3

Classification with conformal predictors

In this chapter we concentrate on the problem of classification, where the label space $\mathbf{Y}$ is finite. We start in §3.1 by giving two more examples of nonconformity measures, this time specifically for the case of classification, and reporting on the empirical performance of one of them. In the next section, §3.2, we state the main result of the chapter: there exists a "universal" smoothed conformal predictor whose asymptotic efficiency[1] is not worse than that of any other asymptotically valid randomized confidence predictor, regardless of the probability distribution Q generating individual examples. In particular, even if for a given probability distribution Q we construct the optimal, or "Bayes"[2], confidence predictor Γ, our universal predictor will be as efficient as Γ asymptotically, even though the former "knows nothing" about Q. In §3.4 we make the first step towards the proof of the main result looking closely at the Bayes confidence predictor; this will allow us to set the target for the universal predictor. The universal smoothed conformal predictor is constructed in §3.3. As usual, most of the actual proofs will be given in a separate section, §3.5.

The learning protocol of this chapter is the same as in Chap. 2, but we will state it again this time including not only the variables Err_n^ϵ (the total number of errors made up to and including trial n at significance level ϵ) and err_n^ϵ (the binary variable showing whether an error is made at trial n) but also the analogous variables Mult_n^ϵ, mult_n^ϵ, Emp_n^ϵ, emp_n^ϵ for multiple (containing more than one label) and empty (containing no labels) predictions:

```
$\mathrm{Err}_0^\epsilon := 0$, $\mathrm{Mult}_0^\epsilon := 0$, $\mathrm{Emp}_0^\epsilon := 0$ for all $\epsilon \in (0,1)$;
FOR $n = 1, 2, \dots$:
    Reality outputs $x_n \in \mathbf{X}$;
```

[1] We use the expression "asymptotic efficiency" only informally, to refer to the asymptotic optimality of the predictions made. In this book we never consider the (very interesting) question of how fast optimality is approached.

[2] We are following standard usage (see Devroye et al. 1996), despite the lack of connection with Bayesian learning (as discussed in, e.g., Chap. 10).

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

In this chapter, the label space $\mathbf{Y}$ is finite and equipped with the discrete σ-algebra.

3.1 More ways of computing nonconformity scores

First of all we notice that the general scheme discussed at the end of §2.2 is applicable generally, including the case of classification. For classification, it is especially important to allow the case $\hat{\mathbf{Y}} \neq \mathbf{Y}$.

Another general remark is that any procedure of computing nonconformity scores for regression can be used for computing nonconformity scores in binary classification (and there are standard ways to reduce general classification problems to binary ones, as we will see at the end of this section). Indeed, if $\mathbf{Y}$ consists of just two elements, we can encode them by two different real numbers and run the regression procedure for computing nonconformity scores. In particular, we can use the nonconformity scores produced by ridge regression and by nearest neighbors regression, as discussed in the previous chapter, in classification problems.

Nonconformity scores from nearest neighbors

There is, however, a much more direct way of applying the nearest neighbors idea to obtain a nonconformity measure: assuming the objects are vectors in a Euclidean space, the nonconformity scores can be defined, in the spirit of the 1-nearest neighbor algorithm, as

$$A(\Lbag (x_1, y_1), \dots, (x_l, y_l) \Rbag, (x, y)) := \frac{\min_{i=1,\dots,l: y_i = y} d(x, x_i)}{\min_{i=1,\dots,l: y_i \neq y} d(x, x_i)}, \tag{3.1}$$

where d is the Euclidean distance (i.e., an object is considered nonconforming if it is close to an object labeled in a different way and far from any object

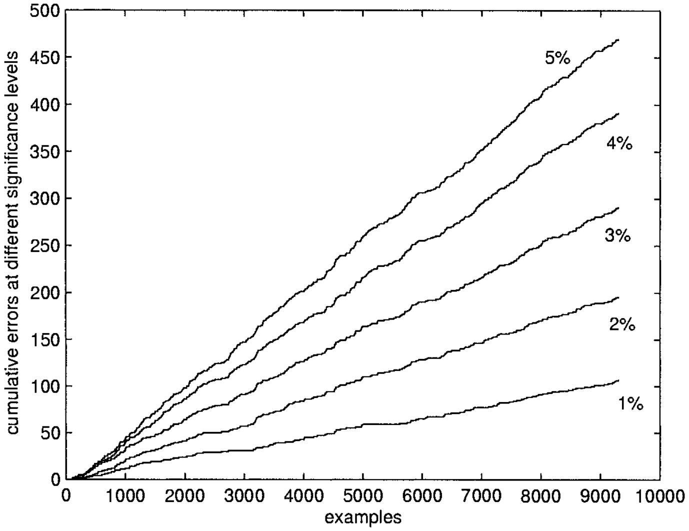

Fig. 3.1. Cumulative errors Err_n^ϵ suffered by the 1-nearest neighbor conformal predictor on the USPS data set (9298 hand-written digits, randomly permuted) plotted against n for the significance levels from $\epsilon = 1\%$ to $\epsilon = 5\%$

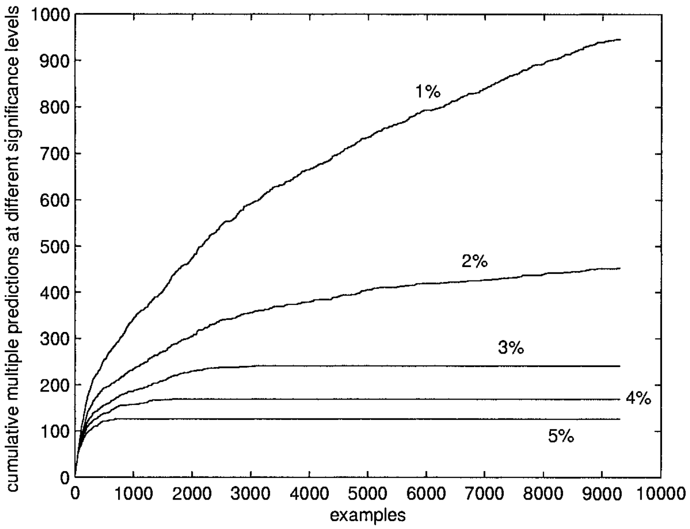

Fig. 3.2. Cumulative number of multiple predictions Mult_n^ϵ output by the 1-nearest neighbor conformal predictor on the USPS data set

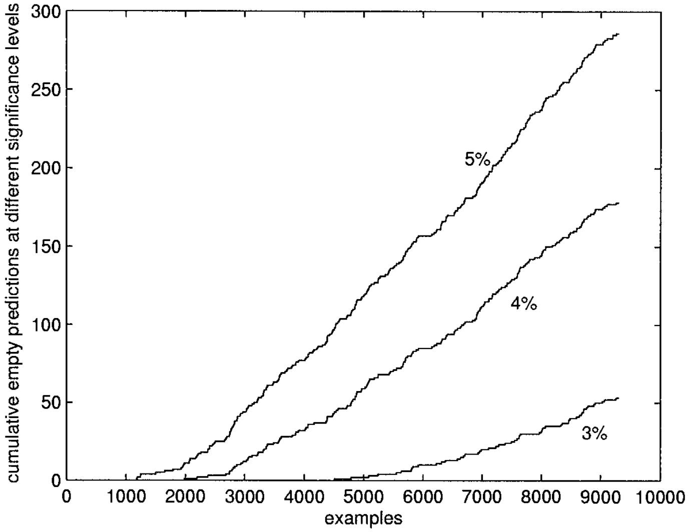

Fig. 3.3. Cumulative number of empty predictions Emp_n^ϵ output by the 1-nearest neighbor conformal predictor on the USPS data set

labeled in the same way). It is possible for (3.1) to be equal to ∞ (if the denominator in (3.1) is zero).

Figures 3.1–3.3 show the on-line performance of the *1-nearest neighbor conformal predictor* (determined by (3.1)) on the USPS data set (the original 9298 hand-written digits, as described in Appendix B, but randomly permuted) for the confidence levels 95–99%. Figure 3.1 again confirms empirically the validity of conformal predictors; the cumulative numbers of errors at $\epsilon = 1\%$ and $\epsilon = 5\%$ were already given in Chap. 1 (Fig. 1.5 on p. 10). Figures 3.2 and 3.3 show that for a vast majority of examples the prediction set contains precisely one label at the considered significance levels. Figure 3.4 illustrates a feature of Figs. 3.2 and 3.3 that is not very noticeable since it requires examination of both figures simultaneously: at a fixed significance level, empty predictions appear only after multiple predictions disappear. (This figure cannot be directly compared to the error rate of 2.5% for humans reported in Vapnik 1998, since our 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.)

Nonconformity scores from support vector machines

Support vector machines were proposed by Vapnik (1998, Part II); the standard abbreviation of "support vector machine" is SVM. We concentrate on the problem of binary classification, assuming that the set $\mathbf{Y}$ of possible labels

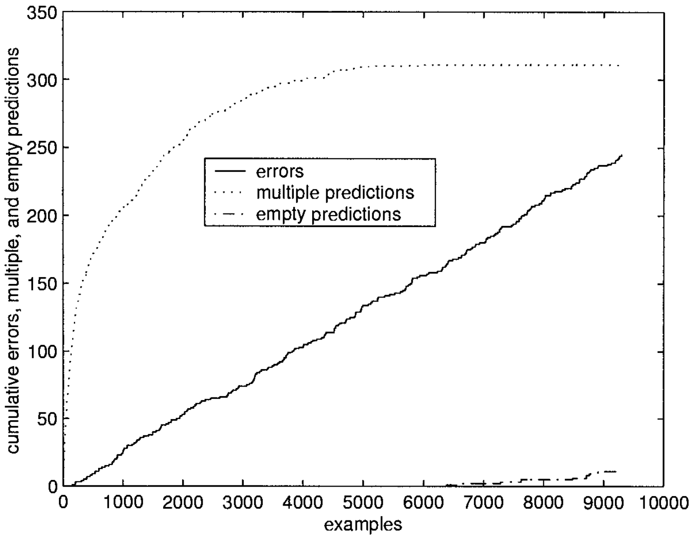

Fig. 3.4. On-line performance of the 1-nearest neighbor conformal predictor on the USPS data set for the significance level 2.5%. The solid line shows the cumulative number of errors, dotted the cumulative number of multiple predictions, and dashdot the cumulative number of empty predictions

is $\{-1, 1\}$. For a given bag

$$\Lbag z_1, \dots, z_n \Rbag = \Lbag (x_1, y_1), \dots, (x_n, y_n) \Rbag$$

and its element z_i, each SVM, to be defined shortly, provides a very natural definition of the nonconformity score

$$\alpha_i = A_n \left(\Lbag z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n \Rbag, z_i \right) . \tag{3.2}$$

Suppose the objects are vectors in a dot product space $\mathbf{H}$ and consider the quadratic optimization problem

$$\frac{1}{2}(w \cdot w) + C \left(\sum_{i=1}^{n} \xi_i \right) \to \min , \tag{3.3}$$

where C is an *a priori* fixed positive constant and the variables $w \in \mathbf{H}$, $\xi = (\xi_1, \dots, \xi_n)' \in \mathbb{R}^n$, $b \in \mathbb{R}$ (the last variable not entering (3.3)) are subject to the constraints

$$y_i \left(w \cdot x_i + b \right) \geq 1 - \xi_i, \quad i = 1, \dots, n, \tag{3.4}$$

$$\xi_i \geq 0, \qquad i = 1, \dots, n . \tag{3.5}$$

If this optimization problem has a solution, it is unique and is also denoted $w, (\xi_1, \dots, \xi_n)', b$. The hyperplane $w \cdot x + b = 0$, called the *optimal separating hyperplane*, is the boundary of the prediction rule produced by the corresponding SVM: the prediction $\hat{y}$ for a new object x is 1 if $w \cdot x + b > 0$ and -1 if $w \cdot x + b < 0$ (with an arbitrary convention if $w \cdot x + b = 0$).

The next step in the development of SVM is to consider an arbitrary object space $\mathbf{X}$ (not necessarily a linear space) and apply a transformation $F : \mathbf{X} \to \mathbf{H}$ mapping the objects x_i into the "feature vectors" $F(x_i) \in \mathbf{H}$, where $\mathbf{H}$ is a dot product space. This replaces x_i by $F(x_i)$ in the optimization problem (3.3)–(3.5); as before, w ranges over $\mathbf{H}$, but the latter is now different from the object space. After that the Lagrange method is applied to the modified problem; to each inequality in (3.4) corresponds a Lagrange multiplier α_i. The optimal values of α_i, obtained by solving the *dual problem*

$$\begin{gathered} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \mathcal{K}(x_i, x_j) \to \max, \\ \sum_{i=1}^{n} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n , \end{gathered} \tag{3.6}$$

where $\mathcal{K}(x_i, x_j) := F(x_i) \cdot F(x_j)$, can be interpreted as follows:

- the examples z_i with $\alpha_i = 0$ are typical;
- the examples with $\alpha_i = C$ are the most extreme (under the given choice of C) outliers;
- the examples with $0 < \alpha_i < C$ are intermediate, with a possible interpretation of α_i as a measure of nonconformity of the corresponding example.

This makes the solutions to the dual problem ideal for use as nonconformity scores (3.2).

Remark Actually, it is quite possible that the Lagrange multipliers computed by a given computer implementation of SVM will not provide a bona fide nonconformity measure, with α_i in (3.2) depending on the order in which the examples $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ are presented. The order of the examples may be especially important in so-called "chunking", a standard feature of SVM implementations. To ensure the invariance of α_i w.r.t. to permutations of $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ (and so the validity of the resulting conformal predictor), the examples $z_1, \dots, z_n$ can be sorted in some way (e.g., in the lexicographic order of their ASCII representations) obtaining $z_{\pi(1)}, \dots, z_{\pi(n)}$, where π is a permutation of the set $\{1, \dots, n\}$. The Lagrange multipliers computed from $z_{\pi(1)}, \dots, z_{\pi(n)}$ should then be permuted using π^{-1} to obtain $\alpha_1, \dots, \alpha_n$.

Reducing classification problems to the binary case

The original SVM method can only deal with binary classification problems, but we will now see that there are ways to use it for solving *multilabel* classification problems (i.e., those with $|\mathbf{Y}| > 2$).

There are two standard ways to reduce multilabel classification problems to the binary case: the "one-against-the-rest" procedure and the "one-against-one" procedure. Suppose we have a reasonable nonconformity measure A for binary classification but are confronted with a multilabel classification problem. For concreteness we will assume that the label space in the binary classification problem is $\{0,1\}$; if it is $\{a,b\}$ (e.g., $a=-1$ and $b=1$ in the case of SVM), the reduction will be achieved by a further scaling, $y \mapsto a+(b-a)y$.

The *one-against-the-rest procedure* gives the nonconformity measure

$$\begin{aligned} A^{(1)}&\left(\Lbag(x_1,y_1),\dots,(x_l,y_l)\Rbag,(x,y)\right)\\ &:= \lambda A\left(\Lbag(x_1,\mathbb{I}_{y_1=y}),\dots,(x_l,\mathbb{I}_{y_l=y})\Rbag,(x,1)\right)\\ &+\frac{1-\lambda}{|\mathbf{Y}|-1}\sum_{y'\neq y} A\left(\Lbag(x_1,\mathbb{I}_{y_1=y'}),\dots,(x_l,\mathbb{I}_{y_l=y'})\Rbag,(x,0)\right)\ , \end{aligned} \tag{3.7}$$

where $\lambda \in [0,1]$ is a constant (parameter of the procedure) and $\mathbb{I}$ is the indicator function (i.e., $\mathbb{I}_E = 1$ if E holds and $\mathbb{I}_E = 0$ if not). Intuitively, we consider $|\mathbf{Y}|$ auxiliary binary classification problems and compute the nonconformity score of an example (x,y) as the weighted average of the scores this example receives in the auxiliary problems.

The *one-against-one procedure* gives the nonconformity measure

$$A^{(2)}\left(\Lbag(x_1,y_1),\dots,(x_l,y_l)\Rbag,(x,y)\right) := \frac{1}{|\mathbf{Y}|-1}\sum_{y'\neq y} A\left(B_{y,y'},(x,1)\right)\ , \tag{3.8}$$

where $B_{y,y'}$ is the bag obtained from $\Lbag(x_1,y_1),\dots,(x_l,y_l)\Rbag$ as follows: remove all (x_i,y_i) with $y_i \notin \{y,y'\}$; replace each (x_i,y) by $(x_i,1)$; replace each (x_i,y') by $(x_i,0)$. We now have $|\mathbf{Y}|-1$ auxiliary binary classification problems.

