Asymptotic optimality of Transductive Confidence Machine

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-02-02

27th May 2002

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

Abstract

Transductive Confidence Machine (TCM) is a way of converting standard machine-learning algorithms into algorithms that output predictive regions rather than point predictions. It has been shown recently that TCM is well-calibrated when used in the on-line mode: at any confidence level $1 - \delta$, the long-run relative frequency of errors is guaranteed not to exceed δ provided the examples are generated independently from the same probability distribution P. Therefore, the number of "uncertain" predictive regions (i.e., those containing more than one label) becomes the sole measure of performance. The main result of this paper is that when the probability distribution P generating the examples is known, it is possible to construct a TCM (guaranteed to be well-calibrated under *any* distribution) that performs asymptotically as well as the best region predictor.

1 Region predictors

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:

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

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 family (indexed by $\gamma \in [0, 1]$) of Predictor's strategies Γ_γ such that $\Gamma_\gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ is a measurable function of Nature's moves and

$$\Gamma_{\gamma_1}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma_{\gamma_2}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$$

when $\gamma_1 \le \gamma_2$. Since we are interested in prediction with confidence, the 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*.

To provide the reader with an intuition about region prediction, we briefly present results of Predictor's particular strategy ("1-Nearest Neighbour TCM"; for details, see [8]) to the USPS data set (see [6], §12.2). Figures 1 and 2 show the cumulative number of errors Err_n plotted against $n = 0, 1, \ldots, 9298$ (solid line), the cumulative number Unc_n of uncertain predictions Γ_n, and that of empty predictions (for which $|\Gamma_n| = 0$). Figure 3 gives the *empirical calibration curve*

$$\delta \mapsto \frac{\mathrm{Err}_N(\mathrm{USPS}, \Gamma_{1-\delta})}{N}$$

and the *empirical performance curve*

$$\delta \mapsto \frac{\mathrm{Unc}_N(\mathrm{USPS}, \Gamma_{1-\delta})}{N}$$

for this region predictor; we use the strategies followed by Nature (the USPS data set) and Predictor (corresponding to significance level δ) as arguments for Err and Unc; $N = 9298$ is the size of the USPS data set.

2 Well-calibrated and asymptotically optimal region predictors

Suppose we know the true distribution P in $\mathbf{Z}$ generating the examples. In this section we will construct a region predictor optimal under P; we will often omit P from our notation.

Let $P_{\mathbf{X}}$ be the marginal distribution of P in $\mathbf{X}$ (i.e., $P_{\mathbf{X}}(E) := P(E \times \mathbf{Y})$) and $P_{\mathbf{Y}\,|\,\mathbf{X}}(y\,|\,x)$ be 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; the existence of regular conditional probability is obvious in this case and also follows from general results, such as [5], Theorem II.7.5). We will often omit subindices $\mathbf{X}$ and $\mathbf{Y}\,|\,\mathbf{X}$.

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 5; the graph of F is the thick line, and the unit box is also shown.

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 $[0, 1] \to [0, 1]$. (Why the terminology introduced here and below is natural will become clear from Theorems 1 and 2.) Geometrically, $\mathbf{S}$ is defined from the graph of F as follows (see Figure 5): move 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*

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

(the area under the thick curve in Figure 5; we will later see that this coincides with what is sometimes called *Bayes error* or *Bayes risk*—see, e.g., [2], §2.1).

So far we have defined some characteristics of the distribution P itself; now we will give definitions related to individual region predictors. The most natural class of region predictors is that of *permutationally invariant region predictors* Γ, for which $\Gamma_{1-\delta}(z_1, \ldots, z_n, x)$ does not depend on the order of $z_1, \ldots, z_n$ (we know the examples are i.i.d., so knowing the order should not help).

The *calibration curve* of a region predictor Γ under P is the following function of the type $[0,1] \to [0,1]$:

$$\mathbf{C}(\delta) := \inf \left\{ \beta : \mathbb{P} \left\{ \limsup_{n\to\infty} \frac{\mathrm{Err}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \beta \right\} = 1 \right\} \tag{2}$$

($\mathbb{P}(E)$ stands for the probability of event E). By the Hewitt–Savage zero-one law (see, e.g., [5], Theorem IV.1.3) in the case of permutationally invariant region predictors this definition will not change if "$= 1$" is replaced by "> 0" in (2). The *performance curve* of Γ under P is defined by

$$\mathbf{P}(\delta) := \inf \left\{ \beta : \mathbb{P} \left\{ \limsup_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \beta \right\} = 1 \right\}; \tag{3}$$

this is again a function of the type $[0,1] \to [0,1]$. The Hewitt–Savage zero-one law again implies that for permutationally invariant Γ this will not change if "$= 1$" is replaced by "> 0".

We will say that a region predictor Γ is *well-calibrated* under P if its calibration curve $\mathbf{C}(\delta)$ is below the diagonal: $\mathbf{C}(\delta) \le \delta$ for any significance level δ. It is *asymptotically optimal* under P if its performance curve coincides with the success curve: $\mathbf{P}(\delta) = \mathbf{S}(\delta)$ for all δ.

Theorem 1 *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, its performance curve $\mathbf{P}$ is above $\mathbf{S}$: for any δ, $\mathbf{P}(\delta) \ge \mathbf{S}(\delta)$. Moreover, for any significance level δ,*

$$\liminf_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \ge \mathbf{S}(\delta) \qquad a.s. \tag{4}$$

Let us now assume, for simplicity, that the distribution P is *regular*, in the sense that the predictability distribution function F is continuous. (The general case will be considered in §5 and will involve randomised region predictors.)

The main result of this paper (Theorem 2, strengthened in Theorem 2r) is that one can construct an asymptotically optimal TCM (which is well-calibrated automatically, by [8]). If, however, we know for sure that the true distribution is P it is very easy to construct a well-calibrated and asymptotically optimal region predictor. Fix a *choice function* $\hat{y} : \mathbf{X} \to \mathbf{Y}$ such that

$$\forall x \in \mathbf{X} : f(x) = P(\hat{y}(x) \mid x)$$

(to put it differently, $\hat{y}(x) \in \arg\max_y P(y \mid x)$). Define the *$P$-Bayesian* region predictor Γ by

$$\Gamma_{1-\delta}(z_1, \ldots, z_n, x) := \begin{cases} \{\hat{y}(x)\} & \text{if } F(f(x)) \geq \mathbf{S}(\delta) \\ \mathbf{Y} & \text{otherwise,} \end{cases}$$

for all significance levels δ and data sequences $(z_1, \ldots, z_n, x) \in \mathbf{Z}^n \times \mathbf{X}$, $n = 0, 1, \ldots$. It can be shown that the P-Bayesian region predictor is well-calibrated and asymptotically optimal under P. (Our definition of the P-Bayesian region predictor is arbitrary in several respects; in principle, different choice functions can be used at different trials, the prediction can be arbitrary when $F(f(x)) = \mathbf{S}(x)$, and $\mathbf{Y}$ can be replaced by any $E \subseteq \mathbf{Y}$ such that $P(E \mid x) := \sum_{y \in E} P(y \mid x) = 1$.)

