Universal Well-Calibrated Algorithm for On-line Classification

Vladimir Vovk
Computer Learning Research Center
Department of Computer Science
Royal Holloway, University of London
vovk@cs.rhul.ac.uk

Technical Report

CLRC-TR-02-04

11th November 2002

Royal Holloway
University of London

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

Abstract

We study the problem of on-line classification in which the prediction algorithm is given a "confidence level" $1-\delta$ and is required to output as its prediction a range of labels (intuitively, those labels deemed compatible with the available data at the level δ) rather than just one label; as usual, the examples are assumed to be generated independently from the same probability distribution P. The prediction algorithm is said to be "well-calibrated" for P and δ if the long-run relative frequency of errors does not exceed δ P-almost surely. For well-calibrated algorithms we take the number of "uncertain" predictions (i.e., those containing more than one label) as the principal measure of predictive performance. The main result of this paper is the construction of a prediction algorithm which, for any (unknown) P and any δ: (a) makes errors independently and with probability δ at every trial (in particular, is well-calibrated for P and δ); (b) makes in the long run no more uncertain predictions than any other prediction algorithm that is well-calibrated for P and δ; (c) processes example n in time $O(\log n)$.

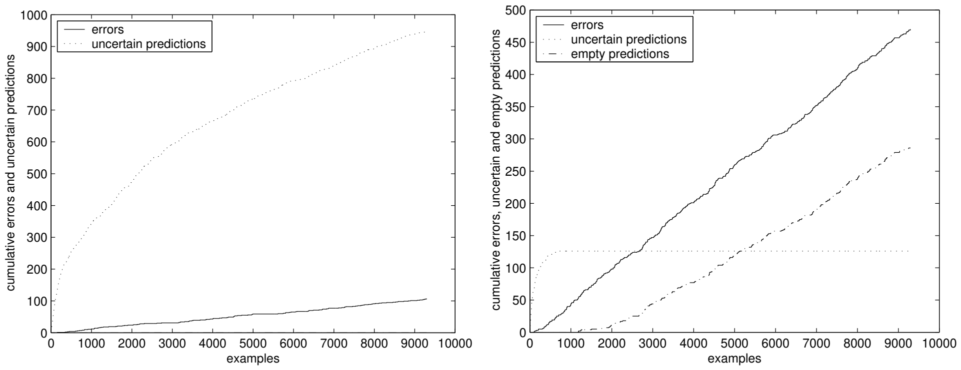

Figure 1 TCM at 99% (left) and 95% (right) on the USPS data set

1 Introduction

Typical machine learning algorithms output a point prediction for the label of an unknown object. This paper continues study of an algorithm, called Transductive Confidence Machine (TCM) and first introduced in [4], that complements its predictions with some measures of confidence. There are different ways of presenting TCM's output; in this paper (as in the related [3]) we use TCM as a "region predictor", in the sense that for any confidence level $1-\delta$ it outputs a predictive region rather than a point prediction.

Paper [3] shows that any TCM is well-calibrated when used in the on-line mode: for any confidence level $1-\delta$ the long-run relative frequency of erroneous predictions does not exceed δ. What makes this property of TCM especially appealing is that it is far from being just an asymptotic phenomenon: a slight modification of TCM called randomized TCM (rTCM; randomization is needed to break ties and deal efficiently with borderline cases) makes errors independently at different trials and with probability δ at each trial; well-calibratedness then immediately follows by the Borel strong law of large numbers.

The justification of the study of TCM given in [3] was its good performance on real-world and standard benchmark data sets. For example, Figure 1 shows that for the standard confidence levels 99% and 95% most examples in the well-known USPS data set (randomly permuted) can be predicted categorically (by a simple Nearest Neighbors TCM): the predictive region contains only one label.

This paper presents theoretical results about TCM's performance; we show that there exists a *universal* rTCM, which, for any confidence level $1-\delta$ and without knowing the true distribution P generating the examples:

- produces, asymptotically, no more uncertain predictions than any other prediction algorithm that is well-calibrated for P and δ;

- produces, asymptotically, at least as many empty predictions as any other prediction algorithm that is well-calibrated for P and δ and whose percentage of uncertain predictions is optimal (in the sense of the previous item).

The importance of the first item is obvious: we want to minimize the number of uncertain predictions. This criterion ceases to work, however, when the number of uncertain predictions stabilizes, as in Figure 1 (right). In such cases the number of empty predictions becomes important: empty predictions (automatically leading to an error) provide a warning that the object is difficult to predict, and one would like to be warned as often as possible, taking into account that the frequency of errors (including empty predictions) is guaranteed not to exceed δ in the long run.

This paper's result elaborates on [2], where it was shown that an optimal TCM exists when the distribution P generating the examples is known. Here we consider only randomized TCM, so we drop the adjective "randomized".

The two areas of mainstream machine learning that are most closely connected with this paper are PAC learning theory and Bayesian learning theory. Whereas we often use the rich arsenal of mathematical tools developed in these fields, they do not provide the same kind of guarantees (a prespecified probability of error, with errors at different trials independent) under unknown P; for more details, see [3] and references therein.

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 whether an error is made at trial n); it also includes analogous variables Unc_n, unc_n, Emp_n, emp_n for uncertain and empty predictions:

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

We will use the notation $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ for the *example space*; Γ_n will be called the *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 randomized 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 99% and 95%); 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 randomized strategies.