The numbers $|\mathbf{Y}|$ and $|\mathbf{Y}|-1$ of auxiliary binary classification problems given above refer to computing only one nonconformity score. When using nonconformity measures for conformal prediction, we have to compute all n nonconformity scores (2.19) (p. 26) for all $y \in \mathbf{Y}$. With the one-against-the-rest procedure, we have to consider $2|\mathbf{Y}|$ auxiliary binary classification problems altogether, whereas with the one-against-one procedure $|\mathbf{Y}|(|\mathbf{Y}|-1)$ auxiliary binary classification problems are required. When $|\mathbf{Y}|=3$, the numbers of auxiliary problems coincide, $2|\mathbf{Y}| = |\mathbf{Y}|(|\mathbf{Y}|-1)$, but for $|\mathbf{Y}|>3$ the one-against-the-rest procedure requires fewer auxiliary problems, $2|\mathbf{Y}| < |\mathbf{Y}|(|\mathbf{Y}|-1)$.

3.2 Universal predictor

We first describe the main idea of a universal confidence predictor. Let us fix an exchangeable probability distribution P on $\mathbf{Z}^\infty$ generating the examples $z_1, z_2, \dots$, and let us fix a significance level ϵ. Remember that $\mathbf{Y}$ is finite and $|\mathbf{Y}|>1$.

Slightly elaborating on the notion introduced in Chap. 2, we say that a confidence predictor is *asymptotically conservative for* P and ϵ if the long-run frequency of errors does not exceed ϵ almost surely w.r.t. to P; we know that each conformal predictor satisfies this property. For asymptotically conservative predictors we take the number Mult_n of multiple predictions as the principal measure of predictive efficiency. The main result of this chapter is the construction of a confidence predictor (a smoothed conformal predictor) which, for any (unknown) P and any ϵ: (a) makes errors independently and with probability ϵ at every trial (in particular, is asymptotically conservative for P and ϵ); (b) makes in the long run no more multiple predictions than any other randomized confidence predictor that is asymptotically conservative for P and ϵ; (c) processes example n in time $O(\log n)$.

There is a slight complication for item (b), dealing with predictive efficiency: we also have to deal carefully with empty predictions. The full picture is that our universal predictor, for any significance level ϵ and without knowing the true distribution P generating the examples:

- produces, asymptotically, no more multiple predictions than any other randomized confidence predictor that is asymptotically conservative for P and ϵ;
- produces, asymptotically, at least as many empty predictions as any other randomized confidence predictor that is asymptotically conservative for P and ϵ and whose percentage of multiple 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 multiple predictions. This criterion ceases to work, however, when the number of multiple predictions stabilizes, as in the case of significance levels 3%–5% in Fig. 3.2. In such cases the number of empty predictions becomes important: empty predictions (automatically leading to an error) provide a warning that the object is untypical (looks very different from the previous objects), 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 ϵ in the long run.

We now start the formal exposition, only considering randomized confidence predictors. We will often use the notation mult_n^ϵ, emp_n^ϵ, etc., in the case where Predictor and Reality are using given randomized strategies, as was already done in the previous chapter for err_n^ϵ and Err_n^ϵ; for example, $\mathrm{mult}_n^\epsilon(\Gamma, P)$ is the random variable equal to 1 if Predictor makes a multiple prediction 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 Reality are independent.

We say that a randomized confidence predictor Γ is *asymptotically conservative for* a probability distribution P on $\mathbf{Z}^\infty$ and a significance level $\epsilon \in (0,1)$ if

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

We say that Γ is *asymptotically optimal* for P and ϵ if, for any randomized confidence predictor $\Gamma^\dagger$ which is asymptotically conservative for P and ϵ,

$$\limsup_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma, P)}{n} \le \liminf_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma^\dagger, P)}{n} \quad \text{a.s.} \tag{3.9}$$

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

A randomized confidence predictor Γ is a *universal predictor* if:

- it is asymptotically conservative for any exchangeable P and ϵ;
- it is asymptotically optimal for any exchangeable P and ϵ;
- for any exchangeable P, any ϵ, and any randomized confidence predictor $\Gamma^\dagger$ which is asymptotically conservative and optimal for P and ϵ,

$$\liminf_{n\to\infty} \frac{\mathrm{Emp}_n^\epsilon(\Gamma, P)}{n} \ge \limsup_{n\to\infty} \frac{\mathrm{Emp}_n^\epsilon(\Gamma^\dagger, P)}{n} \quad \text{a.s.}$$

Now we can state the main result of this chapter.

Theorem 3.1. *Suppose the object space* $\mathbf{X}$ *is Borel. There exists a universal predictor.*

In the next section we construct a universal predictor (processing example n in time $O(\log n)$).

3.3 Construction of a universal predictor

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.

In this chapter we will especially often apply our procedures (such as nonconformity measures and prediction rules) not to the original objects $x \in \mathbf{X}$ but to the 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). Along with examples (x, y) we will thus 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

$$|x_1 - x_3, \sigma_1 - \sigma_3| < |x_2 - x_3, \sigma_2 - \sigma_3| \,, \tag{3.10}$$

where

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

If (x_1, σ_1) and (x_2, σ_2) are extended objects, we will sometimes refer to $|x_1 - x_2, \sigma_1 - \sigma_2|$ as the *distance* between (x_1, σ_1) and (x_2, σ_2), even though this distance is a two-dimensional object (what is important is that the distances are linearly ordered according to (3.10)).

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., Devroye et al. 1996, Chap. 11); the random components σ are needed for tie-breaking. It is still theoretically possible for the expression "the kth nearest neighbor" not to have a precise meaning: two extended objects in a training set can be at the same distance from a given extended object. This case, which happens with probability zero, will be always treated separately.

Conformal prediction in the current context

The smoothed conformal predictors we are going to construct will work on extended examples; otherwise it will be our standard notion. (It might have been better to call them "doubly smoothed conformal predictors", but we will not make such fine distinctions.) Therefore, a nonconformity measure is a mapping $A : \tilde{\mathbf{Z}}^{(*)} \times \tilde{\mathbf{Z}} \to \overline{\mathbb{R}}$. The smoothed conformal predictor determined by the nonconformity measure A is the following randomized confidence predictor

$$\Gamma^\epsilon(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) \;: \tag{3.11}$$

at each trial n and for each $y \in \mathbf{Y}$, define

$$\begin{aligned} \alpha_i := A\Big(&\Lbag (x_1, \tau'_1, y_1), \ldots, (x_{i-1}, \tau'_{i-1}, y_{i-1}), \\ &(x_{i+1}, \tau'_{i+1}, y_{i+1}), \ldots, (x_{n-1}, \tau'_{n-1}, y_{n-1}), (x_n, \tau'_n, y) \Rbag, (x_i, \tau'_i, y_i)\Big), \\ &\qquad\qquad\qquad i = 1, \ldots, n-1\,, \end{aligned}$$

and

$$\alpha_n := A\left(\Lbag (x_1, \tau'_1, y_1), \ldots, (x_{n-1}, \tau'_{n-1}, y_{n-1}) \Rbag, (x_n, \tau'_n, y)\right)\;;$$

include y in (3.11) if and only if

$$\frac{|\{i = 1, \ldots, n : \alpha_i > \alpha_n\}| + \tau''_n\, |\{i = 1, \ldots, n : \alpha_i = \alpha_n\}|}{n} > \epsilon\,. \tag{3.12}$$

We already know that every (smoothed) conformal predictor is asymptotically conservative for every P and ϵ.

Universal predictor

Fix a strictly increasing sequence of integers K_n, $n = 1, 2, \dots$, such that

$$K_n \to \infty,\ K_n = o\left(\sqrt{n/\ln n}\right) \tag{3.13}$$

as $n \to \infty$. Let $B = \lbag w_1, \dots, w_n \rbag$ be a bag of extended examples $w_i = (x_i, \sigma_i, y_i)$. The nearest neighbors approximation $\hat{Q}_B(y \mid x, \sigma)$ to the true (but unknown) conditional probability that an object x's label is y is defined as

$$\hat{Q}_B(y \mid x, \sigma) := N_B(x, \sigma, y)/K_n\ , \tag{3.14}$$

where $n := l + 1$ and $N_B(x, \sigma, y)$ is the number of $i = 1, \dots, l$ such that $y_i = y$ and (x_i, σ_i) is one of the K_n nearest neighbors of (x, σ) in the sequence $((x_1, \sigma_1), \dots, (x_l, \sigma_l))$. If $K_n \geq n$ or $K_n \leq 0$, this definition does not work, so set, e.g., $\hat{Q}_B(y \mid x, \sigma) := 1/|\mathbf{Y}|$ for all y and (x, σ) (this particular convention is not essential since, by (3.13), $0 < K_n < n$ from some n on). It is also possible that the phrase "K_n nearest neighbors of (x, σ)" is not defined because of distance ties; in this case we again set $\hat{Q}_B(y \mid x, \sigma) := 1/|\mathbf{Y}|$ for all y.

Define the "empirical predictability function" $\hat{f}_B$ by

$$\hat{f}_B(x, \sigma) := \max_{y \in \mathbf{Y}} \hat{Q}_B(y \mid x, \sigma)\ .$$

For all B and (x, σ) fix some

$$\hat{y}_B(x, \sigma) \in \arg\max_y \hat{Q}_B(y \mid x, \sigma)$$

(e.g., take the first element of $\arg\max_y \hat{Q}_B(y \mid x, \sigma)$ in a fixed ordering of $\mathbf{Y}$) and define the *nearest neighbors nonconformity measure* by

$$A(B, (x, \sigma, y)) := \begin{cases} -\hat{f}_B(x, \sigma) & \text{if } y = \hat{y}_B(x, \sigma) \\ \hat{f}_B(x, \sigma) & \text{otherwise}\ , \end{cases} \tag{3.15}$$

B ranging over the bags of extended examples. The *nearest neighbors smoothed conformal predictor* is defined to be the smoothed conformal predictor determined by the nearest neighbors nonconformity measure. The nearest neighbors smoothed conformal predictor will later be shown to be universal.

Proposition 3.2. *Let $\{\epsilon_1, \dots, \epsilon_K\} \subseteq (0, 1)$ be a finite set. If $\mathbf{X} = [0, 1]$ and $K_n \to \infty$ sufficiently slowly, the nearest neighbors smoothed conformal predictor can be implemented for significance levels $\epsilon = \epsilon_1, \dots, \epsilon_K$ so that computations at trial n are performed in time $O(\log n)$.*

Proposition 3.2 assumes a computational model that allows operations (such as comparison) with real numbers. If $\mathbf{X}$ is an arbitrary Borel space, for this proposition to be applicable $\mathbf{X}$ should be embedded in $[0, 1]$ first; e.g., if $\mathbf{X} \subseteq [0, 1]^n$, an $x = (x_1, \dots, x_n) \in \mathbf{X}$ can be represented as

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

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

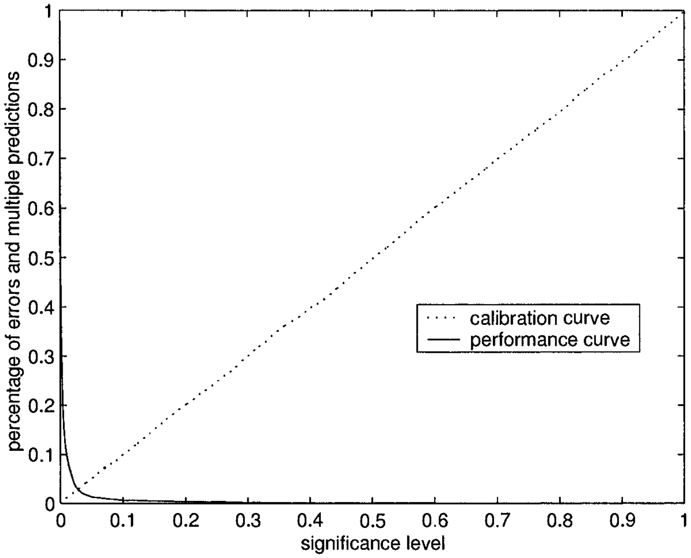

Fig. 3.5. The empirical calibration and performance curves for the 1-nearest neighbor conformal predictor on the USPS data set (randomly permuted)

3.4 Fine details of confidence prediction

In this section we make first steps towards the proof of Theorem 3.1. By de Finetti's theorem (see §A.5), each exchangeable distribution on $\mathbf{Z}^\infty$ (which is a Borel space as long as $\mathbf{Z}$ is Borel: see, e.g. Schervish 1995, Lemma B.41) is a mixture of power distributions. Therefore, without loss of generality we assume that $P = Q^\infty$ for a probability distribution Q on $\mathbf{Z}$.

To provide the reader with extra intuition about confidence prediction in the case of classification, we first briefly discuss further empirical results for the 1-nearest neighbor conformal predictor and the USPS data set. Recall that Figs. 3.1–3.3 show the cumulative number of errors, the cumulative number of multiple predictions, and that of empty predictions. Figure 3.5 gives the *empirical calibration curve*

$$\epsilon \mapsto \frac{\mathrm{Err}^\epsilon_N(\Gamma, \mathrm{USPS})}{N} \tag{3.16}$$

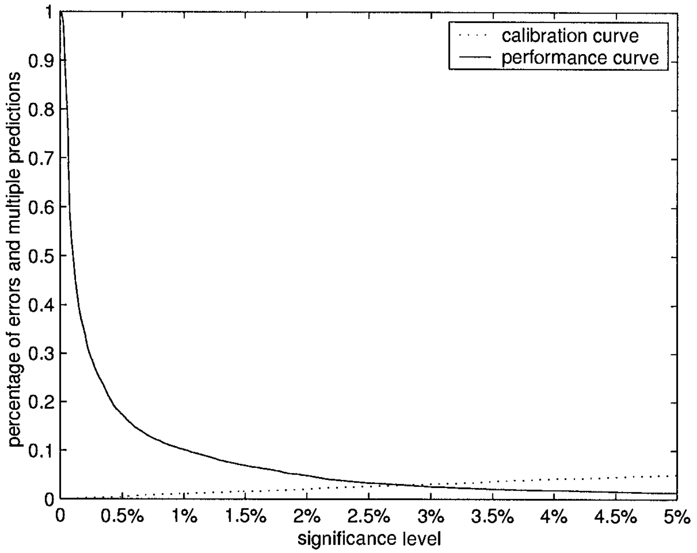

Fig. 3.6. The left edge of the previous figure stretched horizontally

and the *empirical performance curve*

$$\epsilon \mapsto \frac{\mathrm{Mult}_N^{\epsilon}(\Gamma, \mathrm{USPS})}{N}$$

for this confidence predictor; we use the strategies followed by Reality (the USPS data set, randomly permuted) and Predictor as arguments for Err and Mult. Remember that the size of the USPS data set is $N = 9298$.

We denote by $Q_{\mathbf{X}}$ the marginal distribution of Q on $\mathbf{X}$ (i.e., $Q_{\mathbf{X}}(E) := Q(E \times \mathbf{Y})$) and by $Q_{\mathbf{Y}|\mathbf{X}}(y \mid x)$ the conditional probability that, for a random example (X, Y) drawn from Q, $Y = y$ provided $X = x$ (we fix arbitrarily a regular version of this conditional probability; the existence of regular conditional probability is obvious in our case of finite $\mathbf{Y}$ and also follows from general results: see §A.3). We will often omit lower indices $\mathbf{X}$ and $\mathbf{Y} \mid \mathbf{X}$.

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

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

and the *predictability distribution function* is the increasing[3] function $F : [0,1] \to [0,1]$ defined by

[3] In this book "increasing" and "decreasing" are used in Bourbaki's weak sense: e.g., $F(\beta)$ is called increasing if $F(\beta_1) \le F(\beta_2)$ whenever $\beta_1 \le \beta_2$.

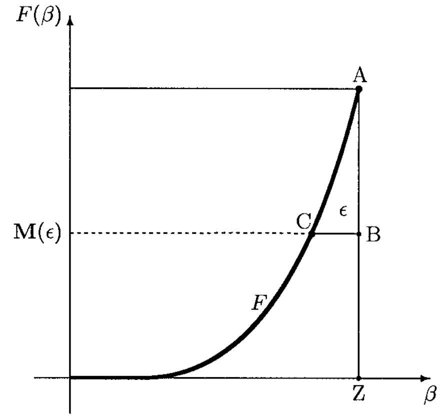

Fig. 3.7. The predictability distribution function F and how it determines the multiplicity curve $\mathrm{M}(\epsilon)$. The function F is increasing, continuous on the right, and $F(1/|\mathbf{Y}|-) = 0$. For a possibly more realistic example of a predictability distribution function, see Fig. 3.12

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

(essentially, it is the distribution function of the image Qf^{-1} of the probability distribution Q under the mapping f). An example of such a function F is given in Fig. 3.7; the graph of F is the thick line, and the unit box is also shown. (The intuition behind some constructions in this chapter will become clearer if the case of finite $\mathbf{X}$ with equiprobable objects is considered first; see Fig. 3.8.)

