On-Line Confidence Machines Are Well-Calibrated

Vladimir Vovk

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

On-line Compression Modelling Project

Working Paper #1

April 28, 2002

Project web site:
http://vovk.net/kp

Abstract

Transductive Confidence Machine (TCM) and its computationally efficient modification, Inductive Confidence Machine (ICM), are ways of complementing machine-learning algorithms with practically useful measures of confidence. We show that when TCM and ICM are used in the on-line mode, their confidence measures are well-calibrated, in the sense that predictive regions at confidence level $1 - \delta$ will be wrong with relative frequency at most δ (approaching δ in the case of randomised TCM and ICM) in the long run. This is not just an asymptotic phenomenon: actually the error probability of randomised TCM and ICM is δ at every trial and errors happen independently at different trials.

Contents

1 Introduction

The bulk of work in computational learning theory is done under the *i.i.d. assumption*: the random examples fed to the learning algorithm are independent and identically distributed. This is the assumption that we make in this paper (but no other assumptions are made). The modern theory was started by Vapnik and Chervonenkis (see [12] for a recent review) and, much later but independently, by Valiant [11]; nowadays, this theory is often referred to as the theory of PAC learning. It produced not only deep mathematical results but also efficient learning algorithms that work very well in practice.

An apparent drawback of the theory is that it only studies algorithms generating "bare predictions", i.e., algorithms predicting labels for new objects without saying how reliable these predictions are. A major concern of the theory of PAC learning, however, is estimation of the probability of erroneous predictions, and, in principle, a low bound on the error probability would mean high confidence in the prediction. A serious real drawback of the theory is the weakness of error bounds it produces in practice, even for relatively clean data sets: e.g., for the standard USPS data set of hand-written digits (described in [12], p. 496) typical PAC bounds on error probability exceed one [5].

The notion of Transductive Confidence Machine (TCM) was introduced in [8, 15] to provide a different framework that would allow practically useful confidence measures. As reported in [7], TCM indeed produces practically meaningful results on the USPS data set: the "Nearest Neighbour TCM" (defined in §3 below) is able to predict the vast majority (approximately 95%) of test examples at the confidence level 99%.

The problem of prediction with confidence can be formalised, as done in this paper, as that of computing "predictive regions" (sets of labels) rather than bare predictions. (Although this is not the only way to package the output of prediction with confidence; e.g., in [8, 15] we preferred to present TCM's output as a bare prediction plus two measures of its quality, "confidence" and "credibility".) A prediction algorithm takes as input a "confidence level" $1-\delta$ and for each new object outputs as its prediction a predictive region rather than a point prediction. There are two natural desiderata for such algorithms:

- they should be *well-calibrated*, in the sense that in the long run the predictions are wrong with relative frequency at most δ;

- they should *perform* well, in the sense that the number of *uncertain* (containing more than one label) predictions should be as small as possible.

The first desideratum is the priority: without it, the meaning of predictive regions is lost, and it becomes easy to achieve the best possible performance. This paper constructs the first non-trivial prediction algorithm (randomised TCM) which is shown to be well-calibrated without using any assumptions beyond i.i.d.; moreover, this algorithm is well-calibrated in a very strong non-asymptotic sense: the probability of error is always δ and errors are independent of each other.

The qualification "non-trivial" is essential, since it is easy to construct trivial well-calibrated algorithms (e.g., always output the predictive region containing all possible labels). The theory of this paper will be exclusively about the first of the two desiderata listed above, but to convince the reader that it is natural to care whether TCM is well-calibrated we briefly report some experimental results.

Figures 1 and 2 show the on-line performance of the Nearest Neighbour TCM on the USPS data set (the original 9298 hand-written digits, but randomly permuted) for the confidence levels 95% and 99%, respectively. For every new hand-written digit TCM predicts a set of possible labels (0 to 9) for this digit (the predictive region). The solid line shows the cumulative number of errors, dotted the cumulative number of uncertain predictive regions, and dashdot the cumulative number of empty predictive regions (inevitably leading to an error). In Figure 2, the dashdot line coincides with the horizontal axis (there are no empty predictions) and so is invisible. (The four figures in this paper are not significantly affected by statistical variation due to the random choice of the permutation of the data set.) We can see that the performance of this particular TCM on this particular data set is quite good: for most examples the predictive region has at most one label. For theoretical results about TCM's performance, see [13] and [14].

Notice that Figures 1 and 2 also provide empirical evidence that TCM is well-calibrated. It can be seen that the number of errors made grows linearly, and the slope is approximately 5% for the confidence level 95% and 1% for the confidence level 99%.

A popular alternative to PAC learning is Bayesian learning. Bayesian algorithms, however, will not be well-calibrated: we require calibration under *any* i.i.d. distribution, whereas Bayesian algorithms are constructed for a

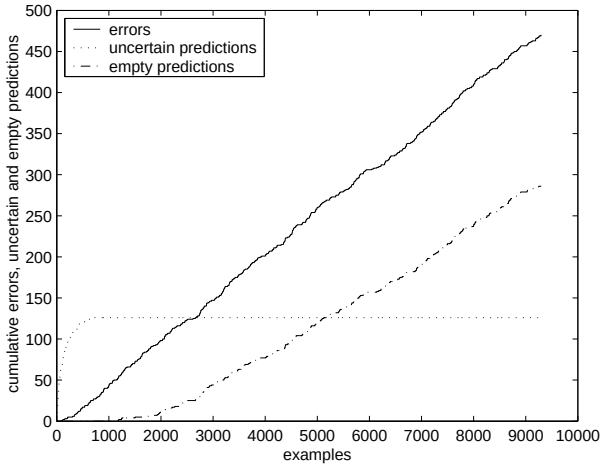

Figure 1: TCM at 95%

given complete specification of the probability distribution generating the data. See [3] for a comparison between the TCM and Bayesian approaches.

This paper is completely self-contained and independent of the previous papers about TCM (which were of experimental nature and about the batch setting, in contrast to this paper's on-line setting).

2 Region predictors

Our basic protocol is as follows. Nature outputs pairs

$$(x_1, y_1), (x_2, y_2), \dots \tag{1}$$

called *examples*. Each example (x_i, y_i) consists of an *object* x_i and its *label* y_i; e.g., the objects can be hand-written digits and y_i their classifications (numbers from 0 to 9). The objects are elements of a measurable space $\mathbf{X}$ called the *object space* and the labels are elements of a measurable space $\mathbf{Y}$ called the *label space*. We will use the notation $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ for the *example space*; therefore, the infinite data sequence (1) will be an element of the measurable space $\mathbf{Z}^\infty$.

