2K_n \left(\left(a + \frac{2B}{K_n} \right) \left(1 - b\frac{B^{-2/K_n}}{B^{2/K_n}} \right) - a(1-b) \right) \right) \\
&= \max_{a,b\in[0,1]} \left(2K_n \left(\frac{2B}{K_n} - abB^{-4/K_n} - \frac{2B}{K_n} bB^{-4/K_n} + ab \right) \right) \\
&\le 2K_n \left(\frac{2B}{K_n} + 1 - B^{-4/K_n} \right) \le C(B),
\end{aligned} \tag{15}$$

where $C(B)$ is a constant depending only on B (to see that $K_n(1 - B^{-4/K_n})$ is bounded, notice that

$$K(1 - B^{-4/K}) \to 4 \ln B$$

as $K \to \infty$).

Therefore, the change $z_j \mapsto z_j^*$ does not change the expression (14) by more than $C(B)$, some constant depending only on B.

Next we consider

$$\begin{aligned}
&\mathbb{E} \left(I(x_i) \left(1 - \frac{B^{P_n(y_i \mid x_i)}}{\sum_{y \in \mathbf{Y}} B^{P_n(y \mid x_i)}} \right) \right) \\
&= \mathbb{E} \left(I(x_n) \left(1 - \frac{B^{P_n(y_n \mid x_n)}}{\sum_{y \in \mathbf{Y}} B^{P_n(y \mid x_n)}} \right) \right)
\end{aligned}$$

(the equality holds because of the symmetry). It is clear that if B is large, this expectation is arbitrarily close to being smaller than

$$\delta^* := \int_{c+\epsilon}^{1} (F(\beta) - F(c+\epsilon))\, d\beta < \delta.$$

McDiarmid's theorem ([2], Theorem 9.2) implies that the probability that (14) will be greater than $n^{\frac{\delta+\delta^*}{2}}$ is exponentially small; in combination with the Borel–Cantelli lemma this shows that (14) does not exceed $n^{\frac{\delta+\delta^*}{2}}$ from some n on. This implies, by a similar application of McDiarmid's theorem (whose "bounded differences" condition is now obviously satisfied) and the Borel–Cantelli lemma, that the event (b) happens from some n on.

5 Conclusion

We have shown that there exist ϵ-universal well-calibrated region predictors, thus satisfying, to some degree, both desiderata mentioned in §1. 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 δ and are independent, whereas the universality is an asymptotic property.

Acknowledgments

This work was partially supported by EPSRC (grant GR/R46670/01 "Complexity Approximation Principle and Predictive Complexity: Analysis and Applications"), BBSRC (grant 111/BIO14428 "Pattern Recognition Techniques for Gene and Promoter Identification and Classification in Plant Genomic Sequences"), and EU (grant IST-1999-10226 "EurEdit: The Development and Evaluation of New Methods for Editing and Imputation").

References

[1] 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.

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

[3] Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

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

[5] Vladimir Vovk. Asymptotic optimality of Transductive Confidence Machine. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory*, 2002.

[6] Vladimir Vovk. On-line Confidence Machines are well-calibrated. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 2002. Full version published as Technical Report CLRC-TR-02-01, Computer Learning Research Centre, Department of Computer Science, Royal Holloway, University of London, April 2002, which can be downloaded from `http://www.clrc.rhul.ac.uk`. Additional information can be found at `http://www.cs.rhul.ac.uk/~vovk/cm`.