The *multiplicity curve* $\mathbf{M} = \mathbf{M}_Q$ of Q is defined by the equality

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

where t^+ stands for $\max(t, 0)$; the function $\mathbf{M}$ is also of the type $[0,1] \to [0,1]$. (Why the terminology introduced here and below is natural will become clear from Propositions 3.3 and 3.5.) Geometrically, $\mathbf{M}$ is defined from the graph of F as follows (see Fig. 3.7): move the point B from A to Z until the area of the curvilinear triangle ABC becomes ϵ (assuming this area does become ϵ eventually, i.e., ϵ is not too large); the ordinate of B is then $\mathrm{M}(\epsilon)$. The intuition in the case of finite $\mathbf{X}$ (see Fig. 3.8) is that $1 - \mathrm{M}(\epsilon)$ is the maximum fraction of objects that are "easily predictable" in the sense that their cumulative lack of predictability does not exceed ϵ (where the lack of predictability $1 - f(x)$ of each object is taken with the weight $1/|\mathbf{X}|$).

The *emptiness curve* $\mathbf{E} = \mathbf{E}_Q$ of Q is defined by

$$\mathrm{E}(\epsilon) = \sup\left\{B \in [0,1] : B + \int_0^1 (F(\beta) - B)^+ \mathrm{d}\beta \le \epsilon\right\},$$

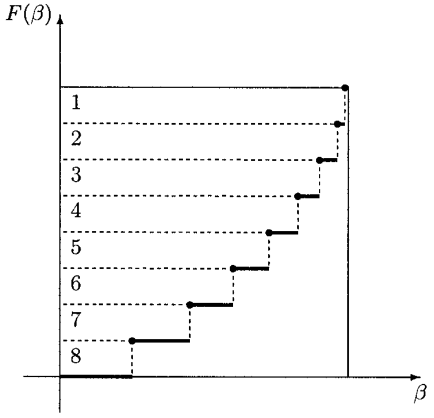

Fig. 3.8. The predictability distribution function (thick line) in the case where the object space $\mathbf{X}$ is finite and all objects $x \in \mathbf{X}$ have the same probability. The objects are numbered, from 1 to 8, in the order of decreasing predictability (equal to the length of the corresponding rectangle)

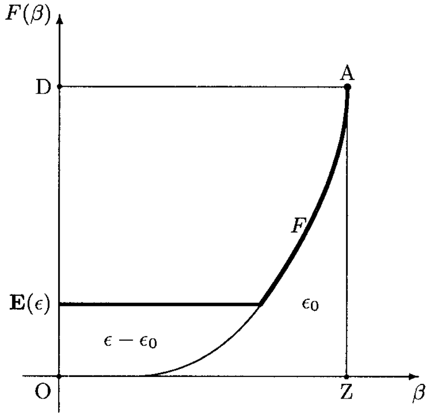

Fig. 3.9. The predictability distribution function F and how it determines the emptiness curve $\mathbf{E}(\epsilon)$

with $\sup \emptyset$ interpreted as 0. Similarly to the case of $\mathbf{M}(\epsilon)$, $\mathbf{E}(\epsilon)$ is defined as the value such that the area of the part of the box AZOD below the thick line in Fig. 3.9 is ϵ ($\mathbf{E}(\epsilon) = 0$ if such a value does not exist).

Define the *critical significance level* ϵ_0 as

$$\epsilon_0 := \int_0^1 F(\beta)\mathrm{d}\beta \tag{3.17}$$

(the area under the thick curve in Fig. 3.7; we will later see that this coincides with what is sometimes called the *Bayes error* – see, e.g., Devroye et al. 1996, §2.1). It is clear that

$$\epsilon \le \epsilon_0 \Longrightarrow \int_0^1 (F(\beta) - \mathbf{M}(\epsilon))^+ \mathrm{d}\beta = \epsilon \;\&\; \mathbf{E}(\epsilon) = 0$$

$$\epsilon \ge \epsilon_0 \Longrightarrow \mathbf{M}(\epsilon) = 0 \;\&\; \mathbf{E}(\epsilon) + \int_0^1 (F(\beta) - \mathbf{E}(\epsilon))^+ \mathrm{d}\beta = \epsilon\,.$$

So far we have defined some characteristics of the distribution Q itself; now we will give definitions pertaining to individual confidence predictors. The most natural class of confidence predictors consists of what we called in Chap. 2 *invariant confidence predictors*: those confidence predictors Γ for which $\Gamma^\epsilon(z_1, \dots, z_l, x)$ does not depend on the order of $z_1, \dots, z_l$. This includes the definition of randomized confidence predictors as a special case (where z_i range over $\mathbf{X} \times [0,1] \times \mathbf{Y}$ instead of $\mathbf{Z}$ and x ranges over $\mathbf{X} \times [0,1]$ instead of $\mathbf{X}$).

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

$$\mathbf{C}_{\Gamma,Q}(\epsilon) := \inf\left\{\beta : \mathbb{P}\left\{\limsup_{n\to\infty} \frac{\mathrm{Err}_n^\epsilon(\Gamma, Q^\infty)}{n} \le \beta\right\} = 1\right\} \tag{3.18}$$

($\mathbb{P}(E)$ stands for the probability of event E). By the Hewitt–Savage zero-one law (see, e.g., Shiryaev 1996, Theorem IV.1.3) in the case of invariant predictors this definition will not change if "$= 1$" is replaced by "> 0" in (3.18). The *performance curve* of Γ under Q is defined by

$$\mathbf{P}_{\Gamma,Q}(\epsilon) := \inf\left\{\beta : \mathbb{P}\left\{\limsup_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \le \beta\right\} = 1\right\}\;; \tag{3.19}$$

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

Notice that a randomized confidence predictor Γ is asymptotically conservative for Q^∞ and any $\epsilon \in (0,1)$ if its calibration curve $\mathbf{C}_{\Gamma,Q}$ is below the diagonal: $\mathbf{C}_{\Gamma,Q}(\epsilon) \le \epsilon$ for any significance level ϵ. The next proposition shows that it is asymptotically optimal for Q^∞ and any $\epsilon \in (0,1)$ if its performance curve coincides with the multiplicity curve: $\mathbf{P}_{\Gamma,Q}(\epsilon) = \mathbf{M}_Q(\epsilon)$ for all ϵ (and we will later see that Γ is asymptotically optimal for Q^∞ and any $\epsilon \in (0,1)$ only if this condition holds). We will often omit the lower indices in (3.18) and (3.19).

Proposition 3.3. *Let Q be a probability distribution on $\mathbf{Z}$ with the multiplicity curve $\mathbf{M}$ and let $\epsilon \in (0,1)$. If a randomized confidence predictor Γ is asymptotically conservative for Q^∞ and ϵ, its performance curve $\mathbf{P}_{\Gamma,Q}$ is above $\mathbf{M}$ at ϵ: $\mathbf{P}_{\Gamma,Q}(\epsilon) \ge \mathbf{M}(\epsilon)$. Moreover,*

$$\liminf_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \ge \mathbf{M}(\epsilon) \quad a.s. \tag{3.20}$$

Of course, this proposition will continue to hold if the word "randomized" is omitted. The "a.s." in (3.20) refers to the probability distribution $(Q \times \mathbf{U})^\infty$ generating the sequence $z_1, \tau_1, z_2, \tau_2, \ldots$, with $\mathbf{U}$ standing for the uniform distribution on $[0,1]$.

Since we are also interested in the number of empty predictions made, we complement Proposition 3.3 with

Proposition 3.4. *Let Q be a probability distribution on $\mathbf{Z}$ with multiplicity curve $\mathbf{M}$ and emptiness curve $\mathbf{E}$ and let $\epsilon \in (0,1)$ be a significance level. If a randomized confidence predictor Γ is asymptotically conservative for Q and ϵ and satisfies*

$$\limsup_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \le \mathbf{M}(\epsilon) \quad a.s.\,, \tag{3.21}$$

then

$$\limsup_{n\to\infty} \frac{\mathrm{Emp}_n^\epsilon(\Gamma, Q^\infty)}{n} \le \mathbf{E}(\epsilon) \quad a.s.$$

Theorem 3.1 immediately follows from Propositions 3.3, 3.4 and the following proposition.

Proposition 3.5. *Suppose $\mathbf{X}$ is a Borel space. For any $Q \in \mathbf{P}(\mathbf{Z})$ and any significance level ϵ, the nearest neighbors smoothed conformal predictor Γ constructed in §3.3 satisfies*

$$\limsup_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \le \mathbf{M}_Q(\epsilon) \quad a.s. \tag{3.22}$$

and

$$\liminf_{n\to\infty} \frac{\mathrm{Emp}_n^\epsilon(\Gamma, Q^\infty)}{n} \ge \mathbf{E}_Q(\epsilon) \quad a.s. \tag{3.23}$$

Bayes confidence predictor

Let us now assume, for simplicity, that the distribution Q is *regular*, in the sense that the predictability distribution function F is continuous.

In this chapter we prove that one can construct an asymptotically optimal smoothed conformal predictor. If, however, we know for sure that Q is the true distribution on $\mathbf{Z}$, it is very easy to construct an asymptotically conservative and optimal confidence predictor. Fix a *choice function* $\hat{y} : \mathbf{X} \to \mathbf{Y}$ such that

$$\forall x \in \mathbf{X} : f(x) = Q(\hat{y}(x) \mid x) \tag{3.24}$$

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

$$\Gamma^\epsilon(z_1, \ldots, z_l, x) := \begin{cases} \{\hat{y}(x)\} & \text{if } F(f(x)) \ge \max(\mathbf{M}(\epsilon), \mathbf{E}(\epsilon)) \\ \mathbf{Y} & \text{if } F(f(x)) < \mathbf{M}(\epsilon) \\ \emptyset & \text{if } F(f(x)) < \mathbf{E}(\epsilon) \end{cases}$$

for all significance levels ϵ and data sequences $(z_1, \ldots, z_l, x) \in \mathbf{Z}^l \times \mathbf{X}$, $l = 0, 1, \ldots$. It can be shown that the Q-Bayes confidence predictor is asymptotically conservative and optimal for Q^∞ and any $\epsilon \in (0,1)$; in addition, it satisfies (3.23) for all $\epsilon \in (0,1)$ (it also satisfies (3.22), but this is equivalent to the asymptotic optimality). Non-asymptotic analogs of these properties also hold. (Our definition of the Q-Bayes confidence predictor is arbitrary in several respects: in principle, different choice functions can be used at different trials, the prediction can be arbitrary when $F(f(x)) = \max(\mathbf{M}(\epsilon), \mathbf{E}(\epsilon))$, and $\mathbf{Y}$ can be replaced by any $E \subseteq \mathbf{Y}$ such that $Q(E \mid x) := \sum_{y \in E} Q(y \mid x) = 1$.)

The critical significance level (3.17) is an important characteristic of the probability distribution Q generating the individual examples. If $\epsilon > \epsilon_0$, the Q-Bayes confidence predictor will never output multiple predictions and, since it has to achieve the error rate ϵ, will sometimes have to output empty predictions. If, on the other hand, $\epsilon < \epsilon_0$, there will be multiple predictions but no empty predictions. Figures 3.2 and 3.3 suggest that the critical significance level for the permuted USPS data set is between 2% and 3%. This agrees with the observation that the critical significance level is just the error rate of the Bayes simple predictor (which is restricted to outputting prediction sets Γ_n with $|\Gamma_n| = 1$ and minimizes the expected number of errors) and the already mentioned fact (Vapnik 1998) that the error rate achieved by humans on the USPS data set is 2.5%. Notice that in Fig. 3.4 the onset of empty predictions closely follows the point where multiple predictions disappear; see also Figs. 3.10 and 3.11.

3.5 Proofs

First we establish some simple properties of the predictability distribution function and the multiplicity and emptiness curves.

Lemma 3.6. *The predictability distribution function F satisfies the following properties:*

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

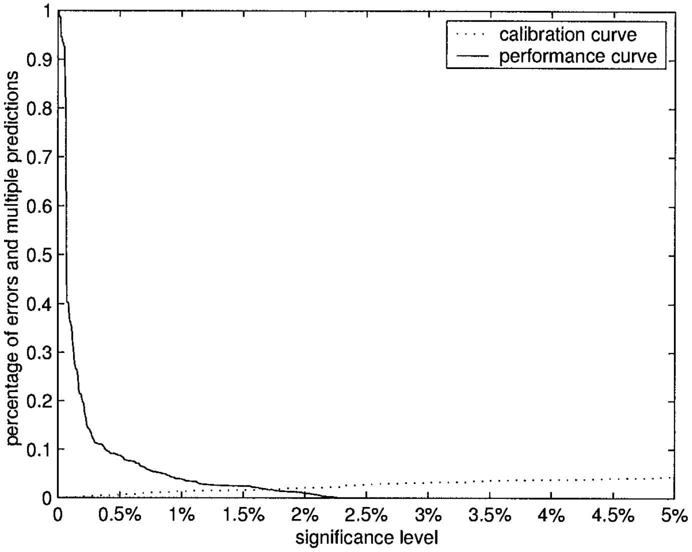

Fig. 3.10. Picture analogous to Fig. 3.6 for the last one thousand examples. Notice a different behavior of the empirical performance curve as it approaches the horizontal axis as compared with Fig. 3.6. The unexpected behavior of the empirical performance curve as it approaches the vertical axis may be explained by the presence of ambiguous and even misclassified examples (LeCun et al. 1990)

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 Q on $\mathbf{X} \times \mathbf{Y}$ for which F is the predictability distribution function.

Proof. Properties 1 (cf. the caption to Fig. 3.7), 2, and 3 are obvious (and the last two are well-known properties of all distribution functions). The fact that these three properties characterize predictability distribution functions easily follows from the fact that the last two properties plus $F(-\infty) = 0$ and $F(\infty) = 1$ characterize distribution functions (see, e.g., Shiryaev 1996, 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 3.7. *The multiplicity curve* $\mathbf{M} : [0,1] \to [0,1]$ *always satisfies these properties:*

1. $\mathbf{M}$ *is convex.*
2. *There is a point* $\epsilon_0 \in [0,1]$ *(the critical significance level) such that* $\mathbf{M}(\epsilon) = 0$ *for* $\epsilon \geq \epsilon_0$ *and* $\mathbf{M}'_{\text{left}}(\epsilon_0) < -1$*; therefore,* $\mathbf{M}'_{\text{left}} < -1$ *and* $\mathbf{M}'_{\text{right}} < -1$ *to the left of* ϵ_0*, and the function* $\mathbf{M}$ *is strictly decreasing before it hits the horizontal axis at* ϵ_0.
3. $\mathbf{M}$ *is continuous at* $\epsilon = 0$*; therefore, it is continuous everywhere in* $[0,1]$.

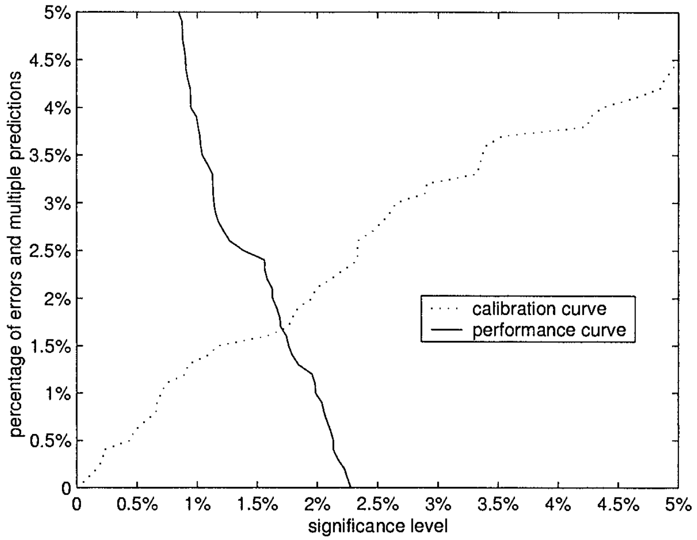

Fig. 3.11. The bottom part of Fig. 3.10 stretched vertically. Notice that the slope of the empirical performance curve is greater than 1 in absolute value before it hits the horizontal axis; this agrees with Lemma 3.7. This figure suggests that, if the 1-nearest neighbor conformal predictor were an optimal confidence predictor, the critical significance level for the permuted USPS data set would be close to 2.3%

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

Proof sketch. For the basic properties of convex functions and their left and right derivatives, see, e.g., Bourbaki 1958 (§I.4). The statement of the lemma follows from the fact that the multiplicity curve $\mathbf{M}$ 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. 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: $F_3(x) := \int_0^x F_2(t)\mathrm{d}t$.
4. Invert F_3: $F_4 := F_3^{-1}$.
5. Flip F_4 around the line $y = 0.5$: $F_5 := 1 - F_4$.