We will be assuming that the data sequence (1) is output according to some probability distribution P in $\mathbf{Z}^\infty$. The usual further assumption made

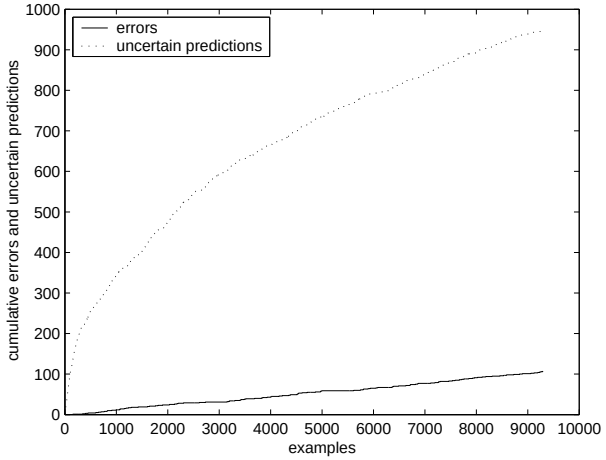

Figure 2: TCM at 99%

in PAC theory is that P is an *i.i.d. distribution*, i.e., $P = Q^\infty$, where Q is a distribution in $\mathbf{Z}$ (in other words, that individual examples are generated by Q independently of each other). For us, it will be sufficient to make the weaker assumption that P is *exchangeable*: for every positive integer n, every permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ of the set $\{1, \dots, n\}$, and every measurable set $E \subseteq \mathbf{Z}^n$,

$$\begin{aligned} &P\{(z_1, z_2, \dots) \in \mathbf{Z}^\infty : (z_1, \dots, z_n) \in E\} \\ &= P\left\{(z_1, z_2, \dots) \in \mathbf{Z}^\infty : (z_{\pi(1)}, \dots, z_{\pi(n)}) \in E\right\}. \end{aligned}$$

The difference is small, however (in the current context of infinite data sequences): according to de Finetti's representation theorem (see, e.g., [9], Theorem 1.49), every exchangeable distribution is a mixture of i.i.d. distributions provided the example space $\mathbf{Z}$ is Borel.

We are interested in algorithms for predicting, at every trial n, the label y_n given the object x_n and all the previous examples, from (x_1, y_1) to (x_{n-1}, y_{n-1}). Since we are interested in prediction with confidence, our algorithms are given an extra input $(1 - \delta) \in (0, 1)$, which we call the *confidence level*. Formally, we define a *region predictor* to be a function

$$\Gamma : \mathbf{Z}^* \times \mathbf{X} \times (0, 1) \to 2^{\mathbf{Y}} \tag{2}$$

($2^{\mathbf{Y}}$ is the set of all subsets of $\mathbf{Y}$; the argument $(1-\delta) \in (0, 1)$ will be written as subindex) which, for every confidence levels $1 - \delta_1 \le 1 - \delta_2$, every positive

integer n, and every *incomplete data sequence*

$$x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n \tag{3}$$

(we often ignore unnecessary parentheses, such as those around (x_i, y_i)) satisfies

$$\begin{aligned} &\Gamma_{1-\delta_1}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \\ &\subseteq \Gamma_{1-\delta_2}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n). \end{aligned} \tag{4}$$

Intuitively, given the incomplete data sequence (3) and a confidence level $1-\delta$, the region predictor Γ predicts that

$$y_n \in \Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n),$$

and the larger $1-\delta$ the more emphatic the prediction; condition (4) is a natural requirement of consistency. Formally, for any infinite data sequence

$$\omega = (x_1, y_1, x_2, y_2, \ldots), \tag{5}$$

confidence level $1-\delta$, and positive integer n, we define the number of errors that Γ makes at the confidence level $1-\delta$ on the sequence ω during the first n trials to be

$$\begin{aligned} &\mathrm{Err}_n(\Gamma_{1-\delta}, \omega) := \#\{i = 1, \ldots, n : \\ &y_i \notin \Gamma_{1-\delta}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i)\}, \end{aligned}$$

where $\#B$ stands for the size of the set B. Sometimes we will also need the individual prediction results

$$\begin{aligned} &\mathrm{err}_n(\Gamma_{1-\delta}, \omega) := \mathrm{Err}_n - \mathrm{Err}_{n-1} \\ &= \begin{cases} 1 & \text{if } y_n \notin \Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

and the whole infinite sequence of prediction results

$$\begin{aligned} &\mathrm{err}(\Gamma_{1-\delta}, \omega) := \\ &(\mathrm{err}_1(\Gamma_{1-\delta}, \omega), \mathrm{err}_2(\Gamma_{1-\delta}, \omega), \ldots). \end{aligned}$$

In §4 we will also consider *randomised region predictors*, which depend, additionally, on an element of an auxiliary probability space.

3 Transductive Confidence Machine

Transductive Confidence Machine (TCM) is a way to define a region predictor from a "bare predictions" algorithm. Formally, it is a way of transition from what we call an "individual strangeness measure" to a region predictor; first we will give formal definitions and then give a simple example of an individual strangeness measure.

A family of measurable functions $\{A_n : n \in \mathbb{N}\}$, where $A_n : \mathbf{Z}^n \to \mathbb{R}^n$ for all n, $\mathbb{N}$ is the set of all positive integers 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 $n \in \mathbb{N}$, any permutation π of $\{1, \dots, n\}$, any $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and any $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$,

$$\begin{aligned}(\alpha_1, \dots, \alpha_n) = A_n(z_1, \dots, z_n) \implies \\ (\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}) = A_n(z_{\pi(1)}, \dots, z_{\pi(n)}).\end{aligned} \tag{7}$$

In other words,

$$A_n : (z_1, \dots, z_n) \mapsto (\alpha_1, \dots, \alpha_n) \tag{8}$$

is called an individual strangeness measure if every α_i is determined by the (unordered) bag $\lbag z_1, \dots, z_n \rbag$ and z_i. (The difference between the bag $\lbag z_1, \dots, z_n \rbag$ and the set $\{z_1, \dots, z_n\}$ is that the former can contain several copies of the same element.)

The *TCM associated with the individual strangeness measure* A_n is the following region predictor:

$$\Gamma_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \tag{9}$$

is defined to be the set of all labels $y \in \mathbf{Y}$ such that

$$\frac{\#\{i = 1, \dots, n : \alpha_i \geq \alpha_n\}}{n} > \delta, \tag{10}$$

where

$$\begin{aligned}(\alpha_1, \dots, \alpha_n) := \\ A_n((x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y)).\end{aligned} \tag{11}$$

In general, a *TCM* is the TCM associated with some individual strangeness measure.