The critical significance level (1) is an important characteristic of the probability distribution P generating the examples. If $\delta > \delta_0$, an optimal region predictor will always output certain predictions and, if forced to achieve the error rate δ, will sometimes have to output empty predictions. If, on the other hand, $\delta < \delta_0$, there will be uncertain predictions but no empty predictions. Figures 1 and 2 suggest that the critical significance level for the USPS data set is between 1% and 5%. This agrees with the observation that the critical significance level is just the error rate of the Bayesian point predictor (which is restricted to outputting Γ_n with $|\Gamma_n| = 1$ and minimises the expected number of errors) and the fact (reported in [6]) that the error rate achieved by humans on the USPS data set is 2.5%. Notice that in Figure 1 the onset of empty predictions closely follows the point where all predictions become certain; see also Figures 6, 7 and 8.

3 Transductive Confidence Machine

The procedure at the end of §2 works well when P is known. If, however, P is only a convenient benchmark, using the Bayesian region predictor can lead to misleading results (see, e.g., [4]). In the rest of this paper we will discuss how to ensure well-calibratedness under any distribution P in $\mathbf{Z}$ without losing the asymptotic performance of the Bayesian predictor.

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 $A_n : \mathbf{Z}^n \to \mathbb{R}^n$ 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

$$A_n : (z_1, \ldots, z_n) \mapsto (\alpha_1, \ldots, \alpha_n) \tag{5}$$

is determined by z_i and the bag $\lbag z_1, \ldots, z_n \rbag$. (The difference between the bag $\lbag z_1, \ldots, z_n \rbag$ and the set $\{z_1, \ldots, z_n\}$ 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: $\Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ is defined to be the set of all labels $y \in \mathbf{Y}$ such that

$$\frac{\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\}}{n} > \delta, \tag{6}$$

where

$$(\alpha_1, \ldots, \alpha_n) := A_n((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_n, y)). \tag{7}$$

In general, a *TCM* is the TCM associated with some individual strangeness measure. It is shown in [8] that TCM is well-calibrated under any P.

4 Asymptotically optimal TCM

If we know (or suspect) that the probability distribution in $\mathbf{Z}$ generating the examples is P, we can define the individual strangeness measure (5) by

$$\alpha_i := \begin{cases} 0 & \text{if } y_i = \hat{y}(x_i) \\ P(\hat{y}(x_i) \mid x_i) & \text{otherwise.} \end{cases} \tag{8}$$

The corresponding TCM will be called the *P-TCM* (cf. [7]). We say that a region predictor is *universally well-calibrated* if it is well-calibrated under any probability distribution P in $\mathbf{Z}$.

Theorem 2 *Let P be a regular probability distribution in $\mathbf{Z}$. The P-TCM is universally well-calibrated and asymptotically optimal under P.*

5 Randomised region predictors

In this section we will remove the assumption that the probability distribution P generating examples is regular. The price we will have to pay is that we will have to generalise the notion of region predictor in general, and TCM in particular, to allow using a generator of random numbers (as was done in [8] as well).

The generator that we consider generates a sequence τ_n, $n = 1, 2, \ldots$, of uniformly distributed independent random numbers in the interval $[0,1]$; τ_n will be used by Predictor at trial n of our basic protocol (see §1). Formally, a *randomised region predictor* Γ is a family, indexed by $n = 1, 2, \ldots$ and $\gamma \in [0,1]$, of measurable functions $\Gamma_\gamma(z_1, \tau_1, \ldots, z_{n-1}, \tau_{n-1}, x_n, \tau_n)$, where the $z_i \in \mathbf{Z}$, $i = 1, \ldots, n-1$, are examples, $\tau_i \in [0,1]$, $i = 1, \ldots, n$, and $x_n \in \mathbf{X}$ is an object, which satisfies

$$\Gamma_{\gamma_1}(z_1, \tau_1, \ldots, z_{n-1}, \tau_{n-1}, x_n, \tau_n) \subseteq \Gamma_{\gamma_2}(z_1, \tau_1, \ldots, z_{n-1}, \tau_{n-1}, x_n, \tau_n)$$

whenever $\gamma_1 \le \gamma_2$. The notation err_n, unc_n, etc., will be continued to be used in the randomised case as well; it should be remembered that these now depend on $\tau_1, \tau_2, \ldots$. We can strengthen Theorem 1 as follows.

Theorem 1r *Let P be a probability distribution in $\mathbf{Z}$ with the success curve $\mathbf{S}$. If a randomised region predictor Γ is well-calibrated under P, then for any significance level δ,*

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

The "a.s." in this theorem refers to the probability distribution $(P \times \mathbf{U})^\infty$ generating the sequence $z_1, \tau_1, z_2, \tau_2, \ldots$, with $\mathbf{U}$ standing for the uniform distribution in $[0,1]$.

Next we introduce a randomised version of TCM. The *randomised Transductive Confidence Machine (rTCM) associated with an individual strangeness measure* A_n is the following randomised region predictor $\Gamma_{1-\delta}$: at any trial n and for any label $y \in \mathbf{Y}$,

1. if $\#\{i = 1, \ldots, n : \alpha_i > \alpha_n\}/n > \delta$ (as before, the αs are defined by (7)), the label y is included in $\Gamma_{1-\delta}$;
2. if $\#\{i = 1, \ldots, n : \alpha_i \ge \alpha_n\}/n \le \delta$, y is not included in $\Gamma_{1-\delta}$;
3. otherwise, y is included in $\Gamma_{1-\delta}$ if

$$\tau_n < \frac{\#\{i = 1, \ldots, n : \alpha_i \ge \alpha_n\} - n\delta}{\#\{i = 1, \ldots, n : \alpha_i = \alpha_n\}}. \tag{9}$$

We say that a randomised region predictor Γ is *perfectly calibrated* if, under any probability distribution P, its calibration curve $\mathbf{C}(\delta)$ coincides with the diagonal: $\mathbf{C}(\delta) = \delta$ for any significance level δ. (Formally, this definition can also be given for deterministic predictors as well, but it would be impossible to satisfy in some cases.) The P-rTCM is defined to be the rTCM associated with the individual strangeness measure (8).

Theorem 2r *Let P be a probability distribution in $\mathbf{Z}$. The P-rTCM is perfectly calibrated and, under P, asymptotically optimal.*

Notice that the rTCM makes at least as many errors as the TCM associated with the same individual strangeness measure. It is shown in [8] that rTCM's errors are independent and happen with probability δ at any confidence level $1-\delta$. The difference between TCM and rTCM is typically negligible after the first several hundred trials; cf. the dotted line in Figure 3.

6 Conclusion

In this paper we defined two desiderata for region predictors: being well-calibrated and asymptotic optimality. Being well-calibrated is the first priority: without it, the meaning of confidence levels is lost and it does not make sense to talk about optimality. If the probability distribution P generating individual examples is known, the Bayesian region predictor is well-calibrated and asymptotically optimal. But even in the situation where P is just a convenient guess that we are unwilling to take too seriously, there exists a region predictor (a TCM) which is universally well-calibrated and asymptotically optimal under P.