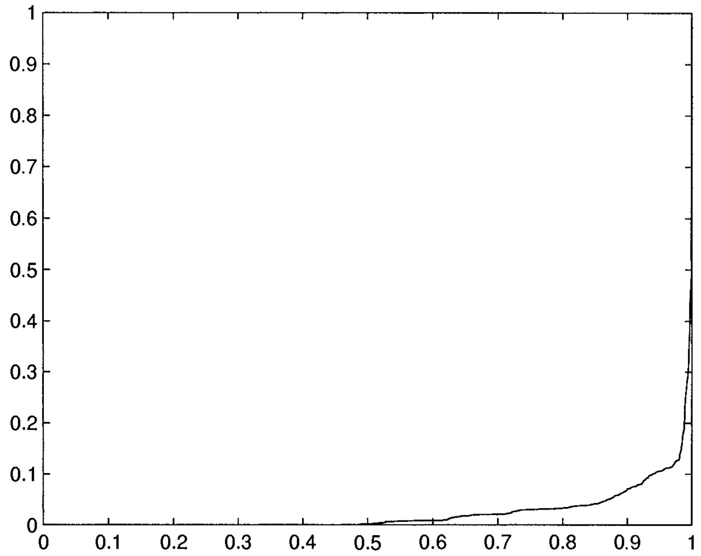

Fig. 3.12. An attempt to reverse engineer the predictability distribution function of the hand-written digits in the USPS data set. This picture was obtained from the solid line in Fig. 3.10 by reversing the list in the proof of Lemma 3.7

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

$$g^{-1}(y) := \sup\{x : g(x) \leq y\} \tag{3.25}$$

for increasing g (so that g^{-1} is continuous on the right). □

Propositions 3.3–3.5 suggest that if the 1-nearest neighbor conformal predictor is close to being optimal on the permuted USPS data set, its empirical performance curve is not far from the multiplicity curve $\mathbf{M}$. Visually the empirical performance curve in Figs. 3.5 and 3.6 seems to satisfy the properties listed in Lemma 3.7 for significance levels that are not too large or too small (approximately in the range 0.1%–5%); for an even better agreement, see Figs. 3.10 and 3.11.

A natural idea is to reverse the process of transforming F into $\mathbf{M}$ and try to obtain an estimate of the predictability distribution function F from an empirical performance curve. Fig. 3.12 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., Vapnik 1998, §1.12).

The following lemma parallels Lemma 3.7:

Lemma 3.8. *The emptiness curve* $\mathbf{E} : [0,1] \to [0,1]$ *always satisfies these properties:*

1. *There is a point* $\epsilon_0 \in [0,1]$ *(namely, the critical significance level) such that* $\mathbf{E}(\epsilon) = 0$ *for* $\epsilon \leq \epsilon_0$ *and* $\mathbf{E}(\epsilon)$ *is concave for* $\epsilon \geq \epsilon_0$.
2. $\mathbf{E}'_{\text{right}}(\epsilon_0) < \infty$ *and* $\mathbf{E}'_{\text{left}}(1) \geq 1$*; therefore, for* $\epsilon \in (\epsilon_0, 1)$, $1 \leq \mathbf{E}'_{\text{right}}(\epsilon) \leq \mathbf{E}'_{\text{left}}(\epsilon) < \infty$ *and the function* $\mathbf{E}(\epsilon)$ *is strictly increasing.*
3. $\mathbf{E}(\epsilon)$ *is continuous at* $\epsilon = \epsilon_0$*; therefore, it is continuous everywhere in* $[0,1]$.

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

Proof sketch. The statement of the lemma follows from the fact that the emptiness curve $\mathbf{E}$ can be obtained from the predictability distribution function F using these steps:

1. Invert F: $F_1 := F^{-1}$.
2. Integrate F_1: $F_2(x) := \int_0^x F_1(t)\mathrm{d}t$.
3. Increase F_2: $F_3(x) := F_2(x) + \epsilon_0$, where $\epsilon_0 := \int_0^1 F(x)\mathrm{d}x$.
4. Invert F_3: $F_4 := F_3^{-1}$.

It can be shown that $\mathbf{E} = F_4$, if we define $g^{-1}(y)$ by (3.25) (with $\sup\emptyset := 0$). □

Proof of Proposition 3.2

Let $z_n = (x_n, y_n)$, $n = 1, 2, \ldots$, be the examples output by Reality and $\tau_1, \tau_2, \ldots$ be the random numbers used by the nearest neighbors smoothed conformal predictor. Let $w_1, w_2, \ldots$ be the sequence of extended examples $w_n := (x_n, \tau'_n, y_n)$. Set

$$Q_n^{\neq}(y \mid x_i, \tau'_i) := \hat{Q}_{\lbag w_1,\ldots,w_{i-1},w_{i+1},w_n \rbag}(y \mid x_i, \tau'_i) \,, \tag{3.26}$$

for all $y \in \mathbf{Y}$ and $i = 1, \ldots, n$. (The upper index $^{\neq}$ reminds us of the fact that (x_i, τ'_i) is not counted as one of its own nearest neighbors in this definition. For the definition of $\hat{Q}$, see (3.14) on p. 63.) We will also use the notation

$$f_n^{\neq}(x_i, \tau'_i) := \max_{y \in \mathbf{Y}} Q_n^{\neq}(y \mid x_i, \tau'_i) = \hat{f}_{\lbag w_1,\ldots,w_{i-1},w_{i+1},\ldots,w_n \rbag}(x_i, \tau'_i)$$

and let $\hat{y}_n(x_i, \tau'_i)$ (without the $^{\neq}$, to make our notation less cumbersome) stand for the first element of $\arg\max_{y \in \mathbf{Y}} Q_n^{\neq}(y \mid x_i, \tau'_i)$ in a fixed ordering of $\mathbf{Y}$.

Without loss of generality we assume that $\{\epsilon_1, \ldots, \epsilon_K\}$ contains only one significance level ϵ, which will be omitted from our notation. We will also assume that all extended objects $(x_i, \tau'_i) \in [0,1]^2$ are different and that all pairwise distances between them are also different (this is true with probability one, since τ_i' are independent random numbers uniformly distributed on $[0,1]$). 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 smoothed conformal predictor:

- a red-black binary *search tree* (see, e.g., Cormen et al. 2001, Chaps. 12–14; the only two operations on red-black trees we need in this book are the query SEARCH and the modifying operation INSERT);
- a growing *array* of nonnegative integers 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 $\{Q_n^{\neq}(y \mid x_i, \tau_i') : y \in \mathbf{Y}\}$ of conditional probability estimates (3.26), 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

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

  with α_i defined by

  $$\begin{aligned}\alpha_i := A\Big(&\Lbag(x_1, \tau_1', y_1), \dots, (x_{i-1}, \tau_{i-1}', y_{i-1}),\\ &(x_{i+1}, \tau_{i+1}', y_{i+1}), \dots, (x_n, \tau_n', y_n)\Rbag, (x_i, \tau_i', y_i)\Big),\\ &\qquad\qquad\qquad\qquad i = 1, \dots, n \ , \qquad (3.27)\end{aligned}$$

  where the nonconformity measure A is defined by (3.15) on p. 63 (with $\hat{f}_B = f_n^{\neq}$ and $\hat{y}_B = \hat{y}_n$).

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'), in the sense of the lexicographic order.

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 estimated conditional probabilities $\{Q_n^{\neq}(y \,|\, x_n, \tau_n') : y \in \mathbf{Y}\}$; this also gives $\hat{y}_n(x_n, \tau_n')$. This involves scanning the vicinity of (x_n, τ_n') for the K_n nearest neighbors of (x_n, τ_n'), which can be done in time $O(K_n)$: the K_n nearest neighbors can be extracted from the vicinity of (x_n, τ_n') sorted in the order of increasing distances from (x_n, τ_n'); since initially the vicinity consists of two sorted lists (to the left and to the right of (x_n, τ_n')), the procedure MERGE used in the merge sort algorithm (see, e.g., Cormen et al. 2001, §2.3.1) will sort the whole vicinity in time $O(K_n)$. Therefore, the required time is $O(K_n) = O(\log n)$.
- For each $y \in \mathbf{Y}$ looks at what happens if the 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 whether $y \in \Gamma_n$. This requires time $O(K_n^2) = O(\log n)$, since there are $O(K_n)$ α_i's in the vicinity of (x_n, τ_n') and each of them can be computed in time $O(K_n)$.
- Outputs the prediction set Γ_n (time $O(1)$).

When the label y_n arrives, the algorithm:

- Inserts the new vertex $(x_n, \tau_n', \tau_n'', y_n, \{Q_n^{\neq}(y \mid x_n, \tau_n') : y \in \mathbf{Y}\})$ 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

  $$\{Q_{n-1}^{\neq}(y \mid x_i, \tau_i') : y \in \mathbf{Y}\} \mapsto \{Q_n^{\neq}(y \mid x_i, \tau_i') : y \in \mathbf{Y}\}$$

  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 the other (x_i, τ_i'), $i = 1, \dots, n-1$, do not change.
- Updates the array, changing $N(K_n \alpha_i)$ for the $(x_i, \tau_i') \neq (x_n, \tau_n')$ in the vicinity of (x_n, 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 the updates required when K_n changes. At the critical trials n when K_n changes the array and all estimated conditional probabilities $Q_n^{\neq}(y|x_i, \tau_i')$ have to be recomputed, which, if done naively, would require time $\Theta(nK_n)$.

The assumption we have made about K_n so far is that $K_n = O(\sqrt{\log n})$. Now we also assume that K_n is strictly increasing and

$$|\{n : K_n < c\}| = O\left(|\{n : K_n = c\}|\right) \qquad (3.28)$$

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

An *epoch* is defined to be a maximal sequence of 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 trial in an epoch we prepare the ground for the next epoch. We will only discuss updating the estimated conditional probabilities $Q^{\neq}_n(y \mid x_i, \tau'_i)$; the array is treated in a similar way.

By the end of epoch $n = A+1, A+2, \dots, B$ we need to change B sets $\{Q^{\neq}_n(y \mid x_i, \tau'_i) : y \in \mathbf{Y}\}$ in $B - A$ trials (the duration of the epoch). Therefore, each vertex of the search tree should contain not only $\{Q^{\neq}_n(y \mid x_i, \tau'_i)\}$ for the current epoch but also $\{Q^{\neq}_n(y \mid x_i, \tau'_i)\}$ for the next epoch (two structures for holding $\{Q^{\neq}_n(y \mid x_i, \tau'_i)\}$ will suffice, one for even epochs and one for odd epochs). Our assumptions of the slow growth of K_n (see (3.28)) imply that $B = O(B - A)$. This means that at each trial $O(1)$ sets $\{Q^{\neq}_n(y \mid x_i, \tau'_i)\}$ for the next epoch should be added. This will take time $O(K_n) = O(\log n)$. As soon as a set $\{Q^{\neq}_n(y \mid x_i, \tau'_i) : y \in \mathbf{Y}\}$ for the next epoch is added at some trial, both sets (for the current and next epoch) will have to be updated for each new example.

Proof of Proposition 3.3

Let us check first that (3.20) indeed implies $\mathbf{P}(\epsilon) \ge \mathbf{M}(\epsilon)$ (we will omit the lower indices Γ, Q). Since probability distributions are σ-additive, (3.19) implies

$$\limsup_{n\to\infty} \frac{\mathrm{Mult}^\epsilon_n(\Gamma, Q^\infty)}{n} \le \mathbf{P}(\epsilon) \quad \text{a.s.},$$

and so we obtain from (3.20):

$$\mathbf{P}(\epsilon) \ge \limsup_{n\to\infty} \frac{\mathrm{Mult}^\epsilon_n(\Gamma, Q^\infty)}{n} \ge \liminf_{n\to\infty} \frac{\mathrm{Mult}^\epsilon_n(\Gamma, Q^\infty)}{n} \ge \mathbf{M}(\epsilon)$$

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

We start the actual proof with alternative definitions of calibration and performance curves. Complement the basic protocol given at the beginning of this chapter (p. 53) in which Reality plays Q^∞ and Predictor plays Γ with the following variables:

$$\overline{\mathrm{err}}^\epsilon_n := (Q \times \mathbf{U})\Big\{(x, y, \tau) \in (\mathbf{Z} \times [0,1]) : \\ y \notin \Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)\Big\}, \tag{3.29}$$

$$\overline{\mathrm{mult}}^\epsilon_n := (Q_{\mathbf{X}} \times \mathbf{U})\Big\{(x, \tau) \in (\mathbf{X} \times [0,1]) : \\ |\Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)| > 1\Big\}, \tag{3.30}$$

$$\overline{\mathrm{Err}}^\epsilon_n := \sum_{i=1}^n \overline{\mathrm{err}}^\epsilon_i, \quad \overline{\mathrm{Mult}}^\epsilon_n := \sum_{i=1}^n \overline{\mathrm{mult}}^\epsilon_i. \tag{3.31}$$

The *predictable calibration curve* of Γ under Q is defined by

$$\overline{\mathbf{C}}(\epsilon) := \inf\left\{\beta : \mathbb{P}\left\{\limsup_{n\to\infty} \frac{\overline{\mathrm{Err}}^\epsilon_n}{n} \le \beta\right\} = 1\right\}$$

and the *predictable performance curve* of Γ under Q by

$$\overline{\mathbf{P}}(\epsilon) := \inf\left\{\beta : \mathbb{P}\left\{\limsup_{n\to\infty} \frac{\overline{\mathrm{Mult}}^\epsilon_n}{n} \le \beta\right\} = 1\right\},$$

where $\mathbb{P}$ refers to the probability distribution $(Q \times \mathbf{U})^\infty$ over the examples $z_1, z_2, \dots$ and random numbers $\tau_1, \tau_2, \dots$. By the martingale strong law of large numbers (see §A.6) the predictable versions of the calibration and performance curves coincide with the original versions: indeed, since $\mathrm{Err}^\epsilon_n - \overline{\mathrm{Err}}^\epsilon_n$ and $\mathrm{Mult}^\epsilon_n - \overline{\mathrm{Mult}}^\epsilon_n$ are martingales (with increments bounded by 1 in absolute value) for all ϵ with respect to the filtration $\mathcal{F}_n$, $n = 0, 1, \dots$, where each $\mathcal{F}_n$ is generated by $z_1, \dots, z_n$ and $\tau_1, \dots, \tau_n$, we have

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

and

$$\lim_{n\to\infty} \frac{\mathrm{Mult}^\epsilon_n - \overline{\mathrm{Mult}}^\epsilon_n}{n} = 0 \quad \mathbb{P}\text{-a.s.}$$

It is also clear that we can replace Mult^ϵ_n by $\overline{\mathrm{Mult}}^\epsilon_n$ in (3.20).

Without loss of generality we can assume that Predictor's move Γ^ϵ_n at trial n is, for each ϵ, either $\{\hat{y}(x_n)\}$ ($\hat{y}$ is defined by (3.24), p. 70) or vacuous, the whole label space $\mathbf{Y}$. Furthermore, we can assume that

$$\overline{\mathrm{mult}}^\epsilon_n = \mathbf{M}(\overline{\mathrm{err}}^\epsilon_n) \tag{3.32}$$

at every trial, since the best way to spend the allowance of $\overline{\mathrm{err}}^\epsilon_n$ is to give non-vacuous predictions for objects x with the largest (topmost in Figs. 3.7 and 3.8) representations $F(f(x))$. (For a formal argument, see the end of this proof.) Using the fact that the multiplicity curve $\mathbf{M}$ is convex, decreasing, and continuous (see Lemma 3.7), we obtain, for any significance level ϵ,

$$\frac{\overline{\mathrm{Mult}}^\epsilon_n}{n} = \frac{1}{n}\sum_{i=1}^n \overline{\mathrm{mult}}^\epsilon_i = \frac{1}{n}\sum_{i=1}^n \mathbf{M}(\overline{\mathrm{err}}^\epsilon_i) \\ \ge \mathbf{M}\left(\frac{1}{n}\sum_{i=1}^n \overline{\mathrm{err}}^\epsilon_i\right) = \mathbf{M}\left(\frac{\overline{\mathrm{Err}}^\epsilon_n}{n}\right) \ge \mathbf{M}(\epsilon) - \delta, \tag{3.33}$$

the last inequality holding almost surely for an arbitrary $\delta > 0$ from some n on.