The definition of TCM can be illustrated by the following simple example of an individual strangeness measure, the one used in producing Figures 1–4: mapping (8) can be defined, in the spirit of the 1-Nearest Neighbour Algorithm, as (assuming the objects are vectors in a Euclidean space)

$$\alpha_i := \frac{\min_{j \neq i: y_j = y_i} d(x_i, x_j)}{\min_{j \neq i: y_j \neq y_i} d(x_i, x_j)}, \tag{12}$$

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).

Of course, there are many other ways of defining individual strangeness measures (e.g., [8, 15] used the Lagrange multipliers in Support Vector Machine as the αs). As soon as we have an individual strangeness measure, the corresponding TCM is defined automatically in a simple way (cf. (9)–(11)); in particular, the learning component of TCM always lies in the individual strangeness measure.

Let us say that a set $E \subseteq \{0,1\}^\infty$ is *monotonic* if, for any two infinite binary sequences $(a_1, a_2, \ldots)$ and $(b_1, b_2, \ldots)$,

$$\left.\begin{array}{r} (a_1, a_2, \ldots) \in E \\ a_i \le b_i, \forall i \end{array}\right\} \Longrightarrow (b_1, b_2, \ldots) \in E. \tag{13}$$

The following result shows that, as far as upper bounds on $P\{\mathrm{err} \in E\}$ for monotonic E are concerned, we can assume that the error probability of Γ is δ at every trial and errors happen independently at different trials.

Theorem 1 *For any confidence level $1 - \delta$, any exchangeable probability distribution P in $\mathbf{Z}^\infty$, and any monotonic $E \subseteq \{0,1\}^\infty$, any TCM Γ satisfies*

$$P\{\omega : \mathrm{err}(\Gamma_{1-\delta}, \omega) \in E\} \le B_\delta^\infty(E), \tag{14}$$

where B_δ is the Bernoulli distribution in $\{0,1\}$ with the parameter δ: $B_\delta\{1\} = \delta$ and $B_\delta\{0\} = 1 - \delta$.

Corollary 1 *Each TCM Γ is* conservatively well-calibrated *in the sense that, for any exchangeable probability distribution P in $\mathbf{Z}^\infty$ and any confidence level $1 - \delta$,*

$$\limsup_{n \to \infty} \frac{\mathrm{Err}_n(\Gamma_{1-\delta}, \omega)}{n} \le \delta \tag{15}$$

for P-almost all $\omega \in \mathbf{Z}^\infty$.

This corollary immediately follows from the usual strong law of large numbers and Theorem 1 since the complement of (15) is monotonic. Using, instead, the law of the iterated logarithm, we can strengthen (15) to

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma_{1-\delta},\omega) - n\delta}{\sqrt{2\delta(1-\delta)n\ln\ln n}} \le 1.$$

We will also state two finite-sample implications of Theorem 1: Hoeffding's inequality (see, e.g., [2], Theorem 8.1) implies that, for any positive integer N and any constant $\epsilon > 0$,

$$P\left\{\omega : \mathrm{Err}_N(\Gamma_{1-\delta},\omega) \ge N(\delta+\epsilon)\right\} \le e^{-2N\epsilon^2};$$

the central limit theorem implies that, for any constant c,

$$\limsup_{N\to\infty} P\left\{\omega : \mathrm{Err}_N(\Gamma_{1-\delta},\omega) \ge N\delta + c\sqrt{N}\right\} \le \frac{1}{\sqrt{2\pi}} \int_{\frac{c}{\sqrt{\delta(1-\delta)}}}^{\infty} e^{-u^2/2} du.$$

4 Randomised Transductive Confidence Machine

In this section we introduce a modification of TCM which will allow us to simplify, strengthen, and prove easily Theorem 1. The *randomised Transductive Confidence Machine (rTCM) associated with the individual strangeness measure* A_n is the following randomised region predictor Γ: for any label $y \in \mathbf{Y}$,

1. if $\#\{i = 1,\dots,n : \alpha_i > \alpha_n\}/n > \delta$ (as before, the αs are defined by (11)), the label y is included in (9);
2. if $\#\{i = 1,\dots,n : \alpha_i \ge \alpha_n\}/n \le \delta$, y is not included in (9);
3. otherwise, y is included in (9) with probability

$$\frac{\#\{i = 1,\dots,n : \alpha_i \ge \alpha_n\} - n\delta}{\#\{i = 1,\dots,n : \alpha_i = \alpha_n\}}. \tag{16}$$

In the typical case where all or almost all $\alpha_1, \ldots, \alpha_n$ are different, there is very little difference between TCM and rTCM (provided n is not too small).

To make the definition of rTCM more formal, we introduce the auxiliary probability space $([0,1]^\infty, U^\infty)$, where $[0,1]^\infty$ is equipped with the standard σ-algebra and U is the uniform probability distribution in $[0,1]$; intuitively, $(\tau_1, \tau_2, \ldots) \in [0,1]^\infty$ are random numbers for use at trials $1, 2, \ldots$, respectively, produced by a random number generator. The rTCM Γ is a function of the type

$$\Gamma : \mathbf{Z}^* \times \mathbf{X} \times (0,1) \times [0,1] \to 2^{\mathbf{Y}}$$

(cf. (2)) where the dependence on the extra argument $\tau \in [0,1]$ (random number) arises because of item 3 of the definition of rTCM; for concreteness, we interpret it as: y is included in $\Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, \tau_n)$ if

$$\tau_n < \frac{\#\{i : \alpha_i \geq \alpha_n\} - n\delta}{\#\{i : \alpha_i = \alpha_n\}}. \tag{17}$$

Notice that functions such as err depend on the extra argument $\tau \in [0,1]^\infty$ in the case of rTCM.

Theorem 2 *For any rTCM Γ, any confidence level $1-\delta$, and any exchangeable probability distribution P in $\mathbf{Z}^\infty$, the image of $P \times U^\infty$ under the mapping*

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \mapsto \mathrm{err}(\Gamma_{1-\delta}, \omega, \tau)$$

is the probability distribution B_δ^∞ of independent Bernoulli trials with parameter δ.

This theorem may appear too strong to be true: it is generally believed that to make categorical assertions about error probabilities some Bayesian-type assumptions are needed and that the general i.i.d. assumption is not sufficient. For example, in the theory of PAC learning an error probability ϵ is only asserted with some probability $1-\delta$. It should be remembered, however, that Theorem 2 does not assert that the probability of error, $\mathrm{err}_n = 1$, is δ conditionally on knowing the whole past (3); it is only asserted that it is δ unconditionally and conditionally on knowing $\mathrm{err}_1, \ldots, \mathrm{err}_{n-1}$. (Actually, it is quite obvious that the probability of error is often not equal to δ if the whole past is known: if the predictive region is empty, the conditional probability of error is 1; to balance this, the conditional probability that a non-empty predictive region is wrong will tend to be less than δ.)