Three important directions in which the results of this paper can be extended are:

- To complement our asymptotic results with inequalities that would hold for finite numbers of trials.
- In the case of regression (in which $\mathbf{Y}$ is the real line) we cannot speak about uncertain predictions, and one would have to consider, e.g., the width of predictive regions instead.
- In this paper we assumed that the probability distribution P generating individual examples independently of each other is known. A typical Bayesian assumption, however, is different: it is assumed that the sequence $z_1, z_2, \ldots$ of examples is produced by the mixture of a known family $\{P_\theta : \theta \in \Theta\}$ of power distributions with the weights given by some prior distribution $\mu(d\theta)$ in Θ. Such mixtures are not power distributions anymore, although they are still exchangeable (and, by de Finetti's theorem, all exchangeable distributions are mixtures of power distributions). It would be very interesting, therefore, to extend this paper's results to more general exchangeable distributions. This should not be difficult in the case of "regular" statistical models $\{P_\theta : \theta \in \Theta\}$ with finite-dimensional Θ, since well-known results of Bayesian statistics assert that, after sufficiently many examples are observed, the conditional distribution of the future examples under the mixture approaches one of the P_θ. Ideal results, however, would be about completely general exchangeable distributions.

Acknowledgments

I am grateful to Alex Gammerman for several helpful suggestions.

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] Nicolas Bourbaki. *Eléments de mathématique, Livre IV, Fonctions d'une variable réelle (théorie élémentaire).* Hermann, Paris, second edition, 1958.

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

[3] E. L. Lehmann. *Testing Statistical Hypotheses.* Wiley, New York, 1959.

[4] 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, 2001. Full version published as Technical Report CLRC-TR-01-05; see `http://www.clrc.rhul.ac.uk`.

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

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

[7] Vladimir Vovk. De-Bayesing Bayesian prediction algorithms. Manuscript, June 2000.

[8] Vladimir Vovk. On-line Confidence Machines are well-calibrated. Technical Report CLRC-TR-02-01, Computer Learning Research Centre, Department of Computer Science, Royal Holloway, University of London, `http://www.clrc.rhul.ac.uk`, April 2002. Additional information can be found at `http://www.cs.rhul.ac.uk/~vovk/cm`.

Appendix A: Figures

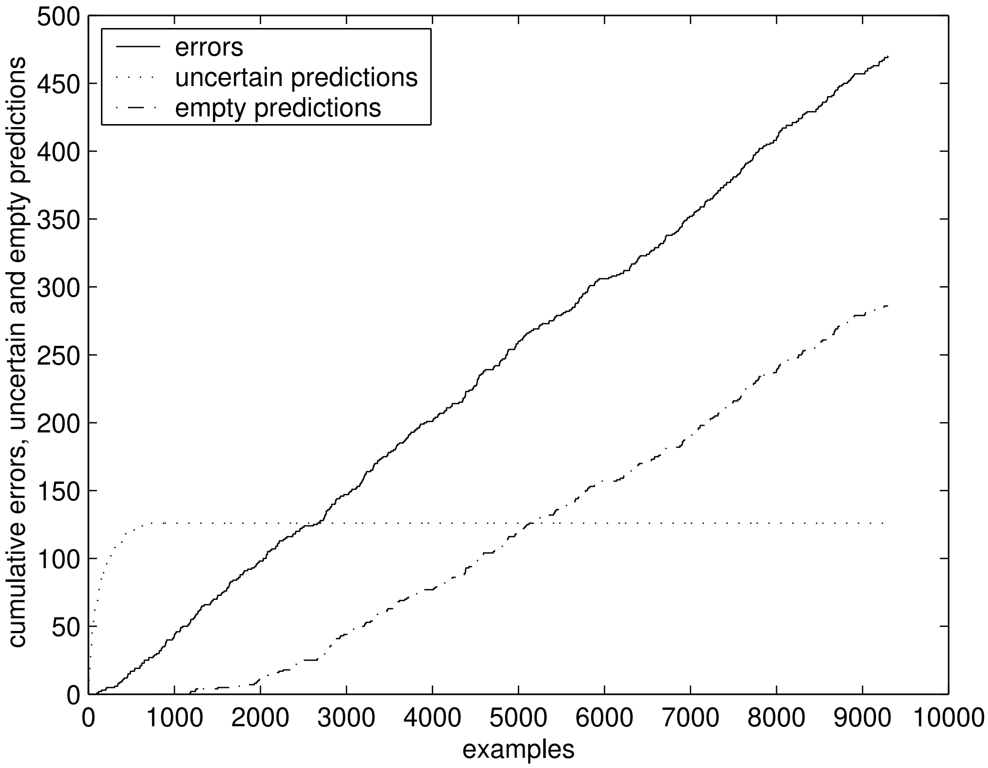

Figure 1 On-line performance of the Nearest Neighbour TCM on the USPS data set (9298 hand-written digits, randomly permuted) for the confidence level 95%. The solid line shows the cumulative number of errors, dotted the cumulative number of uncertain predictions, and dashdot the cumulative number of empty predictions (inevitably leading to an error). This and next figures are not significantly affected by statistical variation (due to the random choice of the permutation of the data set).

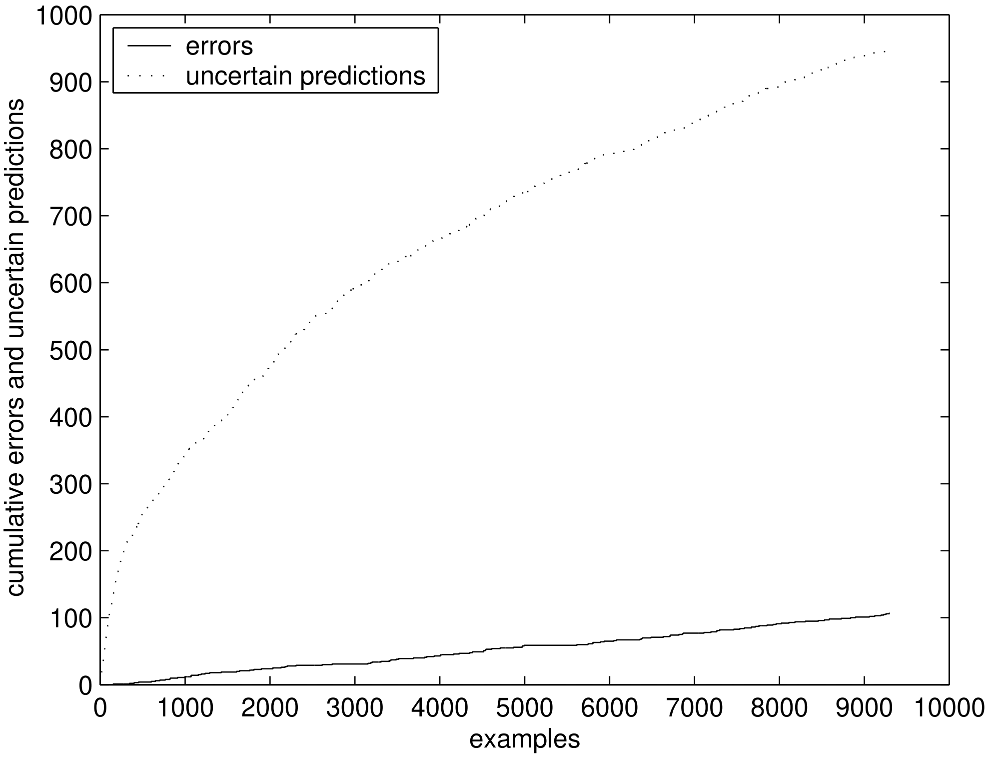

Figure 2 On-line performance of the Nearest Neighbour TCM on the USPS data set for the confidence level 99%. The dashdot line coincides with the horizontal axis (there are no empty predictions) and so is invisible.