It remains to prove formally that $\overline{\mathrm{mult}}^\epsilon_n \geq \mathbf{M}(\overline{\mathrm{err}}^\epsilon_n)$ (which is the part of (3.32) that we actually used). Let us fix ϵ and

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

we will write

$$\Gamma(x, \tau) := \Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau) \ ,$$

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

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

Our goal is to show that $\overline{\mathrm{mult}}_n \geq \mathbf{M}(\delta)$; without loss of generality we assume $\delta < \epsilon_0$, where ϵ_0 is the critical significance level. To put it differently, we are required to show that the value of the optimization problem

$$\int_\mathbf{X} p(x) Q(\mathrm{d}x) \to \max \tag{3.34}$$

subject to the constraint

$$\int_\mathbf{X} (1 - f(x)) p(x) Q(\mathrm{d}x) = \delta$$

is $1 - \mathbf{M}(\delta)$ at best (remember that $f(x)$ is the predictability of x; Q is a shorthand for $Q_\mathbf{X}$). By the Neyman–Pearson lemma (see, e.g., Lehmann 1986, Theorem 3.2.1) for some solution p there exist constants $c > 0$ and $d \in [0, 1]$ such that

$$p(x) = \begin{cases} 1 & \text{if } f(x) > c \\ d & \text{if } f(x) = c \\ 0 & \text{if } f(x) < c \ . \end{cases} \tag{3.35}$$

The constants c and d are defined (c uniquely and d uniquely unless the probability of $f(x) = c$ is zero or $c = 1$; in the latter case, $d = 1$) from the condition

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

which is equivalent (see Fig. 3.7 on p. 66) to

$$\int_0^1 (F(\beta) - F(c))^+ \mathrm{d}\beta + d(1 - c)(F(c) - F(c-)) = \delta \ , \tag{3.36}$$

where $F(c-)$ is defined as $\lim_{\beta \uparrow c} F(\beta)$. From this it is easy to obtain that the value of the optimization problem (3.34) is indeed $1 - \mathbf{M}(\delta)$: using the notation $p_d(x)$ for the right-hand side of (3.35), we have

$$\begin{aligned} \int_\mathbf{X} p_d(x) Q(\mathrm{d}x) &= d \int p_1(x) Q(\mathrm{d}x) + (1 - d) \int p_0(x) Q(\mathrm{d}x) \\ &= d Q\{x : f(x) \geq c\} + (1 - d) Q\{x : f(x) > c\} \\ &= d(1 - F(c-)) + (1 - d)(1 - F(c)) \\ &= 1 - F(c) + d(F(c) - F(c-)) = 1 - \mathbf{M}(\delta) \ , \end{aligned}$$

the last equality following from (3.36) (except for the case $c = 1$, when it is obvious). This completes the proof.

Proof sketch of Proposition 3.4

The proof of Proposition 3.4 is similar to (but more complicated than) the proof of Proposition 3.3; this sketch can be made rigorous using the Neyman–Pearson lemma, as we did in the proof of Proposition 3.3.

Along with the random variables (3.29)–(3.31) we will also need

$$\begin{aligned} \overline{\mathrm{emp}}^\epsilon_n := (Q_\mathbf{X} \times \mathbf{U})\bigl\{(x, \tau) \in (\mathbf{X} \times [0, 1]) : \\ |\Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)| = 0\bigr\} \end{aligned} \tag{3.37}$$

and

$$\overline{\mathrm{Emp}}^\epsilon_n := \sum_{i=1}^n \overline{\mathrm{emp}}^\epsilon_i \ . \tag{3.38}$$

It is clear that

$$\lim_{n \to \infty} \frac{\mathrm{Emp}^\epsilon_n - \overline{\mathrm{Emp}}^\epsilon_n}{n} = 0 \quad \text{a.s.}$$

Without loss of generality we can assume that Predictor's move Γ_n at trial n is $\{\hat{y}(x_n)\}$ or the empty set $\emptyset$ or the whole label space $\mathbf{Y}$. Furthermore, we can assume that, at every trial, the predictions are singular (i.e., contain one label) for the new objects above the straight line BC in Fig. 3.13 (more formally, for new extended objects (x, τ) satisfying

$$F(x, \tau) := F(f(x)-) + \tau\,(F(f(x)+) - F(f(x)-)) \geq \mathbf{M}(\overline{\mathrm{err}}^\epsilon_n - \overline{\mathrm{emp}}^\epsilon_n) \ ;$$

intuitively, considering extended objects makes the vertical axis "infinitely divisible") and that the predictions are empty for the objects below the straight line DG in Fig. 3.13. (Indeed, predictions of this kind are admissible in the sense that we cannot improve $\overline{\mathrm{mult}}^\epsilon_n$ and $\overline{\mathrm{emp}}^\epsilon_n$ simultaneously, and all admissible predictions are equivalent to predictions of this kind. A formal argument for the case where emp^ϵ_n are omitted is given in the proof of Proposition 3.3 above.) It is clear that for the confidence predictor to satisfy (3.21) it must hold that

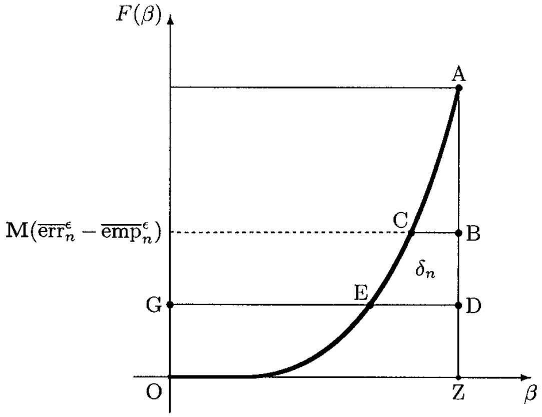
$F(\beta)$
A
$\mathbf{M}(\overline{\mathrm{err}}_n^\epsilon - \overline{\mathrm{emp}}_n^\epsilon)$
C
B
δ_n
E
G
D
O
Z
β

Fig. 3.13. An admissible confidence predictor. The thick line is the predictability distribution function F; the area of the curvilinear triangle ABC is $\overline{\mathrm{err}}_n^\epsilon - \overline{\mathrm{emp}}_n^\epsilon$; the area of the rectangle DZOG is $\overline{\mathrm{emp}}_n^\epsilon$; the (nonnegative) area of the curvilinear quadrangle BDEC is denoted δ_n

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

(otherwise $\overline{\mathrm{Mult}}_n^\epsilon$ can be decreased substantially, which contradicts (3.20); δ_i are defined in the caption of Fig. 3.13), and so we can assume, without loss of generality, that either $\delta_n = 0$ or $\overline{\mathrm{emp}}_n^\epsilon = 0$ at every trial n, i.e., that

$$\overline{\mathrm{mult}}_n^\epsilon = \mathbf{M}(\overline{\mathrm{err}}_n^\epsilon), \quad \overline{\mathrm{emp}}_n^\epsilon = \mathbf{E}(\overline{\mathrm{err}}_n^\epsilon)$$

at every trial. In the sequel we will omit the upper index ϵ.

Let us check that to achieve (3.21) the randomized confidence predictor must satisfy

$$\epsilon < \epsilon_0 \Longrightarrow \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \epsilon_0)^+ = 0 \tag{3.39}$$

$$\epsilon \ge \epsilon_0 \Longrightarrow \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\epsilon_0 - \overline{\mathrm{err}}_i)^+ = 0 \,, \tag{3.40}$$

where the convergence is, as usual, almost certain. We know from Lemma 3.7 that the multiplicity curve $\mathbf{M}$ is convex, decreasing, continuous, and has slope at most -1 before it hits the horizontal axis at $\epsilon = \epsilon_0$. The second implication, (3.40), now immediately follows from the fact that, under $\epsilon \ge \epsilon_0$ and (3.21),

$$0 = \limsup_{n\to\infty} \frac{\overline{\mathrm{Mult}}_n}{n} = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{M}(\overline{\mathrm{err}}_i) \ge \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\epsilon_0 - \overline{\mathrm{err}}_i)^+ \,.$$

The first implication, (3.39), can be extracted from the chain (3.33) in the proof of Proposition 3.3. Indeed, it can be seen from (3.33) that, assuming the predictor satisfies (3.21) and $\epsilon < \epsilon_0$,

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

and, therefore,

$$\begin{aligned}
\mathbf{M}(\epsilon) &\ge \limsup_{n\to\infty} \frac{\overline{\mathrm{Mult}}_n}{n} = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{M}(\overline{\mathrm{err}}_i) \\
&= \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{M}(\overline{\mathrm{err}}_i \wedge \epsilon_0) \ge \limsup_{n\to\infty} \mathbf{M}\left(\frac{1}{n} \sum_{i=1}^{n} (\overline{\mathrm{err}}_i \wedge \epsilon_0) \right) \\
&= \limsup_{n\to\infty} \mathbf{M}\left(\frac{\overline{\mathrm{Err}}_n}{n} - \frac{1}{n} \sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \epsilon_0)^+ \right) \\
&= \limsup_{n\to\infty} \mathbf{M}\left(\epsilon - \frac{1}{n} \sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \epsilon_0)^+ \right) \\
&= \mathbf{M}\left(\epsilon - \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \epsilon_0)^+ \right)
\end{aligned}$$

almost surely. This proves (3.39).

Using (3.39), (3.40), and the fact that the emptiness curve $\mathbf{E}$ is concave, increasing, and (uniformly) continuous for $\epsilon \ge \epsilon_0$ (see Lemma 3.8), we obtain: if $\epsilon < \epsilon_0$,

$$\begin{aligned}
\frac{\overline{\mathrm{Emp}}_n}{n} &= \frac{1}{n} \sum_{i=1}^{n} \overline{\mathrm{emp}}_i = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}(\overline{\mathrm{err}}_i) \\
&\le \frac{1}{n} \mathbf{E}'_{\mathrm{right}}(\epsilon_0) \sum_{i=1}^{n} (\overline{\mathrm{err}}_i - \epsilon_0)^+ \to 0 \quad (n \to \infty) \,;
\end{aligned}$$

if $\epsilon \ge \epsilon_0$,

$$\begin{aligned}
\frac{\overline{\mathrm{Emp}}_n}{n} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}(\overline{\mathrm{err}}_i) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}(\overline{\mathrm{err}}_i \vee \epsilon_0) \\
&\le \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^{n} (\overline{\mathrm{err}}_i \vee \epsilon_0) \right) = \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^{n} \overline{\mathrm{err}}_i + \frac{1}{n} \sum_{i=1}^{n} (\epsilon_0 - \overline{\mathrm{err}}_i)^+ \right) \\
&= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^{n} \overline{\mathrm{err}}_i \right) + o(1) \le \mathbf{E}(\epsilon) + \delta \,,
\end{aligned}$$

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

Proof sketch of Proposition 3.5

It will be convenient to consider a modification and extension of the function $Q_n^{\neq}(y \mid x_i, \tau_i')$ introduced in (3.26). An alternative definition of the nearest neighbors approximations $Q_n(y \,|\, x, \sigma)$ to the conditional probabilities $Q(y \,|\, x)$ is as follows: for every $(x, \sigma, y) \in \tilde{\mathbf{Z}}$,

$$Q_n(y \mid x, \sigma) := \hat{Q}_{\lbag w_1, \ldots, w_n \rbag}(y \mid x, \sigma) .$$

(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

$$f_n(x, \sigma) := \max_{y \in \mathbf{Y}} Q_n(y \,|\, x, \sigma) = \hat{f}_{\lbag w_1, \ldots, w_n \rbag}(x, \sigma) .$$

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

Lemma 3.9. *Suppose $K_n \to \infty$, $K_n = o(n)$, and $\mathbf{Y} = \{0, 1\}$. For any $\delta > 0$ and large enough n,*

$$\mathbb{P}\left\{\int |Q(1 \,|\, x) - Q_n(1 \,|\, x, \sigma)| \, Q_{\mathbf{X}}(\mathrm{d}x)\mathbf{U}(\mathrm{d}\sigma) > \delta\right\} \le \mathrm{e}^{-n\delta^2/40} ,$$

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

Proof. This is almost a special case of Devroye et al.'s (1994) Theorem 1. There is, however, an important difference between the way we break distance ties and the way Devroye et al. (1994) do this: in Devroye et al. 1994, instead of our (3.10),

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

is used. (Our way of breaking ties better agrees with the lexicographic order on $[0, 1]^2$, which is useful in the proof of Proposition 3.2 and, less importantly, in the proof of Lemma 3.11.) It is easy to check that the proof given in Devroye et al. 1994 also works (and becomes simpler) for our way of breaking distance ties. □

Lemma 3.10. *Suppose $K_n \to \infty$ and $K_n = o(n)$. For any $\delta > 0$ there exists a $\delta^* > 0$ such that, for large enough n,*

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

in particular,

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

Proof. We apply Lemma 3.9 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:

$$\mathbb{P}\left\{\int |Q(y \,|\, x) - Q_n(y \,|\, x, \sigma)| \, Q_{\mathbf{X}}(\mathrm{d}x)\mathbf{U}(\mathrm{d}\sigma) > \delta\right\} \le \mathrm{e}^{-n\delta^2/40} .$$

By Markov's inequality this implies

$$\mathbb{P}\left\{(Q_{\mathbf{X}} \times \mathbf{U})\{|Q(y \,|\, x) - Q_n(y \,|\, x, \sigma)| > \sqrt{\delta}\} > \sqrt{\delta}\right\} \le \mathrm{e}^{-n\delta^2/40} ,$$

which, in turn, implies

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

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

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

Lemma 3.11. *Suppose $K_n \to \infty$ and $K_n = o(n)$. For any $\delta > 0$ there exists a $\delta^* > 0$ such that, for large enough n,*

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

In particular,

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

Proof. Since

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

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

Define

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

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

The main tool in this proof (and several other proofs in this section) will be McDiarmid's theorem (see §A.7). 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

$$4K_n\frac{1}{\delta}\frac{1}{K_n}=\frac{4}{\delta}\,. \tag{3.41}$$

The left-hand side of (3.41) reflects the following facts: the change in $Q_n(y \mid x_i,\tau_i')$ 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 (3.10) 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,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} I_n(x_i,\tau_i') - \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} I_n(x_i,\tau_i')\right) > \delta\right\}$$
$$\le \exp\left(-2\delta^2 n/\left(1+4/\delta\right)^2\right) = \exp\left(-\frac{2\delta^4}{(4+\delta)^2}n\right)\,. \tag{3.42}$$

Next we find:

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

(the penultimate inequality follows from Lemma 3.10) from some n on. In combination with (3.42) this implies

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

in particular

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

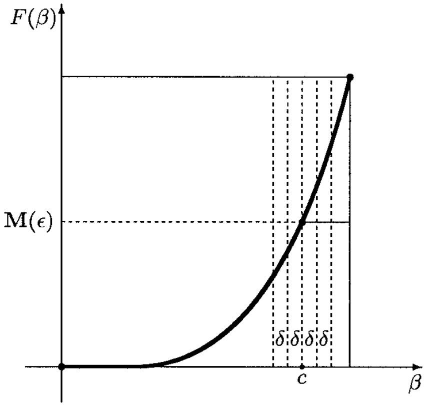

Fig. 3.14. Case $F(c)=\mathbf{M}(\epsilon)$

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

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

which completes the proof. □

We say that an extended example (x_i,τ_i,y_i), $i=1,\dots,n$, is *n-strange* if $y_i \neq \hat{y}_n(x_i,\tau_i')$; otherwise, (x_i,τ_i,y_i) will be called *n-conforming*. We will assume that $(f_n^{\neq}(x_i,\tau_i'),1-\tau_i'')$, $i=1,\dots,n$, are all different for all n; even more than that, we will assume that τ_i'', $i=1,2,\dots$, are all different (we can do so since the probability of this event is one).

Lemma 3.12. *Suppose (3.13) (p. 63) is satisfied and* $\epsilon \le \epsilon_0$. *With probability one, the* $\lfloor(1-\mathbf{M}(\epsilon))n\rfloor$ *extended examples with the largest (in the sense of the lexicographic order)* $(f_n^{\neq}(x_i,\tau_i'),1-\tau_i'')$ *among* $(x_1,\tau_1,y_1),\dots,(x_n,\tau_n,y_n)$ *contain at most* $n\epsilon+o(n)$ *n-strange extended examples as* $n\to\infty$.

Proof. Define

$$c:=\sup\{\beta:F(\beta)\le\mathbf{M}(\epsilon)\}\,.$$

It is clear that $0<c<1$. Our proof will work both in the case where $F(c)=\mathbf{M}(\epsilon)$ and in the case where $F(c)>\mathbf{M}(\epsilon)$, as illustrated in Figs. 3.14 and 3.15.

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

$$\mathbb{P}\left\{f(x_n)=c,(f_{n-1}(x_n,\tau_n'),1-\tau_n'')>(c_n',c_n'')\right\}=F(c)-\mathbf{M}(\epsilon)-\delta \tag{3.43}$$

if $F(c)>\mathbf{M}(\epsilon)$; we assume that δ is small enough for

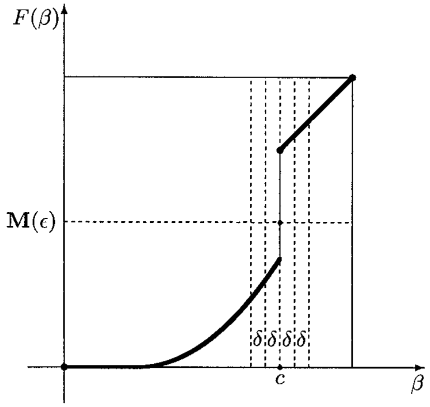

Fig. 3.15. Case $F(c) > \mathbf{M}(\epsilon)$

$$2\delta < F(c) - \mathbf{M}(\epsilon) \tag{3.44}$$

to hold (among other things this will ensure the validity of the definition (3.43)). If $F(c) = \mathbf{M}(\epsilon)$, we set $(c'_n, c''_n) := (c + \delta, 0)$; in any case, we will have

$$\mathbb{P}\{f(x_n) = c, (f_{n-1}(x_n, \tau'_n), 1 - \tau''_n) > (c'_n, c''_n)\} \geq F(c) - \mathbf{M}(\epsilon) - \delta\ . \tag{3.45}$$

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

$$(f^{\neq}_n(x_i, \tau'_i), 1 - \tau''_i) > (c'_n, c''_n)\ ;$$

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 $f(x_i) \leq c - 2\delta$.
Class II: Those that satisfy $f(x_i) = c$ and are below the threshold.
Class III: Those satisfying $c - 2\delta < f(x_i) \leq c + 2\delta$ but not $f(x_i) = c$.
Class IV: Those that satisfy $f(x_i) = c$ and are above the threshold.
Class V: Those satisfying $f(x_i) > c + 2\delta$.