We will often use the notation err_n, unc_n, etc., in the case where Predictor and Nature are using given randomized 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* for a probability distribution P in $\mathbf{Z}^\infty$ and a significance level $\delta \in (0, 1)$ if

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

We say (following [2]) that Γ is *optimal* for P and δ if, for any region predictor $\Gamma^\dagger$ which is well-calibrated for P and δ,

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

(the symbol "a.s." in such a context always assumes that the random numbers used by Γ and $\Gamma^\dagger$ are independent). Of course, the definition of optimality is natural only for well-calibrated Γ.

A region predictor Γ is *universal well-calibrated* if:

- ◇ it is well-calibrated for any P and δ;

- it is optimal for any P and δ;
- for any P, any δ, and any region predictor $\Gamma^\dagger$ which is well-calibrated and optimal for P and δ,

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

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.

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

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

3 Construction of a universal well-calibrated region predictor

3.1 Preliminaries

If τ is a number in $[0,1]$, we split it into two numbers $\tau', \tau'' \in [0,1]$ 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 $\tau' := 0.a_1a_3a_5\ldots$ and $\tau'' := 0.a_2a_4a_6\ldots$. If τ is distributed uniformly in $[0,1]$, then both τ' and τ'' are, and they are independent of each other.

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

Let us set $\mathbf{X} := [0,1]$; we can do this without loss of generality since $\mathbf{X}$ is Borel. This makes the extended object space $\tilde{\mathbf{X}} = [0,1]^2$ a linearly ordered set with the lexicographic order: $(x_1,\sigma_1) < (x_2,\sigma_2)$ means that either $x_1 = x_2$ and $\sigma_1 < \sigma_2$ or $x_1 < x_2$. We say that (x_1,σ_1) is *nearer* to (x_3,σ_3) than (x_2,σ_2) is if

$$|x_1 - x_3, \sigma_1 - \sigma_3| < |x_2 - x_3, \sigma_2 - \sigma_3|, \tag{1}$$

where

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

Our construction will be based on the Nearest Neighbors algorithm, which is known to be strongly universally consistent in the traditional theory of pattern recognition (see, e.g., [1], Chapter 11); the random components σ are needed for tie-breaking. As usual, to give a precise meaning to the expression "the kth nearest neighbor" in a sequence of extended objects to another extended object v, we use the convention that extended objects with smaller indices in the sequence are considered to be nearer to v (this particular convention is not essential, since adding the random components σ ensures that ties will occur with probability zero).

3.2 Transductive Confidence Machines

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

$$A_n : (w_1, \ldots, w_n) \mapsto (\alpha_1, \ldots, \alpha_n) \tag{2}$$

is determined by w_i and the bag $\wr w_1, \ldots, w_n \int$.

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

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

and include y in $\Gamma_{1-\delta}$ if and only if

$$\tau_n'' < \frac{\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\} - n\delta}{\#\{i = 1, \ldots, n : \alpha_i = \alpha_n\}} \tag{3}$$

(in particular, include y in $\Gamma_{1-\delta}$ if $\#\{i = 1, \ldots, n : \alpha_i > \alpha_n\}/n > \delta$ and do not include y in $\Gamma_{1-\delta}$ if $\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\}/n \leq \delta$).

A *TCM* is the TCM associated with some individual strangeness measure. It was shown in [3] that

Proposition 1 *Every TCM is well-calibrated for every P and δ.*

3.3 Universal TCM

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

$$K_n \to \infty,\ K_n = o(n/\ln n) \tag{4}$$

as $n \to \infty$. The *Nearest Neighbors 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 (2)), we first define Nearest Neighbors approximations $P_n^{\neq}(y \mid x_i, \sigma_i)$ to the true (but unknown) conditional probabilities $P(y \mid x_i)$: for every extended example (x_i, σ_i, y_i) in the sequence,

$$P_n^{\neq}(y \mid x_i, \sigma_i) := N^{\neq}(x_i, \sigma_i, y)/K_n, \tag{5}$$

where $N^{\neq}(x_i, \sigma_i, y)$ is the number of $j = 1, \ldots, n$ such that $y_j = y$ and (x_j, σ_j) is one of the K_n nearest neighbors of (x_i, σ_i) in the sequence $((x_1, \sigma_1), \ldots, (x_{i-1}, \sigma_{i-1}), (x_{i+1}, \sigma_{i+1}), \ldots, (x_n, \sigma_n))$. (The upper index $\neq$ reminds us of the fact that (x_i, σ_i) is not counted as one of its own nearest neighbors in this definition.) If $K_n \geq n$ or $K_n \leq 0$, this definition does not work, so set, e.g., $P_n^{\neq}(y \mid x_i, \sigma_i) := 1/|\mathbf{Y}|$ for all y and i (this particular convention is not essential since, by (4), $0 < K_n < n$ from some n on).

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

$$f_n^{\neq}(x_i, \sigma_i) := \max_{y \in \mathbf{Y}} P_n^{\neq}(y \mid x_i, \sigma_i).$$

For each (x_i, σ_i) fix some

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

(e.g., take the first element of $\arg\max_y P_n^{\neq}(y \mid x_i, \sigma_i)$ in a fixed ordering of $\mathbf{Y}$) and define the mapping (2) (where $w_i = (x_i, \sigma_i, y_i)$, $i = 1, \ldots, n$) setting

$$\alpha_i := \begin{cases} -f_n^{\neq}(x_i, \sigma_i) & \text{if } y_i = \hat{y}_n(x_i, \sigma_i) \\ f_n^{\neq}(x_i, \sigma_i) & \text{otherwise.} \end{cases} \tag{6}$$

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

Proposition 2 *If $\mathbf{X} = [0, 1]$ and $K_n \to \infty$ sufficiently slowly, the Nearest Neighbors TCM can be implemented so that computations at step n are performed in time $O(\log n)$.*