Theorem 2 immediately implies Theorem 1: if an rTCM Γ and a TCM $\Gamma^\dagger$ are constructed from the same individual strangeness measure, the latter's errors $\mathrm{err}_n^\dagger$ never exceed the former's errors err_n, $\mathrm{err}_n^\dagger \le \mathrm{err}_n$. Theorem 2 also implies

Corollary 2 *Every rTCM Γ is* precisely well-calibrated *in the sense that*

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma_{1-\delta}, \omega, \tau)}{n} = \delta$$

for $P \times U^\infty$-almost all $\omega \in \mathbf{Z}^\infty$ and $\tau \in [0,1]^\infty$.

5 Inductive Confidence Machine

For large data sets, TCMs can be computationally inefficient. Inductive Confidence Machine (ICM) is a modification of TCM which sacrifices (in typical cases) some predictive accuracy for computational efficiency (for details of ICM in the batch setting, see [6]).

In the case of ICMs, the role of an individual strangeness measure will be played by a pair consisting of an "inductive algorithm" and "discrepancy measure". Let $\hat{\mathbf{Y}}$ be a *prediction space* (an arbitrary measurable space). An *inductive algorithm* D is a measurable function that maps every bag $\lbag z_1, \dots, z_n \rbag$ (of any size) of elements of $\mathbf{Z}$ to a function $D_{\lbag z_1,\dots,z_n\rbag} : \mathbf{X} \to \hat{\mathbf{Y}}$. The usual interpretation of $D_{\lbag z_1,\dots,z_n\rbag}$ is that it is a decision rule, found from the training set $\lbag z_1, \dots, z_n \rbag$, which computes the predicted label $\hat{y} := D_{\lbag z_1,\dots,z_n\rbag}(x)$ for any new object x. Usually, but not always, $\hat{\mathbf{Y}} = \mathbf{Y}$. A *discrepancy measure* is a measurable function $\Delta : \mathbf{Y} \times \hat{\mathbf{Y}} \to \mathbb{R}$; it will be used to measure the discrepancy between the predicted label $\hat{y}$ and the true label y.

Given an inductive algorithm D and a discrepancy measure Δ, we can define an individual strangeness measure $\{A_n\}_{n=1}^\infty$ as follows: for any $((x_1, y_1), \dots, (x_n, y_n))$ in $\mathbf{Z}^*$, the values

$$(\alpha_1, \dots, \alpha_n) = A_n((x_1, y_1), \dots, (x_n, y_n)) \tag{18}$$

can be defined by the formula

$$\alpha_i := \Delta\left(y_i, D_{\lbag (x_1,y_1),\dots,(x_n,y_n)\rbag}(x_i)\right) \tag{19}$$

or the formula

$$\alpha_i := \Delta(y_i, \\ D_{\wr(x_1,y_1),\dots,(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),\dots,(x_n,y_n)\int}(x_i)). \tag{20}$$

This shows that with every inductive algorithm and discrepancy measure we can associate a TCM. Formula (20) is more natural than (19) but typically leads to less computationally efficient TCMs.

At a crude level, one can divide inductive algorithms into two classes: "proper inductive algorithms" and "transductive algorithms" (see [12]; sometimes transductive algorithms are called "instance-based"). For proper inductive algorithms, $D_{\wr z_1,\dots,z_n\int}$ can be computed, in some sense: e.g., $D_{\wr z_1,\dots,z_n\int}$ may be described by a polynomial, and computing $D_{\wr z_1,\dots,z_n\int}$ may mean computing the coefficients of the polynomial; as soon as $D_{\wr z_1,\dots,z_n\int}$ is computed, computing $D_{\wr z_1,\dots,z_n\int}(x)$ for a new object x takes very little time. For transductive algorithms (such as the Nearest Neighbours Algorithms), relatively little can be done before seeing the new object x; even allowing considerable time for pre-processing $\wr z_1,\dots,z_n\int$, computing $D_{\wr z_1,\dots,z_n\int}(x)$ will be a difficult task.

Notice that, even when formula (19) rather than (20) is used and D is a proper inductive algorithm, the TCM based on the individual strangeness measure (19) will still be inefficient: for every new example (x_n, y_n), computing $\Gamma_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ will require constructing a new decision rule.

To define an ICM from an inductive algorithm D and a discrepancy measure Δ, first fix a finite or infinite sequence of positive integer parameters $m_1, m_2, \dots$ (called *training trials*); it is required that $m_1 < m_2 < \cdots$. If the sequence $(m_1, m_2, \dots) = (m_1, \dots, m_l)$ is finite, we set $m_i := \infty$ for $i > l$. The ICM based on D, Δ, and the sequence $m_1, m_2, \dots$ of training trials is defined to be the region predictor Γ such that $\Gamma_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ is computed as follows:

- if $n \le m_1$, $\Gamma_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ is found using a fixed TCM;
- otherwise, find the k such that $m_k < n \le m_{k+1}$ and set

$$\Gamma_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) := \Big\{ y \in \mathbf{Y} : \\ \frac{\#\{j = m_k + 1, \dots, n : \alpha_j \ge \alpha_n\}}{n - m_k} > \delta \Big\}, \tag{21}$$

where the αs are defined by

$$\begin{aligned} \alpha_j &:= \Delta\left(y_j, D_{\wr(x_1,y_1),\ldots,(x_{m_k},y_{m_k})\wr}(x_j)\right), \\ & j = m_k + 1, \ldots, n-1, \\ \alpha_n &:= \Delta\left(y, D_{\wr(x_1,y_1),\ldots,(x_{m_k},y_{m_k})\wr}(x_n)\right). \end{aligned} \tag{22}$$

We can see that ICM requires recomputing the decision rule being used not at every trial but only at the training trials $m_1, m_2, \ldots$; the rate of growth of m_i determines the chosen balance between predictive accuracy and computational efficiency. The most important case is perhaps where there is only one training trial m_1. Randomised ICM (rICM) can be defined analogously to rTCM.

Theorem 3 *Theorems 1 and 2 continue to hold in the case of ICMs and rICMs, respectively.*

Let a and b be positive numbers such that either $a \geq 1$ and $b \geq 1$ or $a > 1$. If an individual strangeness measure A_n is computable in time $\Theta(n^a \log^b n)$, the TCM associated with A_n spends time $\Theta(n^{a+1} \log^b n)$ on the computations needed for the first n trials. On the other hand, if an inductive algorithm D is computable in time $\Theta(n^a \log^b n)$, a discrepancy measure Δ is computable in constant time, and the sequence m_i is infinite and grows exponentially, the ICM based on D, Δ, and (m_i) spends the same, to within a constant factor, time $\Theta(n^a \log^b n)$. (We have been assuming that the TCM or ICM is given A_n or D as an oracle and the label space $\mathbf{Y}$ is finite and fixed.) In the case where the sequence (m_i) is finite, the ICM's computation time becomes $\Theta(n \log n)$ (e.g., use red-black trees for storing α_is; [1], Chapters 14 and 15).