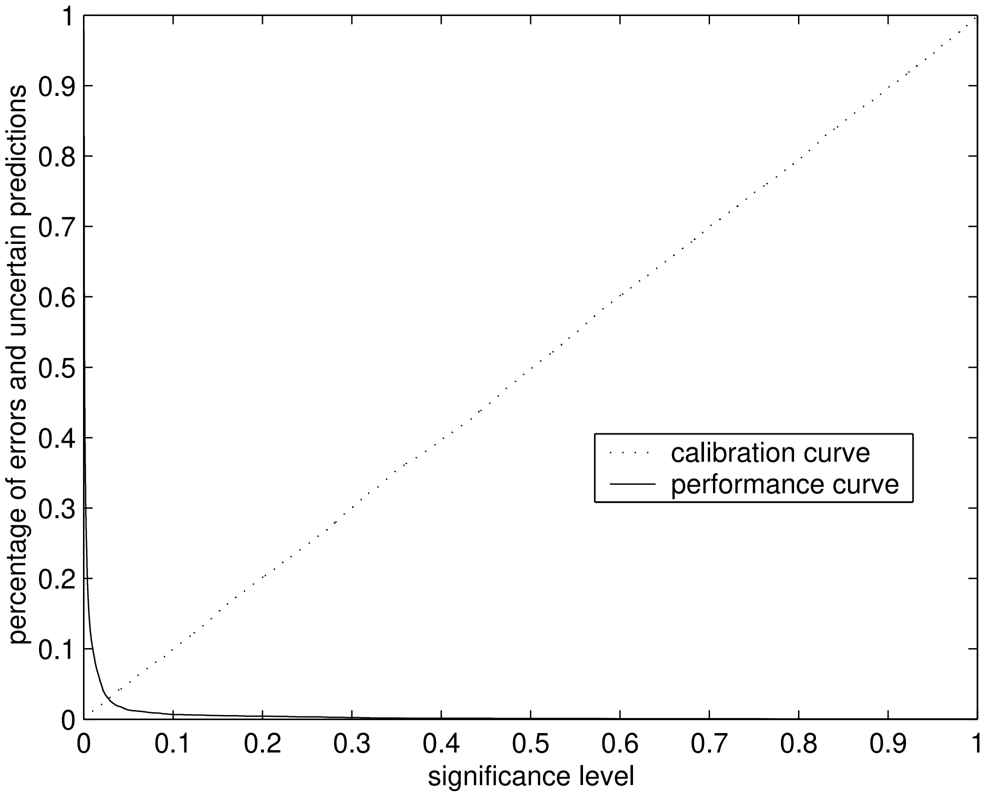

Figure 3 The empirical calibration and performance curves for the Nearest Neighbour TCM on the USPS data set.

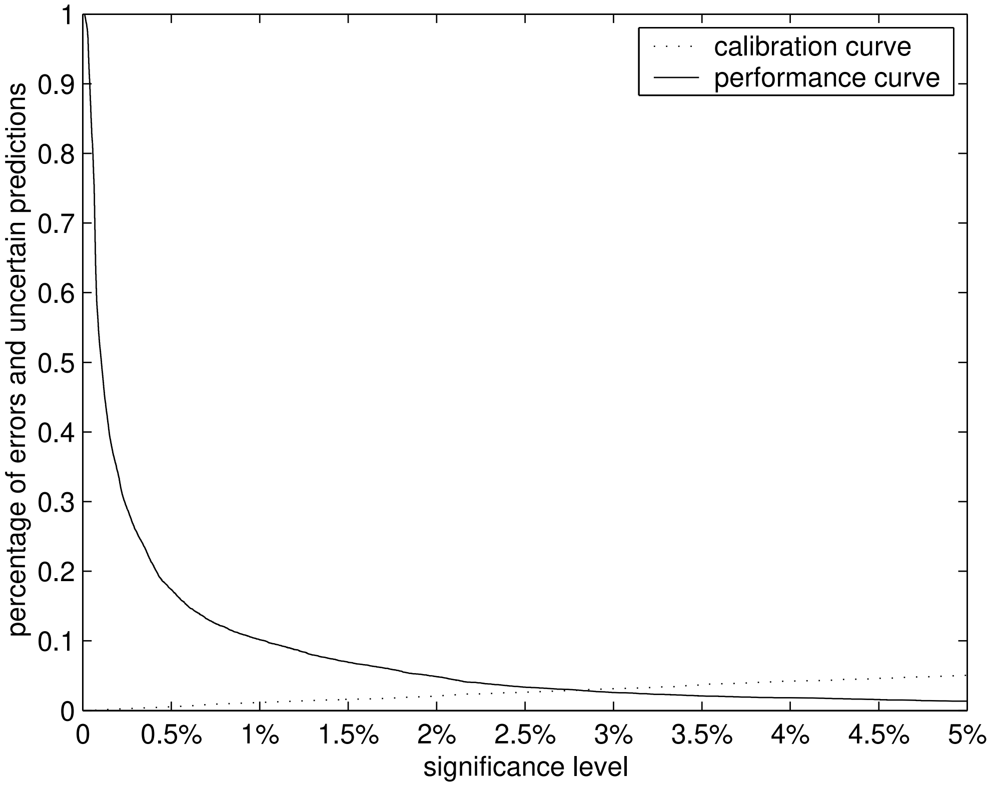

Figure 4 The left edge of the previous figure stretched horizontally.