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

- How many extended examples are there in the class?

- How many of those are above the threshold?
- How many of those above the threshold are n-strange?

If the sum of the answers to the last question does not exceed $n\epsilon$ 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

$$c - \delta \leq c'_n \leq c + \delta \tag{3.46}$$

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

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

(combine Lemma 3.10 with (3.44)). The inequality $c - \delta \leq c'_n$ follows from

$$\begin{aligned}\forall^{\infty} n : &\mathbb{P}\{f(x_n) = c, f_{n-1}(x_n, \tau'_n) \geq c - \delta\}\\ &= \mathbb{P}\{f(x_n) = c\} - \mathbb{P}\{f(x_n) = c, f_{n-1}(x_n, \tau'_n) < c - \delta\}\\ &> F(c) - F(c-) - \delta \geq F(c) - \mathbf{M}(\epsilon) - \delta\ .\end{aligned}$$

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

$$\delta n \tag{3.47}$$

will be above the threshold from some n on almost surely (by Lemma 3.11 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

$$F(c + 2\delta) - F(c) + F(c-) - F(c - 2\delta) \tag{3.48}$$

as $n \to \infty$ almost surely.

To estimate the number N^{IV}_n 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^{IV}_n will change by at most $2K_n + 1$ (since this extended example can affect $f^{\neq}_n(x_i, \tau'_i)$ for at most $2K_n$ other extended examples (x_i, τ_i, y_i)). Therefore,

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

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

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

from some n on almost surely. Since

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

(see (3.45)), we have

$$N_n^{\mathrm{IV}} > (F(c) - \mathbf{M}(\epsilon) - 2\delta) n \tag{3.49}$$

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

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

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

$$\begin{aligned}
\frac{1}{n} \mathbb{E} N_n^{\mathrm{IV,str}} &= \mathbb{P}\left\{ f(x_n) = c, (f_{n-1}(x_n, \tau_n'), 1 - \tau_n'') > (c_n', c_n''), \hat{y}_n(x_n, \tau_n') \neq y_n \right\} \\
&= \mathbb{E}\left((1 - Q\left(\hat{y}_n(x_n, \tau_n') \mid x_n \right)) \mathbb{I}_{\{f(x_n) = c, (f_{n-1}(x_n, \tau_n'), 1 - \tau_n'') > (c_n', c_n'')\}} \right) \\
&\leq \mathrm{e}^{-\delta^*(n-1)} + \delta + (1 - c + 2\delta) \\
&\quad \times \mathbb{P}\{ f(x_n) = c, (f_{n-1}(x_n, \tau_n'), 1 - \tau_n'') > (c_n', c_n'') \} \\
&= \mathrm{e}^{-\delta^*(n-1)} + \delta + (1 - c + 2\delta)(F(c) - \mathbf{M}(\epsilon) - \delta) && (3.50) \\
&\leq (F(c) - \mathbf{M}(\epsilon))(1 - c) + 4\delta && (3.51)
\end{aligned}$$

in the case $F(c) > \mathbf{M}(\epsilon)$; the first inequality in this chain follows from Lemma 3.10: indeed, this lemma implies that, unless an event of the small probability $\mathrm{e}^{-\delta^*(n-1)} + \delta$ happens,

$$\begin{aligned}
Q\left(\hat{y}_n(x_n, \tau_n') \mid x_n \right) &\geq Q_{n-1}\left(\hat{y}_n(x_n, \tau_n') \mid x_n, \tau_n' \right) - \delta \\
&= f_{n-1}\left(x_n, \tau_n' \right) - \delta \geq f(x_n) - 2\delta \,. && (3.52)
\end{aligned}$$

If $F(c) = \mathbf{M}(\epsilon)$, the lines (3.50) and (3.51) of that chain have to be changed to

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

(where the obvious modification of Lemma 3.10 with all "$> \delta$" changed to "$\geq \delta$" is used), but the inequality between the extreme terms of the chain still holds. Therefore, the number of n-strange Class IV extended examples does not exceed

$$((F(c) - \mathbf{M}(\epsilon))(1 - c) + 5\delta)\, n \tag{3.53}$$

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

$$1 - F(c + 2\delta) \tag{3.54}$$

as $n \to \infty$ almost surely. By Lemma 3.11, the Borel–Cantelli lemma, and (3.46), almost surely from some n on at least

$$(1 - F(c + 2\delta) - 2\delta) n \tag{3.55}$$

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,

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

from some n on almost surely. Now

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

from some n on (the first inequality follows from Lemma 3.10, as in (3.52)). Therefore,

$$\frac{1}{n} N_n^{\mathrm{V,str}} < \int_0^1 (F(\beta) - F(c))^+ \mathrm{d}\beta + 5\delta \tag{3.56}$$

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

$$\begin{aligned}
&(F(c) - \mathbf{M}(\epsilon) - 2\delta + 1 - F(c + 2\delta) - 2\delta)\, n \\
&\qquad = (1 - \mathbf{M}(\epsilon) + F(c) - F(c + 2\delta) - 4\delta)\, n && (3.57)
\end{aligned}$$

(see (3.49) and (3.55)) from some n on almost surely. The number of n-strange extended examples among them will not exceed

$$\begin{aligned}&\Bigg(\delta + F(c+2\delta) - F(c) + F(c-) - F(c-2\delta) + \delta \\ &\qquad + (F(c) - \mathbf{M}(\epsilon))(1-c) + 5\delta + \int_0^1 (F(\beta) - F(c))^+ \mathrm{d}\beta + 5\delta\Bigg) n \\ &\qquad\qquad = \Bigg(F(c+2\delta) - F(c) + F(c-) - F(c-2\delta) \\ &\qquad\quad + (F(c) - \mathbf{M}(\epsilon))(1-c) + \int_0^1 (F(\beta) - F(c))^+ \mathrm{d}\beta + 12\delta\Bigg) n \qquad (3.58)\end{aligned}$$

(see (3.47), (3.48), (3.53), and (3.56)) from some n on almost surely. Combining equations (3.57) and (3.58), we can see that the number of n-strange extended examples among the $\lfloor(1 - \mathbf{M}(\epsilon))n\rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), 1 - \tau_i'')$ does not exceed

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

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

$$(F(c) - \mathbf{M}(\epsilon))(1-c) + \int_0^1 (F(\beta) - F(c))^+ \mathrm{d}\beta = \int_0^1 (F(\beta) - \mathbf{M}(\epsilon))^+ \mathrm{d}\beta = \epsilon\,,$$

which completes the proof. □

Lemma 3.13. *Suppose (3.13) is satisfied. The fraction of n-strange extended examples among the first n extended examples (x_i, τ_i, y_i) approaches ϵ_0 asymptotically with probability one.*

Proof sketch. The lemma is not difficult to prove using McDiarmid's theorem and the fact that, by Lemma 3.11, $Q(\hat{y}_n(x_i, \tau_i') \mid x_i)$ will typically differ little from $f(x_i)$. Notice, however, that the part that we really need (that the fraction of n-strange extended examples does not exceed $\epsilon_0 + o(1)$ as $n \to \infty$ with probability one) is just a special case of Lemma 3.12, corresponding to $\epsilon = \epsilon_0$. □

Lemma 3.14. *Suppose that (3.13) is satisfied and $\epsilon > \epsilon_0$. The fraction of n-conforming extended examples among the $\lfloor \mathbf{E}(\epsilon) n \rfloor$ extended examples (x_i, τ_i, y_i), $i = 1, \dots, n$, with the lowest $(f_n^{\neq}(x_i, \tau_i'), 1 - \tau_i'')$ does not exceed $\epsilon - \epsilon_0 + o(1)$ as $n \to \infty$ with probability one.*

Lemma 3.14 can be proved analogously to Lemma 3.12.

Lemma 3.15. *Let $\mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \cdots$ be a decreasing sequence of σ-algebras and $\xi_1, \xi_2 \dots$ be a bounded adapted (in the sense that 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 $\xi_1, \xi_2, \dots$ is a bounded *reverse martingale difference*, in the sense of being adapted and satisfying $\forall n : \mathbb{E}(\xi_n \mid \mathcal{F}_{n+1}) = 0$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i = 0 \quad \text{a.s.} \tag{3.59}$$

Fix a bounded reverse martingale difference $\xi_1, \xi_2, \dots$; our goal is to prove (3.59). By (the martingale version of) Hoeffding's inequality (see §A.7) applied to the martingale difference $(\xi_i, \mathcal{F}_i)$, $i = n, \dots, 1$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i\right| \ge \delta\right\} \le 2\mathrm{e}^{-\delta^2 n/(2C^2)}\,, \tag{3.60}$$

where C is an upper bound on $\sup_n |\xi_n|$. Combined with the Borel–Cantelli–Lévy lemma, (3.60) implies (3.59). □

Now we can sketch the proof of Proposition 3.5. Define $\mathcal{F}_n$, $n = 1, 2, \dots$, 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), $i = 1, \dots, n-1$, and the sequence of extended examples (x_i, τ_i, y_i), $i = n, n+1, \dots$ (starting from the nth extended example).

Suppose first that $\epsilon < \epsilon_0$. Consider the $\lfloor(1 - \mathbf{M}(\epsilon - \delta))n\rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), 1 - \tau_i'')$ among $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$, where $\delta \in (0, \epsilon)$ is a small constant. Let us show that each of these examples will be predicted with a non-multiple prediction from the other extended examples in the sequence $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$, from some n on. We will assume n large enough.

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

$$\left|\{i = 1, \dots, n : \alpha_i^y > \alpha_j^y\}\right| + \tau_j'' \left|\{i = 1, \dots, n : \alpha_i^y = \alpha_j^y\}\right| \le n\epsilon$$

(cf. (3.12) on p. 62), where all α^y are computed as α in (3.27) (p. 75) from the sequence $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$ with y_j replaced by y. 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 $2K_n + 1$ nonconformity scores), it suffices to prove

$$\left|\{i : \alpha_i > f_n^{\neq}(x_j, \tau_j')\}\right| + \tau_j'' \left|\{i : \alpha_i = f_n^{\neq}(x_j, \tau_j')\}\right| \le n(\epsilon - \delta^*) , \qquad (3.61)$$

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

Since $(f_n^{\neq}(x_j, \tau_j'), 1 - \tau_j'') \ge (\alpha_k, 1 - \tau_k'')$ (indeed, by Lemma 3.12, there are less than $(\epsilon - \delta/2)n$ n-strange extended examples among the $\lfloor(1 - \mathbf{M}(\epsilon - \delta))n\rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau_i'), 1 - \tau_i'')$), (3.61) will follow from

$$\left|\{i : \alpha_i > \alpha_k\}\right| + \tau_k'' \left|\{i : \alpha_i = \alpha_k\}\right| \le n(\epsilon - \delta^*) . \qquad (3.62)$$

If $\left|\{i : \alpha_i = \alpha_k\}\right| \le \frac{\delta}{3}n$, the left-hand side of (3.62) does not exceed

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

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

$$\left|\{i : \alpha_i = \alpha_k\}\right| > \frac{\delta}{3} n . \qquad (3.63)$$

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 τ_k'' is about

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

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

$$\tau_k'' \le \frac{(\epsilon - \delta/2)n - \left|\{i : \alpha_i > \alpha_k\}\right|}{\left|\{i : \alpha_i = \alpha_k\}\right|} + \delta^* \qquad (3.64)$$

(remember (3.63)). Equation (3.62) will hold because its left-hand side can be transformed using (3.64) as

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

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

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

Since δ can be arbitrarily small and $\mathbf{M}$ is continuous (Lemma 3.7), this implies

$$\limsup_{n\to\infty} \mathbb{E}(\mathrm{mult}_n \,|\mathcal{F}_{n+1}) \le \mathbf{M}(\epsilon) \quad \text{a.s.}$$

By Lemma 3.15 this implies, in turn,

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

which coincides with (3.22) (p. 69).

If $\epsilon \ge \epsilon_0$, Lemma 3.13 implies that

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

in combination with Lemma 3.15 this again implies (3.22).

Inequality (3.23) is treated in a similar way to (3.22). Lemmas 3.13 and 3.14 imply that

$$\liminf_{n\to\infty} \mathbb{E}(\mathrm{emp}_n \,|\mathcal{F}_{n+1}) \ge \mathbf{E}(\epsilon) \quad \text{a.s.} \qquad (3.65)$$

(this inequality is vacuously true when $\epsilon \le \epsilon_0$). Another application of Lemma 3.15 gives

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

i.e., (3.23).

Remark The derivation of Proposition 3.5 from Lemmas 3.12–3.15 would be very simple if we defined the nonconformity measure by, say,

$$A(B, (x, \sigma, y)) := \begin{cases} (-\hat{f}_B(x, \sigma), \sigma) & \text{if } y = \hat{y}_B(x, \sigma) \\ (\hat{f}_B(x, \sigma), \sigma) & \text{otherwise} \end{cases}$$

(with the lexicographic order on nonconformity scores) instead of (3.15) (in which case the second addend in the numerator of (3.12) would be just τ_n'' almost surely). Our definition (3.15), however, is simpler and, most importantly, facilitates the proof of Proposition 3.2. Another simplification would be to use Lemma 3.12 (applied to $\epsilon := \epsilon - \mathbf{E}(\epsilon)$) instead of Lemma 3.14 in the derivation of (3.65); we preferred a more symmetric picture.

3.6 Bibliographical remarks

Examples of nonconformity measures

For a derivation of the dual problem (3.6) (p. 58), see Vapnik 1998. The objects x_i with $\alpha_i > 0$ are known as *support vectors*; this is the origin of the name "support vector machine".

The idea of reducing binary classification to regression is an old one. In the case of simple prediction the procedure is as follows: encode the labels as real numbers, one negative and the other positive, apply a regression algorithm, and define $\hat{y}_n$ to be the "positive" label if the value predicted for y_n by the regression algorithm is positive, to be the "negative" label if the value predicted for y_n by the regression algorithm is negative, and define $\hat{y}_n$ arbitrarily if the value predicted for y_n by the regression algorithm is zero. Probably the earliest suggestion of this kind was Fisher's *discriminant analysis* (Fisher 1973b, §49.2): if there are, say, l_1 males and l_2 females in the training set and $l_1 + l_2 = l$, encode males as l_2/l and encode females as $-l_1/l$ (so that the mean of the encodings over the training set is 0), and use the least squares algorithm as the regression algorithm.

The precursor of conformal predictor suggested in Gammerman et al. 1998 used the SVM method as the underlying algorithm. Later it was noticed (see Saunders et al. 1998) that the Lagrange method applied to ridge regression in analogy with SVM leads to α_i equivalent to the residuals, and this in turn lead to the realization that almost any machine learning algorithm can be adapted, often in more than one way, to obtain a nonconformity measure. However, the first genuine conformal predictor (then called "transductive confidence machine") introduced in Vovk et al. 1999 and Saunders et al. 1999 still used the Lagrange multipliers α_i corresponding to constraints (3.4) (p. 57) as the nonconformity measure. The original conformal predictors for multilabel classification problems using binary SVMs were based on (3.7) with $\lambda = 1$, but it was quickly noticed that taking $\lambda < 1$ improves results dramatically.

There is a version of SVM for regression (see Vapnik 1998, Chap. 11), which can also be used for computing nonconformity scores.

We mentioned two methods of reducing multilabel classification problems to binary ones: "one-against-the-rest" and "one-against-one". Another popular method, based on error-correcting coding, was proposed by Dietterich and Bakiri (1995).

Instead of reducing a multilabel classification problem to the binary case and then applying the SVM method, it is possible to use directly known multilabel generalizations of SVM. First such generalization was proposed by Blanz and Vapnik (Vapnik 1998, §10.10); later but independently it was found by Watkins and Weston (1999) and Jaakkola.

Universal predictor

The first step towards a universal predictor was done in Vovk 2002a, where it was shown that an optimal smoothed conformal predictor exists when the power distribution Q^∞ generating the examples is known. The full result was announced in Vovk 2003a.