The performance of the Nearest Neighbour ICM on the USPS data set with training trial 4649 (the middle of the data set) is shown in Figures 3 and 4 for the confidence levels 95% and 99%, respectively; in accordance with Theorem 3, starting from scratch at trial 4670 does not affect the error rate (solid line). It can be seen from these figures (and is obvious anyway) that the ICM's performance (measured by the number of uncertain predictions) deteriorates sharply after training trials m_i. Perhaps in practice there should be short spells of "learning" after each training trial, when the ICM is provided with fresh "training examples" and its predictions are not used or evaluated.

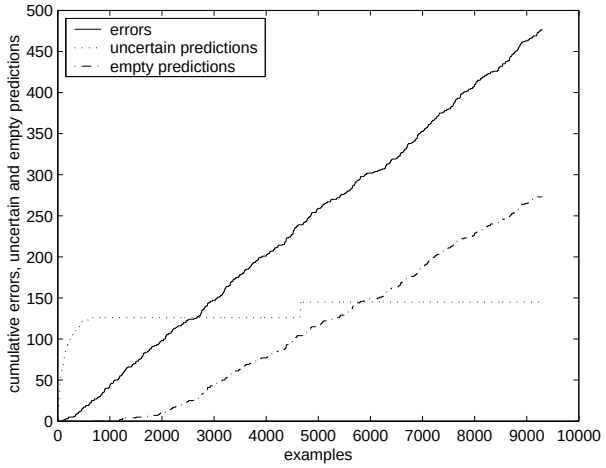

Figure 3: ICM at 95%

6 Conclusion

The main advantages of this paper's approach are:

- As compared to the standard theory of PAC learning, our error bounds are practically meaningful (see §1 and Figures 1–4).
- As compared to the theory of Bayesian learning, we do not assume anything beyond the exchangeability of the underlying probability distribution.
- The usual justification ("validity") of TCM with confidence level $1 - \delta$ is the fact that the error probability at any trial does not exceed δ (see, e.g., [4], Theorem 1). This paper adds the crucial observation that, in the case of rTCM, the events "error at trial n", $n = 1, 2, \ldots$, not only have probability δ, but also are independent of each other. Very little can be said about a sequence of events of probability δ, but combined with independence this gives us a plethora of known properties for Bernoulli trials.

Acknowledgments. I am very grateful to Alex Gammerman, Phil Dawid, Yoav Freund, Ilia Nouretdinov, Leo Gordon, and Kostas Proedrou for their help and valuable discussions and to the members of the Programme Committee and referees for helpful suggestions on improving the presentation.

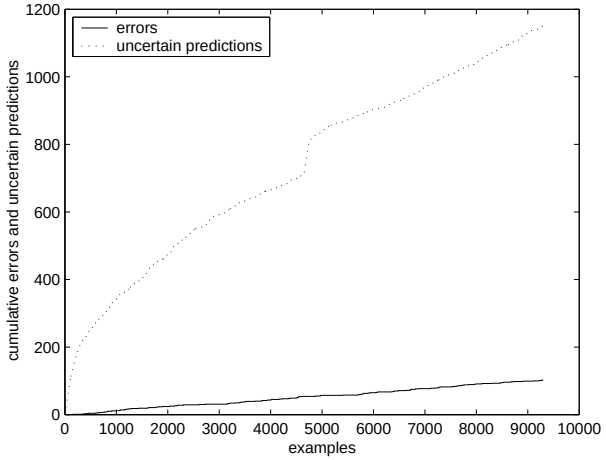

Figure 4: ICM at 99%

This work was partially supported by EPSRC (grant GR/R46670/01), BBSRC (grant 111/BIO14428), and EU (grant IST-1999-10226).

References

[1] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

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

[3] Tom Melluish, Craig Saunders, Ilia Nouretdinov, and Vladimir Vovk. Comparing the Bayes and typicalness frameworks. In L. De Raedt and P. Flash, editors, *Machine Learning: ECML 2001. Proceedings of the 12th European Conference on Machine Learning*, volume 2167 of *Lecture Notes in Artificial Intelligence*, pages 360–371. Springer, 2001.

[4] Ilia Nouretdinov, Tom Melluish, and Vladimir Vovk. Ridge Regression Confidence Machine. In *Proceedings of the 18th International Conference on Machine Learning*, pages 385–392, San Francisco, CA, 2001. Morgan Kaufmann.

[5] Ilia Nouretdinov, Vladimir Vovk, Michael Vyugin, and Alex Gammerman. Pattern recognition and density estimation under the general i.i.d. assumption. In David Helmbold and Bob Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, volume 2111 of *Lecture Notes in Artificial Intelligence*, pages 337–353. Springer, 2001.

[6] Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Qualified predictions for large data sets in the case of pattern recognition. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA'02)*, pages 159–163. CSREA Press, 2002.

[7] Kostas Proedrou, Ilia Nouretdinov, Vladimir Vovk, and Alex Gammerman. Transductive confidence machines for pattern recognition. In *Proceedings of the Thirteenth European Conference on Machine Learning*, 2002.

[8] Craig Saunders, Alex Gammerman, and Vladimir Vovk. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 722–726, 1999.

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

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

[11] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

[12] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[13] Vladimir Vovk. Asymptotic optimality of Transductive Confidence Machine. In *Proceedings of the Thirteenth International Conference on Algorithmic Learning Theory*, 2002. Full version published as Technical Report CLRC-TR-02-02, Computer Learning Research Centre, Department of Computer Science, Royal Holloway, University of London, May 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`.

[14] Vladimir Vovk. Universal well-calibrated algorithm for on-line classification. Technical Report Working Paper 3, The On-line Compression Modelling Project, `http://www.vovk.net/kp`, November 2002.

[15] 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.

A Appendix: Proofs

A.1 Proof of Theorem 2

First we explain the basic idea of the proof. To show that $(\mathrm{err}_1, \dots, \mathrm{err}_N)$ is distributed as B_δ^N (it will be easy to get rid of the assumption of a fixed horizon N), we use the standard idea of reversing the time (see, e.g., the proof of de Finetti's theorem in [9]). We can imagine that the sample $(z_1, \dots, z_N)$ is generated in two steps: first, the bag $\{\!|z_1, \dots, z_N|\!\}$ is generated from some probability distribution (namely, the image of P under the mapping $(z_1, z_2, \dots) \mapsto \{\!|z_1, \dots, z_N|\!\}$), and then the actual sample $(z_1, \dots, z_N)$ is chosen randomly from the set of all orderings of the bag $\{\!|z_1, \dots, z_N|\!\}$. Already the second step ensures that, conditionally on knowing $\{\!|z_1, \dots, z_N|\!\}$ (and, therefore, unconditionally), the sequence $(\mathrm{err}_N, \dots, \mathrm{err}_1)$ is distributed as B_δ^N. Indeed, roughly speaking (i.e., ignoring ties and borderline effects), err_N will be 1 if α_N is among the $N\delta$ largest α_i, and the probability of this is δ since all permutations are equiprobable; when z_N is disclosed, the value err_N will be settled; conditionally on knowing $\{\!|z_1, \dots, z_N|\!\}$ and z_N (and, therefore, knowing $\{\!|z_1, \dots, z_{N-1}|\!\}$), err_{N-1} will also be 1 with probability δ, and so on.