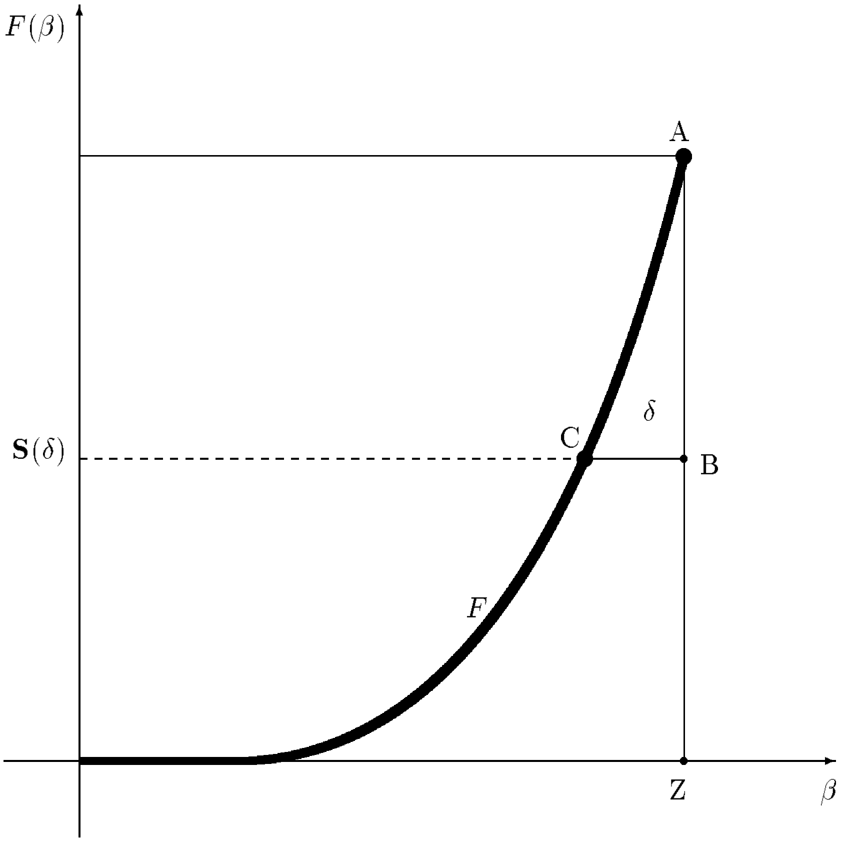

Figure 5 The predictability distribution function F. The function F is non-decreasing, continuous on the right, and $F(1/|\mathbf{Y}|) = 0$. For a possibly more realistic example of a predictability distribution function, see Figure 9.

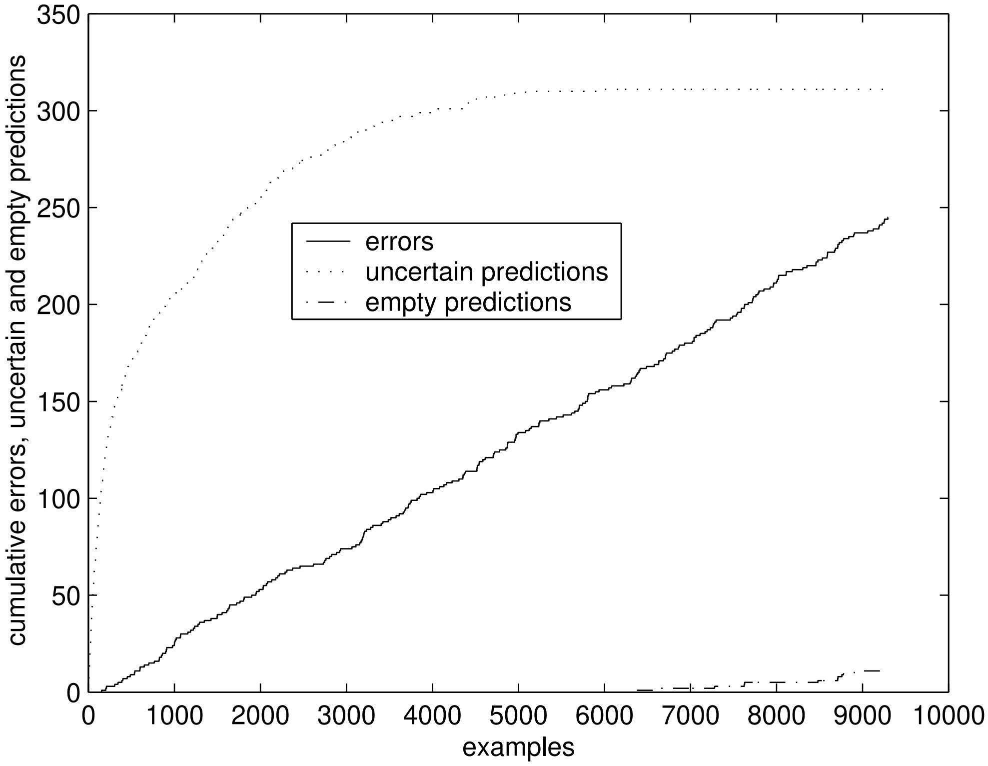

Figure 6 On-line performance of the Nearest Neighbour TCM on the USPS data set for the confidence level 97.5%. (This figure cannot be directly compared to the error rate of 2.5% for humans reported in [6], since this experiment has been carried out on the randomly permuted data set, whereas the test part of the USPS data set is known to be especially hard.)

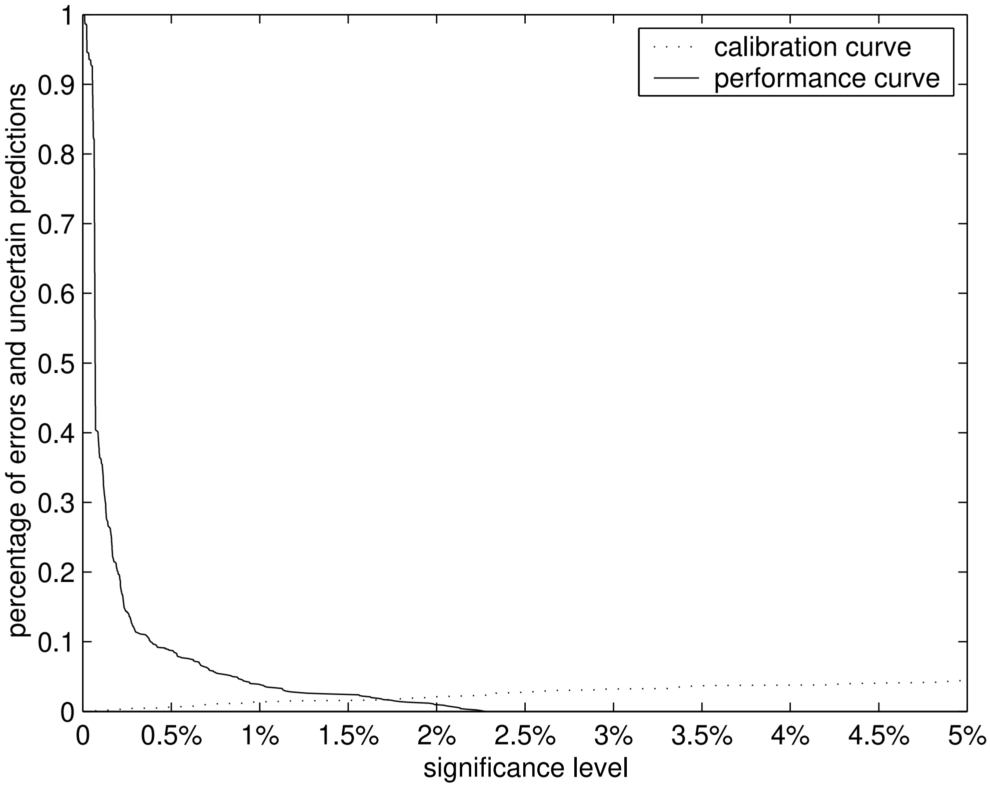

Figure 7 Picture analogous to Figure 4 for the last one thousand examples. Notice a different behaviour of the empirical performance curve as it approaches the horizontal axis as compared with Figure 4. The unexpected behaviour of the empirical performance curve as it approaches the vertical axis may be explained (at least partially) by the "granularity" of TCM: for example, the "realised p-value" given by the left-hand side of (6) can never be less than $1/9298 > 0.01\%$; this behaviour may become more regular for randomised TCM.