Alternative protocols

Several papers (such as Rivest and Sloan 1988, Freund et al. 2004) extend the standard PAC framework by allowing the prediction algorithm to abstain from making a prediction at some trials. Our results show that for any significance level ϵ there exists a prediction algorithm that: (a) makes a wrong prediction with frequency at most ϵ; (b) has an optimal frequency of abstentions among the prediction algorithms that satisfy property (a). The protocol of Rivest and Sloan (1988) and Freund et al. (2004) is in fact a restriction of our protocol, in which Predictor is only allowed to output a one-element set or the whole of $\mathbf{Y}$; the latter is interpreted as abstention. (And in the situation where Err_n and Mult_n are of primary interest, as in this chapter, the difference between these two protocols is not very significant.) The universal predictor can be adapted to the restricted protocol by replacing a multiple prediction with $\mathbf{Y}$ and replacing an empty prediction with a randomly chosen label. In this way we obtain a prediction algorithm in the restricted protocol which is asymptotically conservative and has an optimal frequency of abstentions, in the sense of (3.9) (p. 61), among the asymptotically conservative algorithms.

The methods of Freund et al. (2004) are directly applicable to conformal prediction; in particular, that paper defines a natural nonconformity measure (the "empirical log ratio", taken with appropriate sign) in the situation where a hypothesis class is given.

Confidence and credibility

In the situation where Mult_n and Emp_n are the principal measures of predictive efficiency, it is very natural to summarize the range of possible prediction sets Γ^ϵ, $\epsilon \in (0,1)$, by reporting the *confidence*

$$\sup\{1 - \epsilon : |\Gamma^\epsilon| \leq 1\}\,, \tag{3.66}$$

the *credibility*

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

and the *prediction* Γ^ϵ, where $1 - \epsilon$ is the confidence (in this case Γ^ϵ is never multiple for conformal predictors and usually contains exactly one label). Reporting the prediction, confidence, and credibility was suggested in Vovk et al. 1999 and Saunders et al. 1999; it is analogous to reporting the observed level of significance (Cox and Hinkley 1974, p. 66) in statistics.

4

Modifications of conformal predictors

So far we have emphasized desirable properties of conformal predictors: validity (Chap. 2), asymptotic efficiency (Chap. 3), and flexibility (ability to incorporate a wide range of machine-learning methods); we have also mentioned that the hedged predictions output by good conformal predictors are "conditional", in the sense that they take full account of the object to be predicted. In this chapter we will discuss some limitations of conformal prediction and ways to overcome or alleviate these limitations.

The first problem, dealt with in §4.1, is the relative computational inefficiency of conformal predictors. In that section we construct "inductive conformal predictors" (ICP), whose computational efficiency is often much better; the price is some loss in predictive efficiency (which was called simply "efficiency" in the previous chapters). In §4.2 we introduce several new nonconformity measures, which are especially natural when used with ICP.

"Weak teachers", which are allowed to provide the true label with a delay or not to provide it at all, are considered in §4.3. We introduce a formal notion of a "teaching schedule", which is a fairly general protocol for disclosing labels of observed objects including several interesting special cases. The main result of that section is a characterization of teaching schedules under which the method of conformal prediction remains "asymptotically valid in probability".

The protocol with a weak teacher is a relaxation of the pure on-line protocol in the direction of the off-line setting. After showing in §4.3 that conformal predictors retain some properties of validity in the mixed protocol, in §4.4 we discuss simple validity properties of off-line conformal predictors and inductive conformal predictors, and also briefly consider a mixed protocol for inductive conformal predictors.

The issue of conditionality is taken up in §4.5. The potentially serious problem with conformal predictors is that they are not automatically *conditionally valid*: e.g., in the USPS data set some digits (such as "5") are more difficult to recognize correctly than other digits (such as "0"), and it is natural to expect that at the confidence level 95% the error rate will be significantly greater than 5% for the difficult digits; our usual, unconditional, notion of

validity only ensures that the average error rate over all digits will be close to 5%. The notion of Mondrian conformal predictor is introduced to address this concern.

4.1 Inductive conformal predictors

We start by looking more closely at the reasons for the relative computational inefficiency of conformal predictors for large data sets. For concreteness, we will discuss the conformal predictors determined by the simplest nonconformity measures (2.23) and (2.24) (p. 29), but the phenomenon is general. (The notion of ICP itself will be closer to conformal predictors determined by (2.24), in that ICPs never use the value of a prediction rule on examples from which the rule was found.)

As discussed in Chap. 1, one can usually assign a simple predictor to one of two types: "inductive" or "transductive". For inductive predictors, $D_{\lbag z_1,\dots,z_n \rbag}$ can be computed, in some sense: e.g., $D_{\lbag z_1,\dots,z_n \rbag}$ may be described by a polynomial, and computing $D_{\lbag z_1,\dots,z_n \rbag}$ may mean computing the coefficients of the polynomial; as soon as $D_{\lbag z_1,\dots,z_n \rbag}$ is computed, computing $D_{\lbag z_1,\dots,z_n \rbag}(x)$ for a new object x takes very little time. For transductive predictors, relatively little can be done before seeing the new object x; even allowing considerable time for pre-processing $\lbag z_1,\dots,z_n \rbag$, computing $D_{\lbag z_1,\dots,z_n \rbag}(x)$ will be a difficult task.

Notice that, even when D is an inductive algorithm, the confidence predictor based on the generic nonconformity measure (2.23) (and, even more so, on (2.24)) will still be computationally inefficient: for every new object x_n, computing $\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ will require constructing new prediction rules. Inductive conformal predictors will be defined in such a way that they can make significant computational savings when the underlying simple predictor D is inductive.

To define an ICP from a nonconformity measure (A_n) first fix a finite or infinite sequence of positive integer parameters $m_1, m_2, \dots$ (called *update trials*); it is required that $m_1 < m_2 < \cdots$. If the sequence $m_1, m_2, \dots$ is finite, $(m_1, m_2, \dots) = (m_1, \dots, m_r)$, we set $m_i := \infty$ for $i > r$. The *ICP* determined by (A_n) and the sequence $m_1, m_2, \dots$ of update trials is defined to be the confidence predictor Γ such that the prediction sets $\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ are computed as follows:

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

$$\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) := \left\{ y \in \mathbf{Y} : \frac{|\{j = m_k + 1, \dots, n : \alpha_j \ge \alpha_n\}|}{n - m_k} > \epsilon \right\}, \quad (4.1)$$

where the nonconformity scores α_j are defined by

$$\begin{aligned} \alpha_j &:= A_{m_k+1}\left(\Lbag (x_1,y_1),\ldots,(x_{m_k},y_{m_k})\Rbag, (x_j,y_j)\right), \\ &\qquad \text{for } j = m_k+1,\ldots,n-1, \\ \alpha_n &:= A_{m_k+1}\left(\Lbag (x_1,y_1),\ldots,(x_{m_k},y_{m_k})\Rbag, (x_n,y)\right) . \end{aligned} \tag{4.2}$$

Smoothed ICPs can be defined analogously to smoothed conformal predictors: instead of (4.1) we have

$$\begin{aligned} &\Gamma^\epsilon(x_1,\tau_1,y_1,\ldots,x_{n-1},\tau_{n-1},y_{n-1},x_n,\tau_n) := \\ &\qquad \left\{ y \in \mathbf{Y} : \frac{|\{j : \alpha_j > \alpha_n\}| + \tau_n |\{j : \alpha_j = \alpha_n\}|}{n - m_k} > \epsilon \right\} , \end{aligned} \tag{4.3}$$

where $j = m_k+1,\ldots,n$ and $\tau_n \in [0,1]$ are the random numbers. The following result (which is also a special case of Theorem 8.1 on p. 193) shows that Propositions 2.3 and 2.4 continue to hold in the case of ICPs and smoothed ICPs, respectively.

Proposition 4.1. *All ICPs are conservatively valid. All smoothed ICPs are exactly valid.*

The general scheme for defining nonconformity

For use with inductive confidence predictors, it is convenient to rewrite the definitions (2.23) and (2.24) more explicitly as

$$A\left(\Lbag (x_1,y_1),\ldots,(x_l,y_l)\Rbag, (x,y)\right) := \Delta\left(y, D_{\Lbag (x_1,y_1),\ldots,(x_l,y_l),(x,y)\Rbag}(x)\right) \tag{4.4}$$

and

$$A\left(\Lbag (x_1,y_1),\ldots,(x_l,y_l)\Rbag, (x,y)\right) := \Delta\left(y, D_{\Lbag (x_1,y_1),\ldots,(x_l,y_l)\Rbag}(x)\right) , \tag{4.5}$$

respectively. In the case where A is defined by (4.5), we can see that the ICP requires recomputing the prediction rule being used not at every trial but only at the update trials $m_1, m_2, \ldots$; the rate of growth of m_i determines the chosen balance between predictive and computational efficiency. The simplest nontrivial case, where there is only one update trial m_1, is discussed in §4.4 below.

Let a and b be positive numbers such that either $a \ge 1$ and $b \ge 1$ or $a > 1$, and suppose that the prediction rule $D_{\Lbag z_1,\ldots,z_n\Rbag}$ is computable in time $\Theta(n^a \log^b n)$ and the discrepancy measure Δ is computable in constant time. Then the conformal predictor determined by (4.5) spends time $\Theta(n^{a+1} \log^b n)$ on the computations needed for the first n trials. On the other hand, if the sequence m_i is infinite and grows exponentially fast, the ICP based on D, Δ, and (m_i) spends the same, to within a constant factor, time $\Theta(n^a \log^b n)$. (We have been assuming that the conformal predictor and ICP are given D as an oracle and that the label space $\mathbf{Y}$ is finite and fixed.) In the case where the sequence m_i is finite, the ICP's computation time becomes

$$\Theta(n \log n) \tag{4.6}$$

(e.g., use red-black trees for storing the nonconformity scores, as in the proof of Proposition 3.2 on p. 75, but augment them with information needed to find the rank of an element in time $O(\log n)$ – see Cormen et al. 2001, §14.1).

4.2 Further ways of computing nonconformity scores

All nonconformity measures described in the previous chapters can be used in inductive conformal prediction, and all nonconformity measures that will be introduced in this section can be used in conformal prediction. The nonconformity measures of this section are, however, especially convenient in the case of ICPs.

Suppose we are given a bag

$$\Lbag z_1,\ldots,z_l\Rbag \in \mathbf{Z}^{(*)} \tag{4.7}$$

and an example $z \in \mathbf{Z}$, fixed for the rest of this section. The problem is to define the nonconformity score

$$A\left(\Lbag z_1,\ldots,z_l\Rbag, z\right) , \tag{4.8}$$

which we will usually abbreviate to $A(z)$. In the context of inductive conformal prediction, we are interested in $l = m_1, m_2, \ldots$. Sometimes it will be more convenient to define the conformity score $B(z)$ instead. As usual, we write (x_i, y_i) for z_i and (x,y) for z when we need separate notations for the objects and labels.

We start from applying two nonconformal measures introduced earlier to ICP. The nonconformity measure used to define the deleted LSCM, $\alpha_i = |e_{(i)}|$ with the deleted residuals $e_{(i)}$ defined by (2.35) (p. 34), can be rewritten in our present context as

$$A\left(\Lbag z_1,\ldots,z_l\Rbag, z\right) = |y - \hat{y}| , \tag{4.9}$$

where $\hat{y}$ is the least squares prediction for y as computed from the training set $\Lbag z_1,\ldots,z_l\Rbag$ and x. The nonconformity measure (2.36) used to define the studentized LSCM can be rewritten as

$$A\left(\Lbag z_1,\ldots,z_l\Rbag, z\right) = \frac{|y - \hat{y}|}{\sqrt{1 + x'(X'X)^{-1}x}} , \tag{4.10}$$

where, in addition, X is the $l \times p$ matrix $(x_1,\ldots,x_l)'$. This can be checked using the standard formula

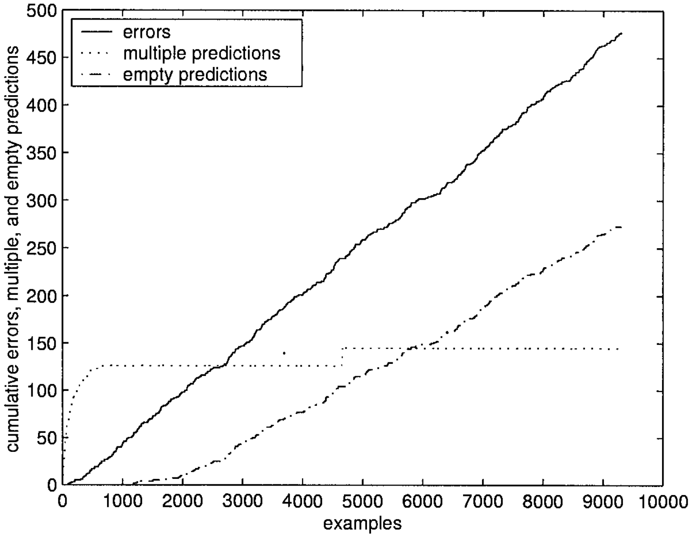

Fig. 4.1. On-line performance of the 1-nearest neighbor ICP with the update trial 4649 (= 9298/2) on the USPS data set for the confidence level 95%. In accordance with Proposition 4.1, starting from scratch at trial 4670 does not affect the error rate (solid line)

$$(K + uv')^{-1} = K^{-1} - \frac{K^{-1}uv'K^{-1}}{1 + v'K^{-1}u}, \tag{4.11}$$

where K is a square matrix and u and v are vectors.

The definition (3.1) (p. 54) of nonconformity measures based on the nearest neighbors classification is already given in the form (4.8) convenient for use with ICP. In the case of regression (p. 38), $A(x, y)$ is defined as $|y - \hat{y}|$, where $\hat{y}$ is the k-NNR prediction for x computed from the bag (4.7).

The performance of the 1-nearest neighbor ICP on the USPS data set with update trial 4649 (the middle of the data set) is shown in Figs. 4.1 and 4.2. It can be seen from these figures (and is obvious anyway) that the ICP's performance (measured by the number of multiple predictions) deteriorates sharply after update trials m_i. (There is a hike of approximately $1/\epsilon$ in the number of multiple predictions, where ϵ is the significance level used.) Perhaps in practice there should be short spells of "learning" after each update trial, when the ICP is provided with fresh "training examples" and its predictions are not used or evaluated.

It is not clear how the way of computing nonconformity scores from SVM, as given in §3.1, could be used by ICP in a computationally efficient way. The easiest solution is perhaps to compute the SVM prediction rule based on the bag (4.7) and define $A(x, y)$ to be the distance (perhaps in a feature space) between x and the optimal separating hyperplane (taken with the minus sign if the SVM prediction for x is different from y).

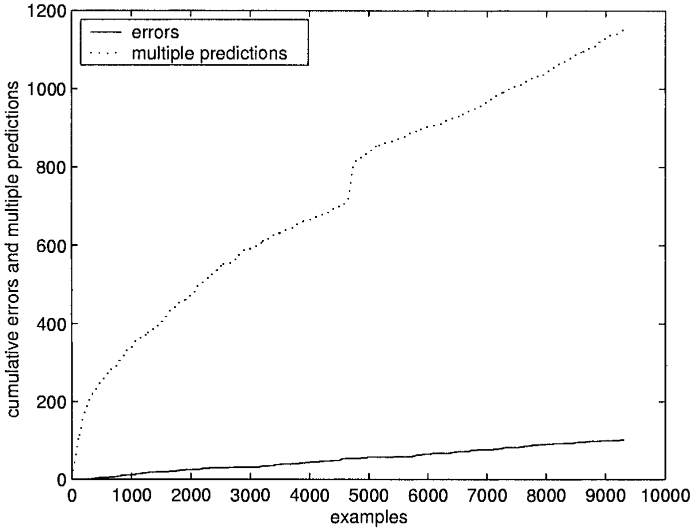

Fig. 4.2. On-line performance of the 1-nearest neighbor ICP with the update trial 4649 on the USPS data set for the confidence level 99%

In the rest of this section we will describe new ways of detecting nonconformity which are especially natural in the case of inductive conformal prediction.

De-Bayesing

Suppose we have a Bayesian model (compatible with the randomness assumption) for the process of generating the label y given the object x. If we fully trust the model, we can use it for computing, e.g., predictive densities and prediction sets in the form of highest probability density regions (see, e.g., Bernardo and Smith 1994, §5.1). We are, however, interested in the case where the Bayesian model is plausible, but we do not really believe it. If it happens to be true, we would like our confidence predictor to be efficient. But we also want it to be always valid, even if the Bayesian model is wrong.