4 Fine details of region prediction

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

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

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

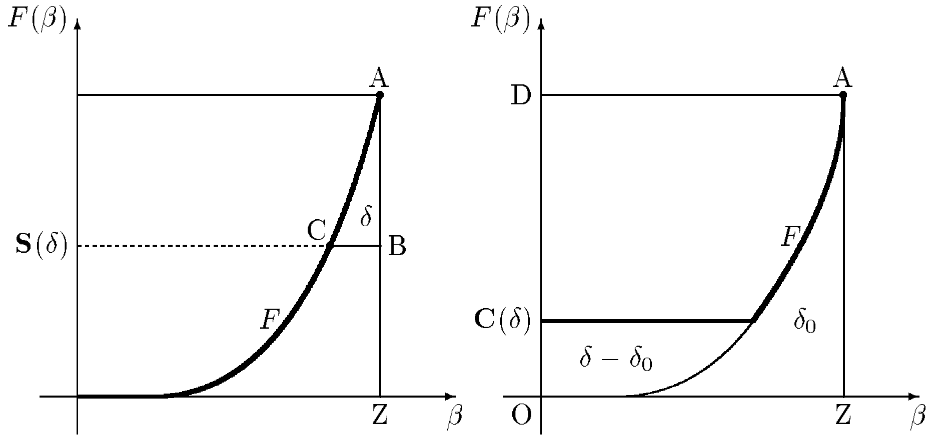

Figure 2 The predictability distribution function F and the success curve $\mathbf{S}(\delta)$ (left); the complementary success curve $\mathbf{C}(\delta)$ (right)

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

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

An example of such a function F is given in Figure 2 (left), where the graph of F is the thick line.

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

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

where t^+ stands for $\max(t,0)$; the function $\mathbf{S}$ is also of the type $[0,1] \to [0,1]$. Geometrically, $\mathbf{S}_P(\delta)$ is defined from the graph of F as follows (see Figure 2, left; we usually drop the lower index $_P$): move the point B from A to Z until the area of the curvilinear triangle ABC becomes δ or B reaches Z; the ordinate of B is then $\mathbf{S}(\delta)$.

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

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

Similarly to the case of $\mathbf{S}(\delta)$, $\mathbf{C}(\delta)$ is defined as the value such that the area of the part of the box AZOD below the thick line in Figure 2 (right) is δ ($\mathbf{C}(\delta) = 0$ if such a value does not exist).

Define the *critical significance level* δ_0 as

$$\delta_0 := \int_0^1 F(\beta) d\beta.$$

It is clear that

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

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

The following result is proven in [2].

Proposition 3 *Let P be a probability distribution in $\mathbf{Z}$ with success curve $\mathbf{S}$ and $\delta > 0$ be a significance level. If a region predictor Γ is well-calibrated for P and δ, then*

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

In this paper we complement Proposition 3 with

Proposition 4 *Let P be a probability distribution in $\mathbf{Z}$ with success curve $\mathbf{S}$ and complementary success curve $\mathbf{C}$ and $\delta > 0$ be a significance level. If a region predictor Γ is well-calibrated for P and δ and satisfies*

$$\limsup_{n\to\infty} \frac{\mathrm{Unc}_n(P^\infty, \Gamma_{1-\delta})}{n} \leq \mathbf{S}(\delta) \quad a.s., \tag{7}$$

then

$$\limsup_{n\to\infty} \frac{\mathrm{Emp}_n(P^\infty, \Gamma_{1-\delta})}{n} \leq \mathbf{C}(\delta) \quad a.s.$$

Theorem 1 immediately follows from Propositions 1, 3, 4 and the following proposition.

Proposition 5 *Suppose $\mathbf{X}$ is Borel. The Nearest Neighbors TCM constructed in §3.3 satisfies, for any P and any significance level δ,*

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

and

$$\liminf_{n\to\infty} \frac{\mathrm{Emp}_n(P^\infty, \Gamma_{1-\delta})}{n} \geq \mathbf{C}_P(\delta) \quad a.s. \tag{9}$$

Proofs of some of these propositions can be found in the appendix.

References

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

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

[3] 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.

[4] Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the 16th International Conference on Machine Learning*, pages 444–453, San Francisco, CA, 1999. Morgan Kaufmann.

Appendix: Proofs

A.1 Proof sketch of Proposition 2

For simplicity we will be assuming that all extended objects $(x_i, \tau_i') \in [0,1]^2$ are different (in any case, this is true with probability one, since τ_i' are independent random numbers distributed as $\mathbf{U}$). Our computational model has an operation of splitting $\tau \in [0,1]$ into τ' and τ'' (or is allowed to generate both τ_n' and τ_n'' at every trial n).

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

- a red-black binary *search tree* (see, e.g., [CLRS01], Chapters 13–15; the only two operations on red-black trees we need in this paper are the query SEARCH and the modifying operation INSERT);
- a growing *array* of nonnegative integer numbers indexed by numbers $k \in \{-K_n, -K_n+1, \dots, K_n\}$ (where n is the ordinal number of the example being processed).