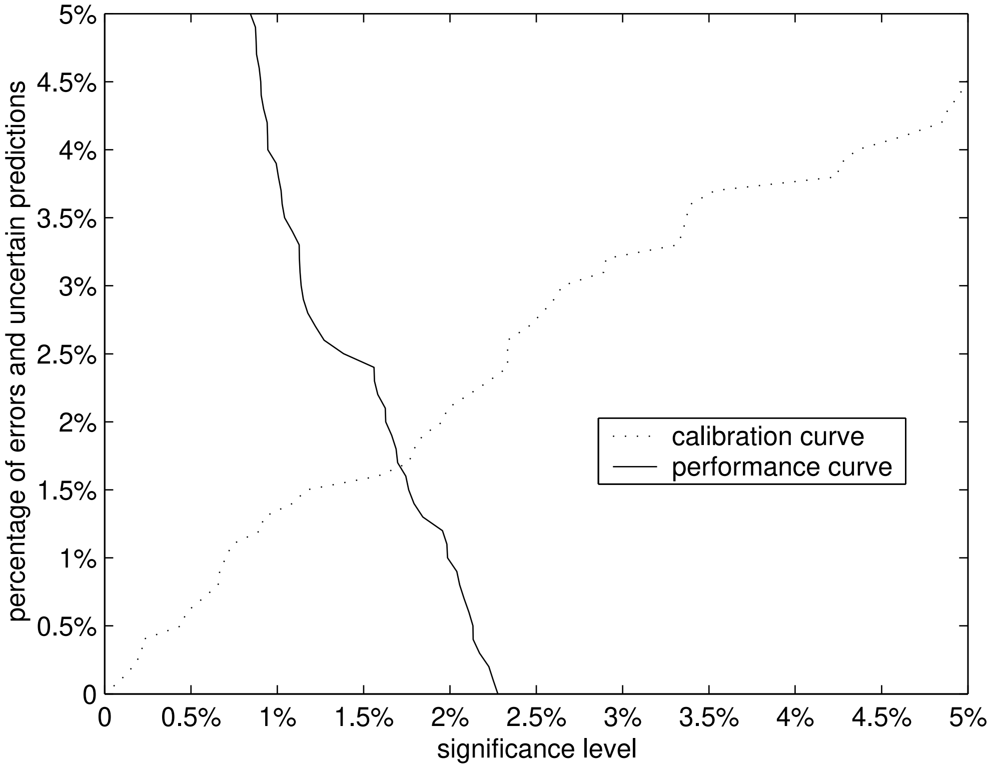

Figure 8 The bottom part of Figure 7 stretched vertically. Notice that the slope of the empirical performance curve is at least 1 in absolute value before it hits the horizontal axis; this agrees with Lemma 2 on p. 18. This figure suggests that, if the 1-NN TCM were an optimal region predictor, the critical significance level for the USPS data set would be close to 2.3.

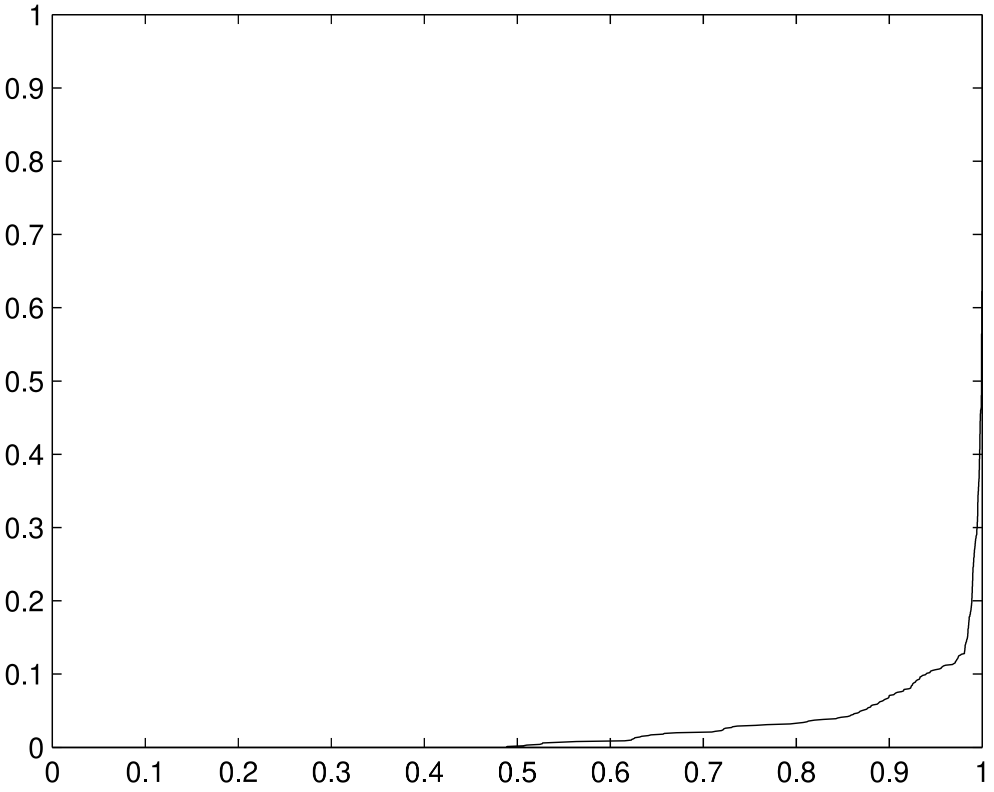

Figure 9 An attempt to reverse engineer the predictability distribution function of the hand-written digits in the USPS data set. This picture was obtained from the solid line in Figure 7 by reversing the list in the proof of Lemma 2.

Appendix B: Proofs

Properties of the predictability distribution function and success curve

Lemma 1 *The predictability distribution function F satisfies the following properties:*

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

If a function $F : [0,1] \to [0,1]$ 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 F is the predictability distribution function.

Proof Properties 1 (cf. the caption to Figure 5), 2, and 3 are obvious (and the last two are well-known properties of all distribution functions). The fact that these three properties characterise predictability distribution functions easily follows from the fact that the last two properties plus $F(-\infty) = 0$ and $F(\infty) = 1$ characterise distribution functions (see, e.g., [5], Theorem II.3.1). ∎

We will use the notations g'_{left} and g'_{right} for the left and right derivatives, respectively, of a function g.

Lemma 2 *The success curve $\mathbf{S} : [0,1] \to [0,1]$ always satisfies these properties:*

1. *$\mathbf{S}$ is convex.*
2. *There is a point $\delta_0 \in [0,1]$ (the* critical significance level*) such that $\mathbf{S}(\delta) = 0$ for $\delta \geq \delta_0$ and $\mathbf{S}'_{\text{left}}(\delta_0) < -1$; therefore, $\mathbf{S}'_{\text{left}} < -1$ and $\mathbf{S}'_{\text{right}} < -1$ to the left of δ_0, and the function $\mathbf{S}$ is decreasing before it hits the δ-axis at δ_0.*
3. *$\mathbf{S}$ is continuous at $\delta = 0$; therefore, it is continuous everywhere in $[0,1]$.*

If a function $\mathbf{S} : [0,1] \to [0,1]$ 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 $\mathbf{S}$ is the success curve.