We start the proof by giving some preliminary definitions. The σ-algebra $\mathcal{G}_n$, $n = 0, 1, 2, \dots$, is the collection of all measurable sets $E \subseteq \mathbf{Z}^\infty$ which satisfy

$$(z_1, z_2, \dots) \in E \implies$$
$$(z_{\pi(1)}, \dots, z_{\pi(n)}, z_{n+1}, z_{n+2}, \dots) \in E$$

for any permutation π of $\{1, \dots, n\}$. In particular, $\mathcal{G}_0$ (the most informative σ-algebra) coincides with the original σ-algebra on $\mathbf{Z}^\infty$; $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \cdots$. We will use the notation $\mathbb{E}_\mathcal{F}$ for the conditional expectation w.r. to a σ-algebra $\mathcal{F}$; if necessary, the underlying probability distribution will be given as an

upper index. Similarly, $\mathbb{P}_{\mathcal{F}}$ will stand for the conditional probability w.r. to $\mathcal{F}$.

Fix a TCM Γ and a confidence level $1-\delta$; these elements will usually be left implicit in our notation. The proof will be based on the following lemma.

Lemma 1 *For any trial n,*

$$\mathbb{E}_{\mathcal{G}_n}^{P\times U^\infty}(\mathrm{err}_n) = \delta. \tag{23}$$

Proof Define

$$\mathrm{error}(z_1,\dots,z_n) := \mathrm{err}_n(\Gamma_{1-\delta},\omega),$$

where $\omega \in \mathbf{Z}^\infty$ is any continuation of the sequence $(z_1,\dots,z_n)$; we do not reflect in the notation the dependence on the random numbers $(\tau_1,\tau_2,\dots) \in [0,1]^\infty$. First we prove that, for any $(z_1,\dots,z_n)$,

$$\frac{1}{n!}\sum_{\pi}\mathbb{E}\,\mathrm{error}(z_{\pi(1)},\dots,z_{\pi(n)}) = \delta, \tag{24}$$

where π ranges over all permutations of $\{1,\dots,n\}$ and $\mathbb{E}$ stands for the expected value w.r. to the distribution U^∞; it is intuitively clear, and will be easy to show formally, that (24) implies (23).

Let us fix a sequence $(z_1,\dots,z_n) \in \mathbf{Z}^n$. As usual we denote by α_i the result of applying the individual strangeness measure underlying Γ to the given examples: $(\alpha_1,\dots,\alpha_n) = A_n(z_1,\dots,z_n)$. For every $i = 1,\dots,n$ define

$$p_i^+ := \frac{\#\{j=1,\dots,n : \alpha_j \ge \alpha_i\}}{n},$$
$$p_i^- := \frac{\#\{j=1,\dots,n : \alpha_j > \alpha_i\}}{n}.$$

It is clear that $p_i^- < p_i^+$ and

$$p_i^+ - p_i^- = \frac{\#\{j : \alpha_j = \alpha_i\}}{n}.$$

Notice that the semi-closed intervals $[p_i^-, p_i^+)$, $i = 1,\dots,n$, either coincide or are disjoint; it is also easy to see that they "lie next to each other", in the sense that their union is also a semi-closed interval (namely, $[0,1)$).

Let us say that an example z_i (more accurately, its index i) is

- *strange* if $p_i^+ \le \delta$

- *ordinary* if $p_i^- > \delta$
- *borderline* if $p_i^- \le \delta < p_i^+$.

We will use the notation $p^- := p_i^-$ and $p^+ := p_i^+$ where i is the index of any borderline example. Notice that the fraction of strange examples (i.e., the number of strange examples divided by n) is p^-, the fraction of ordinary examples is $1 - p^+$, and the fraction of borderline examples is $p^+ - p^-$.

By the definition of rTCM, $\text{error}(z_{\pi(1)}, \dots, z_{\pi(n)})$ is 1 if the last example $z_{\pi(n)}$ is strange, is 0 if the last example is ordinary, and is 0 with probability

$$\frac{p^+ - \delta}{p^+ - p^-} \tag{25}$$

(cf. (16)) if the last example is borderline. Therefore, the expected value of $\text{error}(z_{\pi(1)}, \dots, z_{\pi(n)})$ over the random numbers $(\tau_1 \tau_2 \dots) \in [0,1]^\infty$ and equiprobable permutations π of $\{1, \dots, n\}$ (in other words, the left-hand side of (24)) is

$$p^- + (p^+ - p^-)\frac{\delta - p^-}{p^+ - p^-} = \delta. \tag{26}$$

This proves (24).

It remains to prove that (24) implies (23). Let us assume (24). We say that a set $E \subseteq \mathbf{Z}^n$ is *symmetric* if

$$(z_1, \dots, z_n) \in E \Longrightarrow (z_{\pi(1)}, \dots, z_{\pi(n)}) \in E$$

for any permutation π of $\{1, \dots, n\}$. We are required to prove that

$$\int_E (\mathbb{E}\,\text{error}(z_1, \dots, z_n) - \delta)\, dP = 0 \tag{27}$$

for any symmetric measurable set $E \subseteq \mathbf{Z}^n$.

First we notice that, if $G : \Omega \to \Omega$ is a bijection defined on a measurable space Ω and measurable in both directions, then for every measurable function $f : \Omega \to \mathbb{R}$, measurable set $E \subseteq \Omega$, and measure P on Ω,

$$\int_E f dP = \int_{E'} f' dP', \tag{28}$$

where the set E', function f', and measure P' are defined by

$$E' := G^{-1}(E),\ f'(\omega) := f(G(\omega)),\ P'(A) := P(G(A)).$$

Applying this to $\Omega := \mathbf{Z}^n$,

$$G(z_1, \dots, z_n) := (z_{\pi(1)}, \dots, z_{\pi(n)}),$$

and

$$f(z_1, \dots, z_n) := \mathbb{E}\,\mathrm{error}(z_1, \dots, z_n) - \delta, \tag{29}$$

where π is a permutation, we obtain

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1, \dots, z_n) - \delta)\, dP = \int_E \left(\mathbb{E}\,\mathrm{error}(z_{\pi(1)}, \dots, z_{\pi(n)}) - \delta\right) dP$$

(remember that $E' = E$ and $P' = P$) and so, from (24),

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1, \dots, z_n) - \delta)\, dP = \frac{1}{n!} \sum_\pi \int_E \left(\mathbb{E}\,\mathrm{error}(z_{\pi(1)}, \dots, z_{\pi(n)}) - \delta\right) dP = 0. \qquad \blacksquare$$

The other basic result that we will need is the following simple lemma.