Immediately after processing the nth extended example (x_n, τ_n, y_n) the contents of these data structures are as follows:

- The search tree contains n vertices, corresponding to the extended examples (x_i, τ_i, y_i) seen so far. The key of vertex i is the extended object $(x_i, \tau_i') \in [0,1]^2$; the linear order on the keys is the lexicographic order. The other information contained in vertex i is the random number τ_i'', the label y_i, the set $\{P_n^{\neq}(y \mid x_i, \tau_i') : y \in \mathbf{Y}\}$ of conditional probability estimates (5), the pointer to the following vertex (i.e., the vertex that has the smallest key greater than (x_i, τ_i'); if there is no greater key, the pointer is NIL), and the pointer to the previous vertex (i.e., the vertex that has the greatest key smaller than (x_i, τ_i'); if (x_i, τ_i') is the smallest key, the pointer is NIL).
- The array contains the numbers

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

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

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

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

- ⋄ Generates τ'_n and τ''_n.
- ⋄ Locates the successor and predecessor of (x_n, τ'_n) in the search tree (using the query SEARCH and the pointers to the following and previous vertices); this requires time $O(\log n)$.
- ⋄ Computes the conditional probabilities $\{P_n^{\neq}(y \mid x_n, \tau'_n) : y \in \mathbf{Y}\}$; this also gives $\hat{y}_n(x_n, \tau'_n)$. The required time is $O(K_n) = O(\log n)$.
- ⋄ For each $y \in \mathbf{Y}$ looks at what happens if the nth example is $(x_n, \tau_n, y_n) = (x_n, \tau_n, y)$: computes α_n and updates (if necessary) α_i for (x_i, τ'_i) in the vicinity of (x_n, τ'_n); using the array and τ''_n, finds if $y \in \Gamma_n$. This requires time $O(K_n^2) = O(\log n)$.
- ⋄ Outputs the predictive region Γ_n (time $O(1)$).

When the label y_n arrives, the algorithm:

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

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

  for the $2K_n$ existing vertices (x_i, τ'_i) in the vicinity of (x_n, τ'_n); this requires time $O(K_n^2) = O(\log n)$. The conditional probabilities for other (x_i, τ'_i), $i = 1, \ldots, n-1$, do not change.
- ⋄ Updates the array, changing $N(K_n\alpha_i)$ for the $(x_i, \tau'_i) \neq (x_n, \tau'_n)$ in the vicinity of (x_n, τ'_n) and for both old and new values of α_i and changing $N(K_n\alpha_n)$ (time $O(K_n) = O(\log n)$).

In conclusion we discuss how to do updates required when K_n increases. An *epoch* is defined to be a maximal sequence of ns with the same K_n. Since the changes that need to be done when a new epoch starts are substantial, they will be spread over the whole preceding epoch. An epoch is *odd* if the corresponding K_n is odd and *even* if K_n is even. At every step in an epoch we prepare the ground for the next epoch. By the end of epoch $n = A+1, A+2, \ldots, B$ we need to change B sets $\{P_n^{\neq}(y \mid x_i, \tau_i') : y \in \mathbf{Y}\}$ in $B - A$ steps (the duration of the epoch). Therefore, each vertex of the search tree should contain not only $\{P_n^{\neq}(y \mid x_i, \tau_i')\}$ for the current epoch but also $\{P_n^{\neq}(y \mid x_i, \tau_i')\}$ for the next epoch (two structures for holding $\{P_n^{\neq}(y \mid x_i, \tau_i')\}$ will suffice, one for even epochs and one for odd epochs). If K_n grows slowly enough (say, as $\log n$), $B/A = O(1)$. At each step, $O(1)$ sets $\{P_n^{\neq}(y \mid x_i, \tau_i')\}$ for the next epoch are added. This will take time $O(K_n) = O(\log n)$. As soon as a set $\{P_n^{\neq}(y \mid x_i, \tau_i')\}$ for the next epoch is added at some trial, both sets (for the current and next epoch) will have to be updated for each new example. In a similar way the array for the next epoch is gradually built up.

A.2 Proof sketch of Proposition 4

The proof of Proposition 4 is similar to (but more complicated than) the proof of Theorems 1 and 1r in [Vov02]; this proof sketch can be made rigorous using the Neyman-Pearson lemma, as in [Vov02].

We will use the notations g'_{left} and g'_{right} for the left and right derivatives, respectively, of a function g. The following lemma parallels Lemma 2 in [Vov02], which deals with $\mathbf{S}(\delta)$.

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

1. *There is a point $\delta_0 \in [0,1]$ (viz. the critical significance level) such that $\mathbf{C}(\delta) = 0$ for $\delta \le \delta_0$ and $\mathbf{C}(\delta)$ is concave for $\delta \ge \delta_0$.*
2. *$\mathbf{C}'_{\text{right}}(\delta_0) < \infty$ and $\mathbf{C}'_{\text{left}}(1) \ge 1$; therefore, for $\delta \in (\delta_0, 1)$, $1 \le \mathbf{C}'_{\text{right}}(\delta) \le \mathbf{C}'_{\text{left}}(\delta) < \infty$ and the function $\mathbf{C}(\delta)$ is increasing.*
3. *$\mathbf{C}(\delta)$ is continuous at $\delta = \delta_0$; therefore, it is continuous everywhere in $[0,1]$.*

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

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

1. Invert F: $F_1 := F^{-1}$.
2. Integrate F_1: $F_2(x) := \int_0^x F_1(t)dt$.

3. Increase F_2: $F_3(x) := F_2(x) + \delta_0$, where $\delta_0 := \int_0^1 F(x)dx$.

4. Invert F_3: $F_4 := F_3^{-1}$.

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

Complement the protocol of §2 in which Nature plays P^∞ and Predictor plays $\Gamma_{1-\delta}$ with the following variables:

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

and

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

By the martingale strong law of large numbers it suffices to prove the proposition for these "predictable" versions of Err_n, Unc_n, and Emp_n: 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 $(x_1,\tau_1,y_1),\ldots,(x_n,\tau_n,y_n)$, we have

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

and

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

(see, e.g., [Shi96], Theorem VII.5.4).