Proof For the basic properties of convex functions and their left and right derivatives, see, e.g., [1], §I.4. The statement of the lemma follows from the fact that the success curve $\mathbf{S}$ 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. Flip F_1 around the line $x = 0.5$ and then around the line $y = 0.5$: $F_2(x) := 1 - F_1(1 - x)$.
3. Integrate F_2: $F_3(x) := \int_0^x F_2(t)dt$.
4. Invert F_3: $F_4 := F_3^{-1}$.
5. Flip F_4 around the line $y = 0.5$: $F_5 := 1 - F_4$.

It can be shown that $\mathbf{S} = F_5$, no matter which of the several natural definitions of the operation $g \mapsto g^{-1}$ is used; for concreteness, we can define

$$g^{-1}(y) := \sup\{x : g(x) \le y\}$$

for non-decreasing g (so that g^{-1} is continuous on the right). ∎

Visually the empirical performance curve in Figures 3 and 4 seems to satisfy the properties listed in Lemma 2 for significance levels that are not too large or too small (approximately in the range 0.1%–5%); for an even better agreement, see Figures 7 and 8.

A natural idea is to reverse the process of transforming F into $\mathbf{S}$ and try to obtain an estimate of the predictability distribution function F from an empirical performance curve. Figure 9 shows the result of such an attempt. Such pictures, however, should not be taken too seriously, since the differentiation operation needed in finding F is known to be unstable (see, e.g., [6], §1.12).

Proof of Theorem 1r

Let us check first that (4) indeed implies $\mathbf{P}(\delta) \ge \mathbf{S}(\delta)$. Since probability measures are σ-additive, (3) implies

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

and so we obtain from (4):

$$\mathbf{P}(\delta) \ge \limsup_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \ge \liminf_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \ge \mathbf{S}(\delta)$$

almost surely; since the two extreme terms are deterministic, we have $\mathbf{P}(\delta) \ge \mathbf{S}(\delta)$.

We start the actual proof with alternative definitions of calibration and performance curves. Complement the protocol of §1 in which Nature plays P^∞ and

Predictor plays $\Gamma_{1-\delta}$ with the following variables:

$$\overline{\mathrm{err}}_n := (P \times \mathbf{U})\Big\{(x, y, \tau) \in (\mathbf{Z} \times [0,1]) :$$
$$y \in \Gamma_{1-\delta}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)\Big\},$$
$$\overline{\mathrm{unc}}_n := (P_{\mathbf{X}} \times \mathbf{U})\Big\{(x, \tau) \in (\mathbf{X} \times [0,1]) :$$
$$|\Gamma_{1-\delta}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)| > 1\Big\},$$
$$\overline{\mathrm{Err}}_n := \sum_{i=1}^{n} \overline{\mathrm{err}}_i, \qquad \overline{\mathrm{Unc}}_n := \sum_{i=1}^{n} \overline{\mathrm{unc}}_i$$

(we are not always consistent in the order of arguments of the function $\Gamma_{1-\delta}$). The *prequential calibration curve* of Γ under P is defined by

$$\overline{\mathbf{C}}(\delta) := \inf\left\{\beta : \mathbb{P}\left\{\limsup_{n\to\infty} \frac{\overline{\mathrm{Err}}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \beta\right\} = 1\right\}$$

and the *prequential performance curve* of Γ under P by

$$\overline{\mathbf{P}}(\delta) := \inf\left\{\beta : \mathbb{P}\left\{\limsup_{n\to\infty} \frac{\overline{\mathrm{Unc}}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \beta\right\} = 1\right\},$$

where $\mathbb{P}$ refers to the probability distribution $(P \times \mathbf{U})^\infty$ over the examples $z_1, z_2, \ldots$ and random numbers $\tau_1, \tau_2, \ldots$. By the martingale strong law of large numbers the prequential versions of the calibration and performance curves coincide with the original versions: indeed, since $\mathrm{Err}_n - \overline{\mathrm{Err}}_n$ and $\mathrm{Unc}_n - \overline{\mathrm{Unc}}_n$ 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 $z_1, \ldots, z_n$ and $\tau_1, \ldots, \tau_n$, we have

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

and

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

(see, e.g., [5], Theorem VII.5.4).

It is clear that without loss of generality we can assume that Nature's move Γ_n at trial n is either $\{\hat{y}(x_n)\}$ or the whole label space $\mathbf{Y}$. Furthermore, we can assume that

$$\overline{\mathrm{unc}}_n = \mathbf{S}(\overline{\mathrm{err}}_n) \tag{10}$$

at every trial, since the best way to spend the allowance of $\overline{\mathrm{err}}_n$ is to be certain on objects x with the largest (topmost in Figure 5) representations $F(f(x))$. (For a formal argument, see the end of this proof.) Using the fact that the success curve $\mathbf{S}$ is convex, non-increasing, and continuous (see Lemma 2), we obtain

$$\begin{aligned}\frac{\overline{\mathrm{Unc}}_n}{n} &= \frac{1}{n}\sum_{i=1}^{n}\overline{\mathrm{unc}}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{S}(\overline{\mathrm{err}}_i)\\ &\geq \mathbf{S}\left(\frac{1}{n}\sum_{i=1}^{n}\overline{\mathrm{err}}_i\right) = \mathbf{S}\left(\frac{\overline{\mathrm{Err}}_n}{n}\right) \geq \mathbf{S}(\delta) - \epsilon,\end{aligned}$$

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

It remains to prove formally that $\overline{\mathrm{unc}}_n \geq \mathbf{S}(\overline{\mathrm{err}}_n)$ (which is the part of (10) that we actually used). Let us fix $1-\delta$ and

$$x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1};$$

we will write

$$\Gamma(x,\tau) := \Gamma_{1-\delta}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau),$$

omitting the fixed arguments. Without loss of generality we are assuming that either

$$\Gamma(x,\tau) = \hat{y}(x)$$

or

$$\Gamma(x,\tau) = \mathbf{Y}.$$

Set

$$p(x) := \mathbf{U}\left\{\tau : \Gamma(x,\tau) = \{\hat{y}(x)\}\right\}, \qquad \delta := \overline{\mathrm{err}}_n .$$

Our goal is to show that $\overline{\mathrm{unc}}_n \geq \mathbf{S}(\delta)$. To put it differently, we are required to show that the value of the optimisation problem

$$\int_{\mathbf{X}} p(x) P(dx) \to \max$$

subject to the constraint

$$\int_{\mathbf{X}} (1 - f(x)) p(x) P(dx) = \delta$$

is $1 - \mathbf{S}(\delta)$ at best. By the Neyman–Pearson lemma (see, e.g., [3]) there exist constants $c > 0$ and $d \in [0,1]$ such that

$$p(x) = \begin{cases} 1 & \text{if } f(x) > c \\ d & \text{if } f(x) = c \\ 0 & \text{if } f(x) < c. \end{cases}$$

The constants c and d are defined (c uniquely and d uniquely unless the probability of $f(x) = c$ is zero or $\delta = 0$) from the condition

$$\int_{x:f(x)>c} (1 - f(x)) P(dx) + d\int_{x:f(x)=c} (1-c) P(dx) = \delta,$$

which is equivalent, by Fubini's theorem (applied to the indicator function of the subgraph of F; see Figure 5), to

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

where $F(c-)$ is defined as $\lim_{\beta \uparrow c} F(\beta)$. This makes it obvious that the value of the above optimal problem is indeed $1 - \mathbf{S}(\delta)$.