Lemma 2 *For any trial $n \geq 1$, err_n is $\mathcal{G}_{n-1}$-measurable.*

Proof Fix a trial n. We are required to prove that the event $\{\mathrm{err}_n = 1\}$ is $\mathcal{G}_{n-1}$-measurable, i.e., invariant w.r. to permutations of the first $n-1$ examples. By the definition, (6), this follows from the invariance of $\Gamma_{1-\delta}(z_1, \dots, z_{n-1}, x_n)$ w.r. to permutations of the first $n-1$ examples, which, in its turn, follows (see (10) and (11)) from the invariance of the underlying individual strangeness measure (see (7)). $\blacksquare$

The proof of Theorem 2 will use the following properties of conditional expectations (see, e.g., [10], §II.7.4):

A. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, ξ and η are bounded $\mathcal{F}$-measurable random variables, and η is $\mathcal{G}$-measurable, $\mathbb{E}_\mathcal{G}(\xi\eta) = \eta\, \mathbb{E}_\mathcal{G}(\xi)$ a.s.

B. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, and ξ is a random variable, $\mathbb{E}_\mathcal{G}(\mathbb{E}_\mathcal{F}(\xi)) = \mathbb{E}_\mathcal{G}(\xi)$ a.s.; in particular, $\mathbb{E}(\mathbb{E}_\mathcal{F}(\xi)) = \mathbb{E}(\xi)$.

Fix temporarily positive integer N. To simplify the formulas, we will often omit curly braces. First we prove that, for any $n = 1, \dots, N$,

$$\mathbb{P}_{\mathcal{G}_n}((\mathrm{err}_n, \dots, \mathrm{err}_1) = (\omega_n, \dots, \omega_1)) = \delta^k (1-\delta)^{n-k}, \tag{30}$$

where $\mathbb{P}$ refers to the probability distribution $P \times U^\infty$ and k is the number of 1s in $(\omega_n, \dots, \omega_1)$. The proof is by induction in n. For $n = 1$, (30) immediately follows from Lemma 1. For $n > 1$ we obtain, making use of Lemmas 1 and 2, properties A and B of conditional expectations, and the inductive assumption:

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}_n}((\mathrm{err}_n, \dots, \mathrm{err}_1) = (\omega_n, \dots, \omega_1)) \\
&= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{\mathrm{err}_n=\omega_n}\mathbb{I}_{(\mathrm{err}_{n-1},\dots,\mathrm{err}_1)=(\omega_{n-1},\dots,\omega_1)}\right)\right) \\
&= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{I}_{\mathrm{err}_n=\omega_n}\,\mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{(\mathrm{err}_{n-1},\dots,\mathrm{err}_1)=(\omega_{n-1},\dots,\omega_1)}\right)\right) \\
&= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{I}_{\mathrm{err}_n=\omega_n}\delta^{k^\dagger}(1-\delta)^{(n-1)-k^\dagger}\right) = \delta^k(1-\delta)^{n-k}
\end{aligned}$$

almost surely, where $\mathbb{I}_E$ means the indicator of E, k and $k^\dagger$ are the number of 1s in $(\omega_n, \dots, \omega_1)$ and $(\omega_{n-1}, \dots, \omega_1)$, respectively, and the expected value $\mathbb{E}$ is taken over $\mathbb{P} = P \times U^\infty$.

By property B, (30) immediately implies

$$\mathbb{P}\left((\mathrm{err}_N, \dots, \mathrm{err}_1) = (\omega_N, \dots, \omega_1)\right) = \delta^k(1-\delta)^{N-k},$$

where k is the number of 1s in $(\omega_N \dots \omega_1)$. Therefore, we have proved that the distribution of the random sequence $\mathrm{err} \in \{0,1\}^\infty$ coincides with B_δ^∞ on the σ-algebra $\mathcal{F}_N$ generated by the events $\{(\omega_1, \omega_2, \dots) \in \{0,1\}^\infty : \omega_i = 1\}$, $i = 1, \dots, N$. It is well known (see, e.g., [10], Theorem II.3.3) that this implies that the distribution of err coincides with B_δ^∞ on all measurable sets in $\{0,1\}^\infty$.

A.2 Precise statement of Theorem 3

First we define rICM formally. The rICM based on D, Δ, and the parametric sequence $m_1, m_2, \dots$ (as before, m_i are required to satisfy $m_1 < m_2 < \cdots$) is defined to be the randomised region predictor Γ such that the predictive region $\Gamma_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, \tau_n)$ is computed as follows:

- if $n \le m_1$, the predictive region is found using a fixed rTCM;

- otherwise, find the k such that $m_k < n \le m_{k+1}$, and, for each label $y \in \mathbf{Y}$,
  - include y in the predictive region if
$$\frac{\#\{j = m_k + 1, \dots, n : \alpha_j > \alpha_n\}}{n - m_k} > \delta,$$
  - do not include if
$$\frac{\#\{j = m_k + 1, \dots, n : \alpha_j \ge \alpha_n\}}{n - m_k} \le \delta,$$
  - if neither condition is satisfied, include y in the predictive region if and only if
$$\tau_n < \frac{\#\{j : \alpha_j \ge \alpha_n\} - (n - m_k)\delta}{\#\{j : \alpha_j = \alpha_n\}} \tag{31}$$
    j ranging over $\{m_k + 1, \dots, n\}$,

  where $\alpha_{m_k+1}, \dots, \alpha_n$ are defined by (22).

The main assertion made in Theorem 3 is that the image of the probability distribution $P \times U^\infty$ under the mapping

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \mapsto \mathrm{err}(\Gamma_{1-\delta}, \omega, \tau)$$

is B_δ^∞ for any rICM Γ.

Remark In our definitions of rTCM and rICM we assumed that the same random number τ_n is used for every potential label y of x_n. In fact, assuming $\mathbf{Y}$ is finite, we can also use a separate random number τ_n^y for each $y \in \mathbf{Y}$, with the random numbers τ_n^y, $n = 1, 2, \dots$, $y \in \mathbf{Y}$, independent. On the other hand, an arbitrary correlation between τ_n^y, $y \in \mathbf{Y}$, can be allowed; Theorems 2 and 3 will continue to hold as long as the random numbers $\tau_n^{y_n}$, $n = 1, 2, \dots$, are independent.