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

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

at every trial, since the optimal way (which is essentially the only optimal way) to spend the allowance of $\overline{\mathrm{err}}_n$ is to be certain on objects x with the largest (uppermost in Figure 2) representations $F(f(x))$ and, if part of the allowance is still left when certainty is attained for all objects (i.e., if $\overline{\mathrm{err}}_n > \delta_0$), be empty on objects x with the smallest (lowermost in Figure 2) representations $F(f(x))$. (A formal argument for the statement about $\mathbf{S}$ is given in [Vov02].)

From the argument of [Vov02] (Proof of Theorems 1 and 1r) it is clear that to achieve (7) the region predictor must satisfy

$$\begin{aligned}
\delta < \delta_0 &\implies \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (\overline{\mathrm{err}}_i - \delta_0)^+ = 0\\
\delta \ge \delta_0 &\implies \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (\delta_0 - \overline{\mathrm{err}}_i)^+ = 0.
\end{aligned}$$

Using the fact that the complementary success curve $\mathbf{C}$ is concave, increasing, and (uniformly) continuous for $\delta \geq \delta_0$ (see Lemma 1), we obtain: if $\delta < \delta_0$,

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

if $\delta \geq \delta_0$,

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

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

A.3 Proof sketch of Proposition 5

Let us first modify and extend the notation $P_n^{\neq}(y \mid x_i, \sigma_i)$ introduced in (5). Consider the sequence of extended examples $w_i = (x_i, \tau_i', y_i)$, $i = 1, \ldots, n$ ((x_i, y_i) are the first n examples chosen by Nature and τ_i are the random numbers used by Predictor). We define the Nearest Neighbors approximations $P_n(y \mid x, \sigma)$ to the conditional probabilities $P(y \mid x)$ as follows: for every $(x, \sigma, y) \in \tilde{\mathbf{Z}}$,

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

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

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

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

Lemma 2 *Suppose $K_n \to \infty$, $K_n = o(n)$, and $\mathbf{Y} = \{0, 1\}$. For any $\epsilon > 0$ and large enough n,*

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

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

Proof This is almost a special case of Devroye et al.'s [DGKL94] Theorem 1. There is, however, an important difference between the way we break distance ties and the way Devroye et al. [DGKL94] do this: in [DGKL94], instead of our (1),

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

is used. (Our way of breaking ties better agrees with the lexicographic order on $[0,1]^2$, which is useful in the proof of Proposition 2 and, less importantly, in the proof of Lemma 4.) It is easy to check that the proof given in [DGKL94] also works (and becomes simpler) for our way of breaking distance ties. ∎

Lemma 3 *Suppose $K_n \to \infty$ and $K_n = o(n)$. For any $\epsilon > 0$ there exists an $\epsilon^* > 0$ such that, for large enough n,*

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

in particular,

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

Proof We apply Lemma 2 to the binary classification problem obtained from our classification problem by replacing label $y \in \mathbf{Y}$ with 1 and replacing all other labels with 0:

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

By Markov's inequality this implies

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

which, in turn, implies

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

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

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

Lemma 4 *Suppose $K_n \to \infty$ and $K_n = o(n)$. For any $\epsilon > 0$ there exists an $\epsilon^* > 0$ such that, for large enough n,*

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

In particular,

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

Proof Since

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

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

Define

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

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

The main tool in this proof (and several other proofs in this appendix) will be McDiarmid's theorem (see, e.g., [DGL96], Theorem 9.2). First we check the possibility of its application. If we replace an extended object (x_j, τ_j') by another extended object (x_j^*, τ_j^*), the expression

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

will change as follows:

- the addend $I_n(x_i, \tau_i')$ for $i = j$ changes by 1 at most;
- the addends $I_n(x_i, \tau_i')$ for $i \neq j$ such that neither (x_j, τ_j') nor (x_j^*, τ_j^*) are among the K_n nearest neighbors of (x_i, τ_i') do not change at all;
- the sum over the at most $4K_n$ (see below) addends $I_n(x_i, \tau_i')$ for $i \neq j$ such that either (x_j, τ_j') or (x_j^*, τ_j^*) (or both) are among the K_n nearest neighbors of (x_i, τ_i') can change by at most

$$4K_n \frac{1}{\epsilon} \frac{1}{K_n} = \frac{4}{\epsilon}. \tag{10}$$

The left-hand side of (10) reflects the following facts: the change in $P_n(y \mid x_i, \tau_i')$ for $i \neq j$ is at most $1/K_n$; the number of $i \neq j$ such that (x_j, τ_j') is among the K_n nearest neighbors of (x_i, τ_i') does not exceed $2K_n$ (since the extended objects are linearly ordered and (1) is used for breaking distance ties); analogously, the number of $i \neq j$ such that (x_j^*, τ_j^*) is among the K_n nearest neighbors of (x_i, τ_i') does not exceed $2K_n$.