This completes the proof of Theorem 1r; since Theorem 1 is a special case, it is also proved.

Proof of Theorems 2 and 2r

The fact that every rTCM Γ is perfectly calibrated is proved in [8], so we are only required to show that Γ is asymptotically optimal. Fix a confidence level $1 - \delta$; we will show that

$$\limsup_{n \to \infty} \frac{\overline{\mathrm{Unc}}_n(P^\infty, \Gamma_{1-\delta})}{n} \le \mathbf{S}(\delta) \tag{11}$$

almost surely (the underlying probability distribution $\mathbb{P}$ being the product $(P \times \mathbf{U})^\infty$). Without loss of generality we assume $\mathbf{S}(\delta) < 1$ ((11) holds trivially when $\mathbf{S}(\delta) = 1$). Set

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

The case $c = 1$ is simple: it means that $P\{x : f(x) < 1\} \le \mathbf{S}(\delta)$; since, almost surely, $\mathrm{unc}_n = 0$ at trials where $f(x_n) = 1$, by Borel's strong law of large numbers we immediately obtain (11). Therefore, we assume $c < 1$ in the rest of the proof.

First we consider the case $F(c) = \mathbf{S}(\delta)$ (this will be sufficient to prove Theorem 2, since $F(c) = \mathbf{S}(\delta)$ is implied by $F(c) = F(c-)$ and the rTCM constructed in this part of the proof will be deterministic). Notice that $F(c + \epsilon) > F(c)$ for any $0 < \epsilon \le 1 - c$ (we are assuming $\epsilon \le 1 - c$ so that $F(c + \epsilon)$ is defined). We will prove that, for any $0 < \epsilon \le 1 - c$ and from some n on,

$$\mathbb{P}(\mathrm{unc}_n \mid \mathcal{F}_{n-1}) \le F(c + \epsilon) \qquad \text{a.s.} \tag{12}$$

(we are using the same notation for an event and its indicator function). This will imply

$$\limsup_{n \to \infty} \frac{\overline{\mathrm{Unc}}_n}{n} \le F(c + \epsilon)$$

almost surely; since $\lim_{\epsilon \downarrow 0} F(c + \epsilon) = \mathbf{S}(\delta)$, this will prove (11).

Fix $0 < \epsilon \le 1 - c$; without loss of generality assume that F is continuous at $c + \epsilon$. Let us prove (12), assuming n is large enough. Suppose the examples observed before trial n are

$$(z_1, \dots, z_{n-1}) = ((x_1, y_1), \dots, (x_{n-1}, y_{n-1})).$$

Let us say that an example $(x, y) \in \mathbf{Z}$ is *wrongly classified* if $y \neq \hat{y}(x)$. Remember that according to the individual strangeness measure (8) the strange elements in every bag of examples are those that are wrongly classified, and the more predictable they are the stranger. Notice that the prediction for the new object x_n will be certain if (a) the new object x_n is to the right of $\beta = c + \epsilon$ in Figure 5, in the sense $f(x_n) \geq c+\epsilon$, and (b) the number of the wrongly classified objects x_i, $i = 1, \ldots, n-1$, to the right of $\beta = c + \epsilon$ does not exceed $\lfloor n\delta \rfloor - 10$. The probability (conditional on $\mathcal{F}_{n-1}$) of (a) is $1 - F(c+\epsilon)$, so to prove (12) it is sufficient to show that the event (b) (remember that it is measurable w.r. to $\mathcal{F}_{n-1}$) happens from some n on almost surely. The probability that an object is wrongly classified and to the right of $\beta = c + \epsilon$ is

$$b := \int_0^1 (F(\beta) - F(c+\epsilon)) d\beta < \delta.$$

By Hoeffding's inequality (see, e.g., [2], Theorem 8.1) the probability that the event (b) will fail to happen is bounded from above by

$$e^{-2(\lfloor n\delta \rfloor - 10 - (n-1)b)^2/(n-1)} \leq e^{-\kappa n}, \tag{13}$$

for some positive constant κ and from some n on. Since

$$\sum_n e^{-\kappa n} < \infty,$$

the Borel–Cantelli lemma implies that (b) will almost surely happen from some n on. This completes the proof in the case $F(c) = \mathbf{S}(\delta)$.

Now we consider the case $F(c) > \mathbf{S}(\delta)$ (which is the only remaining possibility). In the remaining part of the proof it will be important that we consider rTCM rather than TCM. The intuition may be helped if the definition of rTCM is restated as follows: rTCM is a TCM applied to the strangeness values (α_i, τ_i) (α_i being the original strangeness values and $\tau_i \in [0,1]$ the random numbers) which are ordered lexicographically: $(\alpha_1, \tau_1) \leq (\alpha_2, \tau_2)$ if either $\alpha_1 < \alpha_2$ or $\alpha_1 = \alpha_2$ & $\tau_1 \leq \tau_2$.

Let $\epsilon > 0$ satisfy $\epsilon < F(c) - \mathbf{S}(\delta)$. We will prove that, from some n on,

$$\mathbb{P}(\mathrm{unc}_n \mid \mathcal{F}_{n-1}) \leq \mathbf{S}(\delta) + \epsilon \qquad \text{a.s.} \tag{14}$$

This will imply

$$\limsup_{n \to \infty} \frac{\overline{\mathrm{Unc}_n}}{n} \leq \mathbf{S}(\delta) + \epsilon$$

almost surely, and so prove (11).

We say that an object and random number $(x, \tau) \in \mathbf{X} \times [0,1]$ (such a pair will be called an *extended object*) is *above the line* $\mathbf{S}(\delta) + \epsilon$ (cf. Figure 5) if either $f(x) > c$ or

$$f(x) = c \ \& \ \tau \geq \frac{\mathbf{S}(\delta) + \epsilon - F(c)}{F(c) - F(c-)}$$

(this definition corresponds to representing each extended object (x, τ) by the point

$$(f(x), \tau F(f(x)) + (1-\tau) F(f(x)-))$$

in Figure 5).

Let us prove (14), assuming n is large enough. Suppose the extended objects observed before trial n are

$$(x_1, \tau_1, \ldots, x_{n-1}, \tau_{n-1}).$$

Now the prediction for the new object x_n will be certain if (a) the new extended object (x_n, τ_n) is above $\mathbf{S}(\delta) + \epsilon$, and (b) the number of the wrongly classified extended objects (x_i, τ_i), $i = 1, \ldots, n-1$, above $\mathbf{S}(\delta) + \epsilon$ does not exceed $\lfloor n\delta \rfloor - 10$. (We say that (x, τ) is wrongly classified if x is.) The probability (conditional on $\mathcal{F}_{n-1}$) of (a) is $1 - \mathbf{S}(\delta) - \epsilon$, so to prove (14) it is sufficient to show that the event (b) happens from some n on almost surely. The probability that an extended object is wrongly classified and above $\mathbf{S}(\delta) + \epsilon$ is

$$b := \int_0^1 (F(\beta) - \mathbf{S}(\delta) - \epsilon) d\beta < \delta.$$

The proof is completed literally as before: apply Hoeffding's inequality to obtain (13) and then apply the Borel–Cantelli lemma.