A.3 Proof of Theorem 3

Let us set $m_0 := 0$. In this proof, the σ-algebra $\mathcal{G}_n$, $n = 1, 2, \ldots$, is the collection of all measurable sets $E \subseteq \mathbf{Z}^\infty$ which satisfy the following:

- if k is the largest non-negative integer such that $m_k < n$,

$$\begin{gathered}(z_1, z_2, \ldots) \in E \Longrightarrow \\ (z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}, \\ z_{n+1}, z_{n+2}, \ldots) \in E\end{gathered}$$

  for any permutation π of $\{m_k + 1, \ldots, n\}$;

- if k is any positive integer such that $m_k < n$,

$$\begin{gathered}(z_1, z_2, \ldots) \in E \Longrightarrow \\ (z_1, \ldots, z_{m_{k-1}}, z_{\pi(m_{k-1}+1)}, \ldots, z_{\pi(m_k)}, \\ z_{m_k+1}, z_{m_k+2}, \ldots) \in E\end{gathered}$$

  for any permutation π of $\{m_{k-1} + 1, \ldots, m_k\}$.

As before, $\mathcal{G}_0$ is the original σ-algebra on $\mathbf{Z}^\infty$. It is obvious that $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \cdots$.

The proof of Theorem 2 obviously works for ICMs as well, with the possible exception of Lemma 1. Therefore, in this subsection we will only show how to prove the following analogue of Lemma 1.

Lemma 3 *Any rICM satisfies (23), for any confidence level $1 - \delta$ and any trial n.*

Proof If $n \le m_1$, an rTCM is used and so we have nothing to prove; we will assume $n > m_1$. The proof is parallel to the proof of Lemma 1.

Analogously to the reduction of the proof of Lemma 1 to proving (24) for any $(z_1, \ldots, z_n)$, we first prove that, for any $(z_1, \ldots, z_n)$,

$$\begin{aligned}\frac{1}{(n - m_k)!} \sum_\pi \mathbb{E}\, \mathrm{error}(z_1, \ldots, z_{m_k}, \\ z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) = \delta,\end{aligned} \tag{32}$$

where k is the largest positive integer for which $m_k < n$ and π ranges over all permutations of $\{m_k + 1, \dots, n\}$. Define α_j, $j = m_k + 1, \dots, n$, by the formula

$$\alpha_j := \Delta\left(y_j, D_{\wr(x_1,y_1),\dots,(x_{m_k},y_{m_k})\int}(x_j)\right)$$

(cf. (22)). For every $i = m_k + 1, \dots, n$ define

$$p_i^+ := \frac{\#\{j = m_k + 1, \dots, n : \alpha_j \geq \alpha_i\}}{n - m_k},$$
$$p_i^- := \frac{\#\{j = m_k + 1, \dots, n : \alpha_j > \alpha_i\}}{n - m_k}.$$

It is clear that $p_i^- < p_i^+$ and

$$p_i^+ - p_i^- = \frac{\#\{j = m_k + 1, \dots, n : \alpha_j = \alpha_i\}}{n - m_k}.$$

Again the semi-closed intervals $[p_i^-, p_i^+)$, $i = m_k + 1, \dots, n$, either coincide or are disjoint; they also "lie next to each other".

Let us say that an example z_i, $i = m_k + 1, \dots, n$, is

- *strange* if $p_i^+ \leq \delta$
- *ordinary* if $p_i^- > \delta$
- *borderline* if $p_i^- \leq \delta < p_i^+$.

We will use the notation $p^- := p_i^-$ and $p^+ := p_i^+$, where i is the index of any borderline example. The fraction of strange examples (i.e., the number of strange examples divided by $n - m_k$) is p^-, the fraction of ordinary examples is $1 - p^+$, and the fraction of borderline examples is $p^+ - p^-$.

By the definition of rICM,

$$\mathrm{error}(z_1, \dots, z_{m_k}, z_{\pi(m_k+1)}, \dots, z_{\pi(n)})$$

is 1 if the last example $z_{\pi(n)}$ is strange, is 0 if the last example is ordinary, and is 0 with probability (25) (cf. (31)) if the last example is borderline. Therefore, the expected value of $\mathrm{error}(z_1, \dots, z_{m_k}, z_{\pi(m_k+1)}, \dots, z_{\pi(n)})$ over all equiprobable permutations π of $\{m_k + 1, \dots, n\}$ and random numbers $(\tau_1\tau_2\dots) \in [0,1]^\infty$ (i.e., the left-hand side of (32)) is (26). This proves (32).

To finish the proof it remains to establish (23). We say that a set $E \subseteq \mathbf{Z}^n$ is *calibration symmetric* if

$$(z_1, \ldots, z_n) \in E \Longrightarrow$$
$$(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) \in E$$

for any permutation π of $\{m_k + 1, \ldots, n\}$. It is sufficient to prove that

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP = 0$$

for any calibration symmetric measurable set $E \subseteq \mathbf{Z}^n$.

Applying (28) to $\Omega := \mathbf{Z}^n$,

$$G(z_1, \ldots, z_n) := (z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}),$$

and (29), where π is a permutation of $\{m_k + 1, \ldots, n\}$, we obtain

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP =$$
$$\int_E \left(\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) - \delta\right) dP$$

(again $E' = E$ and $P' = P$) and so, from (32),

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP =$$
$$\frac{1}{(n - m_k)!} \sum_\pi \int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_{m_k},$$
$$z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) - \delta) dP = 0,$$

π ranging over the permutations of $\{m_k + 1, \ldots, n\}$. ∎

On-line Compression Modelling Project
Working Papers

1. *On-line confidence machines are well-calibrated*, by Vladimir Vovk, April 2002.

2. *Asymptotic optimality of Transductive Confidence Machine*, by Vladimir Vovk, May 2002.

3. *Universal well-calibrated algorithm for on-line classification*, by Vladimir Vovk, November 2002.

4. *Mondrian Confidence Machine*, by Vladimir Vovk, David Lindsay, Ilia Nouretdinov and Alex Gammerman, March 2003.

5. *Testing exchangeability on-line*, by Vladimir Vovk, Ilia Nouretdinov and Alex Gammerman, February 2003.

6. *Criterion of calibration for Transductive Confidence Machine with limited feedback*, by Ilia Nouretdinov and Vladimir Vovk, April 2003.

7. *Online region prediction with real teachers*, by Daniil Ryabko, Vladimir Vovk and Alex Gammerman, March 2003.

8. *Well-calibrated predictions from on-line compression models*, by Vladimir Vovk, April 2003.