Therefore, by McDiarmid's theorem,

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} I_n(x_i,\tau_i') - \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} I_n(x_i,\tau_i')\right) > \epsilon\right\} \\
&\quad \le \exp\left(-2\epsilon^2 n/\left(1+4/\epsilon\right)\right) = \exp\left(-\frac{2\epsilon^3}{4+\epsilon}n\right).
\end{aligned} \tag{11}$$

Next we find:

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

(the penultimate inequality follows from Lemma 3) from some n on. In combination with (11) this implies

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

in particular

$$\mathbb{P}\left\{\frac{\#\left\{i : \max_y |P(y \mid x_i) - P_n(y \mid x_i,\tau_i')| \ge 2\epsilon\right\}}{n} > 3\epsilon\right\} \le \exp\left(-\frac{2\epsilon^3}{4+\epsilon}n\right).$$

Replacing 3ϵ by ϵ, we obtain that, from some n on,

$$\mathbb{P}\left\{\frac{\#\left\{i : \max_y |P(y \mid x_i) - P_n(y \mid x_i,\tau_i')| > \epsilon\right\}}{n} > \epsilon\right\} \le \exp\left(-\frac{2(\epsilon/3)^3}{4+\epsilon/3}n\right),$$

which completes the proof. ∎

We say that an extended example (x_i,τ_i,y_i), $i = 1,\ldots,n$, is *n-strange* if $y_i \ne \hat{y}_n(x_i,\tau_i')$; otherwise, (x_i,τ_i,y_i) will be called *n-ordinary.* For concreteness, let us break ties in the expression "the k largest" by regarding elements corresponding to larger indices as larger for this purpose (although with probability one, ties will never happen in our case).

Lemma 5 *Suppose (4) is satisfied and $\delta \le \delta_0$. With probability one, the $\lfloor(1-\mathbf{S}(\delta))n\rfloor$ extended examples with the largest (in the sense of the lexicographic order) $(f_n^{\ne}(x_i,\tau_i'),\tau_i'')$ among $(x_1,\tau_1,y_1),\ldots,(x_n,\tau_n,y_n)$ contain at most $n\delta + o(n)$ n-strange extended examples as $n \to \infty$.*

Proof sketch Define

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

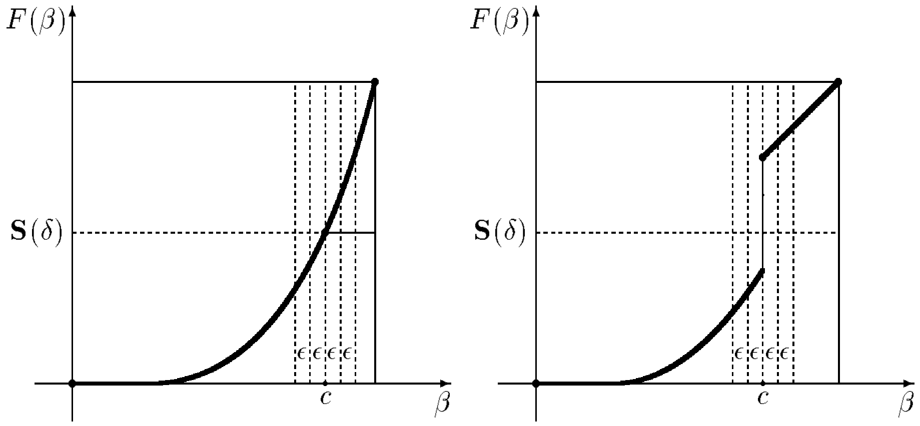

Figure 3 Cases $F(c) = \mathbf{S}(\delta)$ (left) and $F(c) > \mathbf{S}(\delta)$ (right)

It is clear that $0 < c < 1$. Our proof will work both in the case where $F(c) = \mathbf{S}(\delta)$ and in the case where $F(c) > \mathbf{S}(\delta)$, as illustrated in Figure 3.

Let $\epsilon > 0$ be a small constant (we will let $\epsilon \to 0$ eventually). Define a "threshold" $(c'_n, c''_n) \in [0,1]^2$ requiring that

$$\mathbb{P}\left\{f(x_n) = c, (f_{n-1}(x_n, \tau'_n), \tau''_n) > (c'_n, c''_n)\right\} = F(c) - \mathbf{S}(\delta) - \epsilon \qquad (12)$$

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

$$2\epsilon < F(c) - \mathbf{S}(\delta) \qquad (13)$$

to hold (among other things this will ensure the validity of the definition (12)). If $F(c) = \mathbf{S}(\delta)$, we set $(c', c'') := (c + \epsilon, 0)$; in any case, we will have

$$\mathbb{P}\left\{f(x_n) = c, (f_{n-1}(x_n, \tau'_n), \tau''_n) > (c'_n, c''_n)\right\} \geq F(c) - \mathbf{S}(\delta) - \epsilon. \qquad (14)$$

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

$$(f_n^{\neq}(x_i, \tau'_i), \tau''_i) > (c'_n, c''_n);$$

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

Class I: Those satisfying $f(x_i) \leq c - 2\epsilon$.

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

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

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

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

First we explain the general idea of the proof. The threshold (c', c'') was chosen so that approximately $\lfloor (1 - \mathbf{S}(\delta))n \rfloor$ of the available extended examples will be above the threshold. Because of this, the extended examples above the threshold will essentially be the $\lfloor (1 - \mathbf{S}(\delta))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ referred to in Lemma 5. For each of our five classes we will be interested in the following questions:

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

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

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

$$c - \epsilon \le c_n' \le c + \epsilon \tag{15}$$

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

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

(combine Lemma 3 with (13)). The inequality $c - \epsilon \le c_n'$ follows from

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

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

$$\epsilon n \tag{16}$$

will be above the threshold from some n on almost surely (by Lemma 4 and the Borel-Cantelli lemma). None of the Class II extended examples will be above the threshold, by definition. The fraction of Class III extended examples among the first n extended examples will tend to

$$F(c + 2\epsilon) - F(c) + F(c-) - F(c - 2\epsilon) \tag{17}$$

as $n \to \infty$ almost surely.

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

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

the assumption $K_n = o(n/\ln n)$ and the Borel-Cantelli lemma imply that

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

from some n on almost surely. Since

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

(see (14)), we have

$$N_n^{\mathrm{IV}} > (F(c) - \mathbf{S}(\delta) - 2\epsilon)n \tag{18}$$

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

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

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

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

$$\begin{aligned}
\frac{1}{n}\,\mathbb{E}\,N_n^{\mathrm{IV,str}} &= \mathbb{P}\left\{f(x_n) = c, (f_{n-1}(x_n, \tau_n'), \tau_n'') > (c_n', c_n''), \hat{y}_n(x_n, \tau_n') \neq y_n\right\} \\
&= \mathbb{E}\left(\left(1 - P\left(\hat{y}_n(x_n, \tau_n') \mid x_n\right)\right) \mathbb{I}_{\{f(x_n)=c,(f_{n-1}(x_n,\tau_n'),\tau_n'')>(c_n',c_n'')\}}\right) \\
&\leq e^{-\epsilon^* n} + \epsilon + (1 - c + \epsilon) \\
&\quad \times \mathbb{P}\{f(x_n) = c, (f_{n-1}(x_n, \tau_n'), \tau_n'') > (c_n', c_n'')\} \\
&= e^{-\epsilon^* n} + \epsilon + (1 - c + \epsilon)(F(c) - \mathbf{S}(\delta) - \epsilon) && (19) \\
&\leq (F(c) - \mathbf{S}(\delta))(1 - c) + 3\epsilon && (20)
\end{aligned}$$

(the first inequality follows from Lemma 3) in the case $F(c) > \mathbf{S}(\delta)$. If $F(c) = \mathbf{S}(\delta)$, the lines (19) and (20) of this chain have to be changed to

$$\begin{aligned}
&\leq e^{-\epsilon^* n} + \epsilon + (1 - c + \epsilon)\left(e^{-\epsilon^* n} + \epsilon\right) \\
&< 3\epsilon,
\end{aligned}$$

but the inequality between the extreme terms of the chain still holds. Therefore, the number of n-strange Class IV extended examples does not exceed

$$\left((F(c) - \mathbf{S}(\delta))(1 - c) + 4\epsilon\right) n \tag{21}$$

from some n on almost surely.

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

$$1 - F(c + 2\epsilon) \tag{22}$$

as $n \to \infty$ almost surely. By Lemma 4, the Borel-Cantelli lemma, and (15), almost surely from some n on at least

$$(1 - F(c + 2\epsilon) - 2\epsilon)n \tag{23}$$

extended examples in Class V will be above the threshold.

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

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

from some n on almost surely. Now

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

from some n on (the first inequality follows from Lemma 3). Therefore,

$$\frac{1}{n} N_n^{\mathrm{V,str}} < \int_0^1 (F(\beta) - F(c))^+ d\beta + 4\epsilon \tag{24}$$

from some n on almost surely.

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

$$\begin{aligned}
&(F(c) - \mathbf{S}(\delta) - 2\epsilon + 1 - F(c + 2\epsilon) - 2\epsilon)\, n \\
&= (1 - \mathbf{S}(\delta) + F(c) - F(c + 2\epsilon) - 4\epsilon)\, n
\end{aligned} \tag{25}$$

(see (18) and (23)) from some n on almost surely. The number of n-strange extended examples among them will not exceed

$$\begin{aligned}
&\Big(\epsilon + F(c + 2\epsilon) - F(c) + F(c-) - F(c - 2\epsilon) + \epsilon \\
&+ (F(c) - \mathbf{S}(\delta))(1 - c) + 4\epsilon + \int_0^1 (F(\beta) - F(c))^+ d\beta + 4\epsilon \Big) n \\
&= \Big(F(c + 2\epsilon) - F(c) + F(c-) - F(c - 2\epsilon) \\
&+ (F(c) - \mathbf{S}(\delta))(1 - c) + \int_0^1 (F(\beta) - F(c))^+ d\beta + 10\epsilon \Big) n
\end{aligned} \tag{26}$$

(see (16), (17), (21), and (24)) from some n on almost surely. Combining (25) and (26), we can see that the number of n-strange extended examples among the $\lfloor (1-\mathbf{S}(\delta))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i,\tau_i'),\tau_i'')$ does not exceed

$$\begin{aligned}
&\Bigg(F(c+2\epsilon)-F(c)+F(c-)-F(c-2\epsilon)+(F(c)-\mathbf{S}(\delta))(1-c)\\
&\quad+\int_0^1 (F(\beta)-F(c))^+ d\beta + 10\epsilon\Bigg) n + (F(c+2\epsilon)-F(c)+4\epsilon)\, n\\
&=\Bigg(2(F(c+2\epsilon)-F(c))+(F(c-)-F(c-2\epsilon))+(F(c)-\mathbf{S}(\delta))(1-c)\\
&\qquad\qquad+\int_0^1 (F(\beta)-F(c))^+ d\beta + 14\epsilon\Bigg) n
\end{aligned}$$

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

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

which completes the proof. ∎

Lemma 6 *Suppose (4) is satisfied. The fraction of n-strange extended examples among the first n extended examples (x_i,τ_i,y_i) approaches δ_0 asymptotically with probability one.*

Proof sketch The lemma is not difficult to prove using McDiarmid's theorem and the fact that, by Lemma 4, $P(\hat{y}_n(x_i,\tau_i') \mid x_i)$ will typically differ little from $f(x_i)$. Notice, however, that the part that we really need in this paper (that the fraction of n-strange extended examples does not exceed $\delta_0 + o(1)$ as $n\to\infty$ with probability one) is just a special case of Lemma 5, corresponding to $\delta=\delta_0$. ∎

Lemma 7 *Suppose (4) is satisfied and $\delta>\delta_0$. The fraction of n-ordinary extended examples among the $\lfloor \mathbf{C}(\delta)n\rfloor$ extended examples (x_i,τ_i,y_i), $i=1,\ldots,n$, with the lowest $(f_n^{\neq}(x_i,\tau_i'),\tau_i'')$ does not exceed $\delta-\delta_0+o(1)$ as $n\to\infty$ with probability one.*

Lemma 7 can be proven analogously to Lemma 5.

Lemma 8 *Let $\mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \cdots$ be a decreasing sequence of σ-algebras and $\xi_1,\xi_2\ldots$ be a bounded adapted (in the sense that ξ_n is $\mathcal{F}_n$-measurable for all n) sequence of random variables such that*

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

Then

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

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

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i = 0 \quad \text{a.s.} \tag{27}$$

Fix a bounded reverse martingale difference $\xi_1, \xi_2, \ldots$; our goal is to prove (27). By the martingale version of Hoeffding's inequality ([DGL96], Theorem 9.1) applied to the martingale difference $(\xi_i, \mathcal{F}_i)$, $i = n, \ldots, 1$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right| \geq \epsilon\right\} \leq 2e^{-2\epsilon^2 n/(2C)}, \tag{28}$$

where C is an upper bound on $\sup_n |\xi_n|$. Combined with the Borel-Cantelli-Lévy lemma, (28) implies (27). ∎

Now we can sketch the proof of Proposition 5. Define $\mathcal{F}_n$, $n = 1, 2, \ldots$, to be the σ-algebra on $\tilde{\mathbf{Z}}^\infty$ generated by the bag of the first $n-1$ extended examples (x_i, τ_i, y_i), $i = 1, \ldots, n-1$, and the sequence of extended examples (x_i, τ_i, y_i), $i = n, n+1, \ldots$ (starting from the nth extended example).

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

Let (x_k, τ_k, y_k) be the extended example with the $(\lfloor (\delta - \epsilon/2)n \rfloor + 1)$th largest (in the sense of the lexicographic order) $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ among all n-strange extended examples (x_i, τ_i, y_i), $i = 1, \ldots, n$. (The simple case where (x_k, τ_k, y_k) does not exist needs to be considered separately.) Let (x_j, τ_j, y_j) be one of the $\lfloor (1 - \mathbf{S}(\delta - \epsilon))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), \tau_i'')$ and let $y \in \mathbf{Y}$ be a label different from $\hat{y}_n(x_j, \tau_j')$. It suffices to prove that

$$\tau_j'' \geq \frac{\#\{i = 1, \ldots, n : \alpha_i^y \geq \alpha_j^y\} - n\delta}{\#\{i = 1, \ldots, n : \alpha_i^y = \alpha_j^y\}} \tag{29}$$

(cf. (3) on p. 5), where all α^y are computed as α in (6) from the sequence $(x_1, \tau_1, y_1), \ldots, (x_n, \tau_n, y_n)$ with y_j replaced by y. It will be more convenient to write (29) in the form

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

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

$$\#\{i : \alpha_i > f_n^{\neq}(x_j, \tau_j')\} + (1 - \tau_j'')\#\{i : \alpha_i = f_n^{\neq}(x_j, \tau_j')\} \leq n(\delta - \epsilon^*), \tag{30}$$

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

Since $(f_n^{\neq}(x_j, \tau_j'), \tau_j'') \geq (\alpha_k, \tau_k'')$ (indeed, by Lemma 5, there are less that $(\delta - \epsilon/2)n$ n-strange extended examples among the $\lfloor (1 - \mathbf{S}(\delta - \epsilon))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), \tau_i''))$, (30) will follow from

$$\#\{i : \alpha_i > \alpha_k\} + (1 - \tau_k'')\#\{i : \alpha_i = \alpha_k\} \leq n(\delta - \epsilon^*). \tag{31}$$

If $\#\{i : \alpha_i = \alpha_k\} \leq \frac{\epsilon}{3}n$, the left-hand side of (31) does not exceed

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

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

$$\#\{i : \alpha_i = \alpha_k\} > \frac{\epsilon}{3} n. \tag{32}$$

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

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

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

$$1 - \tau_k'' \leq \frac{(\delta - \epsilon/2)n - \#\{i : \alpha_i > \alpha_k\}}{\#\{i : \alpha_i = \alpha_k\}} + \epsilon^* \tag{33}$$

(remember (32)). Equation (31) will hold because its left-hand side can be transformed using (33) as

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

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

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

Since ϵ can be arbitrarily small and $\mathbf{S}$ is continuous ([Vov02], Lemma 2), this implies

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

By Lemma 8 this implies, in turn,

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

which coincides with (8).

If $\delta \geq \delta_0$, Lemma 6 implies that

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

in combination with Lemma 8 this again implies (8).

Inequality (9) is treated in a similar way. Lemmas 6 and 7 imply that

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

(this inequality is vacuously true when $\delta \leq \delta_0$). Another application of Lemma 8 gives

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

i.e., (9).

Remark The derivation of Proposition 5 from Lemmas 5–8 would be very simple if we defined the individual strangeness measure by, say,

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

(with the lexicographic order on α's) instead of (6) (in which case the denominator of (3) would be 1 almost surely). Our definition (6), however, is simpler and, most importantly, facilitates the proof of Proposition 2.

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").

Additional references

[CLRS01] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

[DGKL94] 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.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[Shi96] Albert N. Shiryaev. *Probability.* Springer, New York, second edition, 1996.

[Vov02] Vladimir Vovk. Asymptotic optimality of Transductive Confidence Machine. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory*, 2002.