A natural definition of nonconformity measure (4.8) is as follows: find the posterior (after seeing the old examples $z_1, \ldots, z_l$ and the new object x) conditional distribution p for the label y given x, and define the conformity score for (x, y) as

$$B\left(\Lbag z_1, \ldots, z_l \Rbag, (x, y)\right) := p\{y\} \tag{4.12}$$

in the case of classification ($\mathbf{Y}$ is finite) and

$$B\,(\wr z_1,\dots,z_l \int,(x,y)) := \min(p((-\infty,y]),p[y,\infty)) \tag{4.13}$$

in the case of regression ($\mathbf{Y} = \mathbb{R}$). In both cases, $B(x,y)$ is small when the label is strange under the Bayesian model, so the corresponding ICP is likely to be predictively efficient. The conditional probability distribution for the next example z given $z_1,\dots,z_l$ can be computed before seeing x, which may lead to a computationally efficient ICP. And of course, the ICP will be valid automatically.

The ICP determined by one of these conformity measures may be said to be the result of "de-Bayesing" of the original Bayesian algorithm. More generally, we can also say that the RRCM algorithm of §2.3 is a de-Bayesed version of ridge regression (it is well known, and demonstrated in Chap. 10, that ridge regression is the Bayesian algorithm for a normal prior).

Bootstrap

The basic idea of bootstrap is to use resampling (sampling from the sample, obtaining what is called *bootstrap samples*) to get an idea of the variability of the value of interest (for details, see Efron and Tibshirani 1993, Davison and Hinkley 1997). Let us again consider the case of regression.

One way to implement this idea is as follows. Find the least squares weights $\hat{w} := (X'X)^{-1}X'Y$ from the training set (4.7), where X is the $l \times p$ matrix $(x_1,\dots,x_l)'$ of the objects in the training set (assuming the object space is $\mathbf{X} = \mathbb{R}^p$) and Y is the corresponding $l \times 1$ vector of the labels in the training set. Let $\hat{G}$ be the uniform distribution on the *centered modified residuals* $r_i - \overline{r}$, where

$$r_i := \frac{y_i - \hat{y}_i}{\sqrt{1-h_{ii}}}, \quad \hat{y}_i := \hat{w}\cdot x_i$$

(cf. (2.36) on p. 34), and

$$\overline{r} := \frac{1}{l}\sum_{i=1}^{l} r_i\ .$$

(That is, $\hat{G}$ puts the same weight $1/l$ on each r_i.) Let ξ_r^*, $r = 1,2,\dots$, be a sequence of independent random vectors in $\mathbb{R}^l$ whose components are independent and distributed as $\hat{G}$. Obtain RM (where R should be large enough and $M = 1$ is acceptable) "prediction errors" $\delta^*_{r,m}$ in the following way:

```
FOR $r = 1,\dots,R$:
    $Y_r^* := X\hat{w} + \xi_r^*$;
    $\hat{w}_r^* := (X'X)^{-1}X'Y_r^*$ (least squares estimate from $X$ and $Y_r^*$);
    FOR $m = 1,\dots,M$:
        sample $\epsilon_m^*$ from $\hat{G}$;
        $\delta^*_{r,m} := (\hat{w}\cdot x + \epsilon_m^*) - \hat{w}_r^*\cdot x$
    END FOR
END FOR.
```

From the prediction errors for the new object x we can compute the corresponding conformity score for the full example (x,y) as, e.g.,

$$B(x,y) := \min\left(\left|\{(r,m) : y-\hat{y} \le \delta^*_{r,m}\}\right|, \left|\{(r,m) : y-\hat{y} \ge \delta^*_{r,m}\}\right|\right)\ ,$$

where $\hat{y} := \hat{w}\cdot x$.

Decision trees

A decision tree (for a detailed description see, e.g., Mitchell 1997, Chap. 3) is a way of classifying the objects into a finite number of classes. The classification is performed by testing the values of different attributes, but the details will not be important for us.

There are many methods of constructing a decision tree from a training set of examples. One of the most popular methods is Quinlan's (1993) C4.5, but again we do not need the precise details. We will assume that each class contains at least one object from the training set: if this is not the case, the decision tree can always be "pruned" to make sure this property holds.

After a decision tree is constructed from the training set, we can define a conformity score $B(x,y)$ of the new example (x,y) as the percentage of examples labeled as y among the training examples whose objects are classified in the same way as x by the decision tree.

Boosting

Boosting is, as its name suggests, a method for improving the performance of a given prediction algorithm, usually called the *weak learner*[1]. As usual in the boosting literature, we will assume that the weak learner can be applied to the training set (4.7) in which each example z_i, $i = 1,\dots,l$, is taken with a nonnegative weight w_i, with the weights summing to 1. If a weak learner cannot process weighted examples, a bag of training examples of the same size l should be sampled from the probability distribution $D\{x_i\} := w_i$, and this bag is then used to train the weak learner. The output of the weak learner is a prediction rule $h : \mathbf{X} \to \mathbf{Y}$.

Let us assume, for simplicity, that $\mathbf{Y} = \{-1,1\}$. One of the most popular boosting algorithms, AdaBoost.M1, works as follows:

```
start with the probability distribution $D_1\{i\} := 1/m$, $i = 1,\dots,m$;
FOR $t = 1,\dots,T$:
    call the weak learner providing it with $D_t$;
    get back the prediction rule $h_t : \mathbf{X} \to \mathbf{Y}$;
    compute the error $\epsilon_t := \sum_{i=1,\dots,l:h_t(x_i)\ne y_i} D_t\{i\}$;
    $\alpha_t := \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$;
```

[1] There is no connection between weak learners and our "weak teachers" considered in the next section.

```
        update $D_{t+1}\{i\} := D_t\{i\}\mathrm{e}^{-\alpha_t y_i h_t(x_i)}/Z_t$, $i = 1, \dots, l$,
            where $Z_t$ is the normalizing constant
    END FOR.
```

The normalizing constant Z_t is chosen to make D_{t+1} a probability distribution. The result of the boosting procedure is the function $f : \mathbf{X} \to \mathbb{R}$ defined by

$$f(x) := \frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} \alpha_t} \, .$$

The prediction for a new object x is computed as $\hat{y} := \operatorname{sign} f(x)$. It can also be used to define the conformity score

$$B(x, y) := y f(x) \tag{4.14}$$

for an example (x, y). This is a natural measure from the theoretical point of view (Schapire et al. 1998, Theorem 5) and gives reasonable empirical results on benchmark data sets (Proedrou 2003).

Another natural way to define the conformity score of an example (x, y) is to use a conformity measure for the weak learner. Suppose, e.g., that the weak learner is a method for constructing decision trees. Then we can define the conformity score of (x, y) as

$$B(x, y) := \sum_{t=1}^{T} \alpha_t B_t(x, y) \, , \tag{4.15}$$

where $B_t(x, y)$ is the conformity score of (x, y) computed from h_t, as described in the previous subsection. No significant difference in the empirical performance of (4.14) and (4.15) was found in Proedrou 2003.

Neural networks

Let $|\mathbf{Y}| < \infty$ (neural networks are usually used for classification). When fed with an object $x \in \mathbf{X}$, a neural network outputs a set of numbers o_y, $y \in \mathbf{Y}$, such that o_y reflects the likelihood that y is x's label. (See Mitchell 1997 for details.) Inductive conformal predictors determined by nonconformity scores

$$A(x, y) := \frac{\sum_{y' \in \mathbf{Y} : y' \neq y} o_{y'}}{o_y + \gamma} \, , \tag{4.16}$$

where $\gamma \geq 0$ is a suitably chosen parameter, have been shown to have a reasonable empirical performance (Papadopoulos 2004). Results change little if the $\sum$ in (4.16) is replaced by max.

Logistic regression

The logistic regression model is only applicable in the case $\mathbf{Y} = \{0, 1\}$ and $\mathbf{X} = \mathbb{R}^p$, for some p; according to this model, the conditional probability that $y = 1$ given x for an example (x, y) is given by

$$\frac{\mathrm{e}^{w \cdot x}}{1 + \mathrm{e}^{w \cdot x}}$$

for some weight vector $w \in \mathbb{R}^p$. If $\hat{w}$ is, e.g., the maximum likelihood estimate found from the bag (4.7), it is natural to use the nonconformity measure

$$A(x, y) := \begin{cases} 1 + \mathrm{e}^{-\hat{w} \cdot x} & \text{if } y = 1 \\ 1 + \mathrm{e}^{\hat{w} \cdot x} & \text{if } y = 0 \end{cases} \tag{4.17}$$

(i.e., $1/A(x, y)$ is the estimated probability of the observed y given the observed x for the current example).

Remembering that nonconformity scores can be subjected to a monotonic transformation without changing the prediction sets, we can simplify (4.17) to

$$A(x, y) := \begin{cases} -\hat{w} \cdot x & \text{if } y = 1 \\ \hat{w} \cdot x & \text{if } y = 0 \, . \end{cases}$$

4.3 Weak teachers

In the pure on-line setting, considered so far, we get an immediate feedback (the true label) for every example that we predict. This makes practical applications of this scenario questionable. Imagine, for example, a mail sorting center using an on-line prediction algorithm for zip code recognition; suppose the feedback about the "true" label comes from a human expert. If the feedback is given for every object x_i, there is no point in having the prediction algorithm: we can just as well use the label provided by the expert. It would help if the prediction algorithm could still work well, in particular be valid, if only every, say, tenth object were classified by a human expert. Alternatively, even if the prediction algorithm requires the knowledge of all labels, it might still be useful if the labels were allowed to be given not immediately but with a delay (in our mail sorting example, such a delay might make sure that we hear from local post offices about any mistakes made before giving a feedback to the algorithm). In this section we will see that asymptotic validity still holds in many cases where missing labels and delays are allowed.

In the pure on-line protocol we had validity in the strongest possible sense: at each significance level ϵ each smoothed conformal predictor made errors independently with probability ϵ. Now we will not have validity in this strongest sense, and so we will consider three natural asymptotic definitions, requiring only that $\mathrm{Err}_n^{\epsilon}/n \to \epsilon$ in a certain sense: weak validity, strong validity, and validity in the sense of the law of the iterated logarithm. Finally, we will prove a simple result about asymptotic efficiency.

Imperfectly taught predictors

We are interested in the protocol where the predictor receives the true labels y_n only for a subset of trials n, and even for this subset, y_n may be given with a delay. This is formalized by a function $\mathcal{L} : N \to \mathbb{N}$ defined on an infinite set $N \subseteq \mathbb{N}$ and required to satisfy

$$\mathcal{L}(n) \le n$$

for all $n \in N$ and

$$m \neq n \Longrightarrow \mathcal{L}(m) \neq \mathcal{L}(n)$$

for all $m \in N$ and $n \in N$; a function satisfying these properties will be called a *teaching schedule*. A teaching schedule $\mathcal{L}$ describes the way the data is disclosed to the predictor: at the end of trial $n \in N$ it is given the label $y_{\mathcal{L}(n)}$ for the object $x_{\mathcal{L}(n)}$. The elements of $\mathcal{L}$'s domain N in the increasing order will be denoted n_i: $N = \{n_1, n_2, \dots\}$ and $n_1 < n_2 < \cdots$. We denote the total number of labels disclosed by the beginning of trial n to a predictor taught according to the teaching schedule $\mathcal{L}$ by $s(n) := |\{i : i \in N, i < n\}|$.

Let Γ be a confidence predictor and $\mathcal{L}$ be a teaching schedule. The *$\mathcal{L}$-taught version* $\Gamma^{\mathcal{L}}$ of Γ is

$$\begin{aligned}\Gamma^{\mathcal{L},\epsilon}&(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)\\ &= \Gamma^{\epsilon}(x_{\mathcal{L}(n_1)}, y_{\mathcal{L}(n_1)}, \dots, x_{\mathcal{L}(n_{s(n)})}, y_{\mathcal{L}(n_{s(n)})}, x_n) \,.\end{aligned}$$

Intuitively, at the end of trial n the predictor $\Gamma^{\mathcal{L}}$ learns the label $y_{\mathcal{L}(n)}$ if $n \in N$ and learns nothing otherwise. An *$\mathcal{L}$-taught (smoothed) conformal predictor* is a confidence predictor that can be represented as $\Gamma^{\mathcal{L}}$ for some (smoothed) conformal predictor Γ.

Let us now consider several examples of teaching schedules.

Ideal teacher. If $N = \mathbb{N}$ and $\mathcal{L}(n) = n$ for each $n \in N$, then $\Gamma^{\mathcal{L}} = \Gamma$.

Slow teacher with a fixed lag. If $N = \{l+1, l+2, \dots\}$ for some $l \in \mathbb{N}$ and $\mathcal{L}(n) = n - l$ for all $n \in N$, then $\Gamma^{\mathcal{L}}$ is a predictor which learns true labels with a delay of l.

Slow teacher. The previous example can be generalized as follows. Let $l(n) = n + \text{lag}(n)$ where $\text{lag} : \mathbb{N} \to \mathbb{N}$ is an increasing function. Define $N := l(\mathbb{N})$ and $\mathcal{L}(n) := l^{-1}(n)$, $n \in N$. Then $\Gamma^{\mathcal{L}}$ is a predictor which learns the true label for each object x_n with a delay of $\text{lag}(n)$.

Lazy teacher. If $N \neq \mathbb{N}$ and $\mathcal{L}(n) = n$, $n \in N$, then $\Gamma^{\mathcal{L}}$ is given the true labels immediately but not for every object.

All results of this section (as will be clear from the proofs, given in §4.6) use only the following properties of smoothed conformal predictors Γ:

- At each significance level ϵ, the errors $\text{err}_n^{\epsilon}(\Gamma)$, $n = 1, 2, \dots$, are independent Bernoulli random variables with parameter ϵ.

- The predictions do not depend on the order of the examples learnt so far:

$$\Gamma^{\epsilon}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) = \Gamma^{\epsilon}(x_{\pi(1)}, y_{\pi(1)}, \dots, x_{\pi(n-1)}, y_{\pi(n-1)}, x_n)$$

for any permutation π of $\{1, \dots, n-1\}$. (Remember that we call confidence predictors satisfying this property *invariant.*)

Weak validity

In this subsection we state a necessary and sufficient condition for a teaching schedule to preserve "weak asymptotic validity" of conformal predictors. The condition turns out to be rather weak: feedback should be given at more than a logarithmic fraction of trials.

We start from the definitions, assuming, for simplicity, randomness rather than exchangeability. A randomized confidence predictor Γ is *asymptotically exact in probability* if, for all significance levels ϵ and all probability distributions Q on $\mathbf{Z}$,

$$\frac{1}{n}\sum_{i=1}^{n} \text{Err}_n^{\epsilon}(\Gamma, Q^{\infty}) - \epsilon \to 0$$

in probability. Similarly, a confidence predictor Γ is *asymptotically conservative in probability* if, for all significance levels ϵ and all probability distributions Q on $\mathbf{Z}$,

$$\left(\frac{1}{n}\sum_{i=1}^{n} \text{Err}_n^{\epsilon}(\Gamma, Q^{\infty}) - \epsilon\right)^{+} \to 0$$

in probability.

Theorem 4.2. *Let $\mathcal{L}$ be a teaching schedule with domain $N = \{n_1, n_2, \dots\}$, where $n_1, n_2, \dots$ is a strictly increasing infinite sequence of positive integers.*

- *If $\lim_{k\to\infty}(n_k/n_{k-1}) = 1$, any $\mathcal{L}$-taught smoothed conformal predictor is asymptotically exact in probability.*
- *If $\lim_{k\to\infty}(n_k/n_{k-1}) = 1$ does not hold, there exists an $\mathcal{L}$-taught smoothed conformal predictor which is not asymptotically exact in probability.*

In words, this theorem asserts that an $\mathcal{L}$-taught smoothed conformal predictor is guaranteed to be asymptotically exact in probability if and only if the growth rate of n_k is sub-exponential.

Corollary 4.3. *If $\lim_{k\to\infty}(n_k/n_{k-1}) = 1$, any $\mathcal{L}$-taught conformal predictor is asymptotically conservative in probability.*

Strong validity

Theorem 4.4. *Suppose the example space* $\mathbf{Z}$ *is Borel. Let* Γ *be a smoothed conformal predictor and* $\mathcal{L}$ *be a teaching schedule whose domain is* $N = \{n_1, n_2, \dots\}$, *where* $n_1 < n_2 < \cdots$. *If*

$$\sum_k \left(\frac{n_k}{n_{k-1}} - 1\right)^2 < \infty \tag{4.18}$$

then $\Gamma^{\mathcal{L}}$ *is asymptotically exact.*

This theorem shows that $\Gamma^{\mathcal{L}}$ is asymptotically exact when n_k grows as $\exp(\sqrt{k}/\ln k)$; on the other hand, it does not guarantee that it is asymptotically exact if n_k grows as $\exp(\sqrt{k})$.

Corollary 4.5. *Let* Γ *be a conformal predictor and* $\mathcal{L}$ *be a teaching schedule with domain* $N = \{n_1, n_2, \dots\}$, $n_1 < n_2 < \cdots$. *Under condition (4.18),* $\Gamma^{\mathcal{L}}$ *is an asymptotically conservative confidence predictor.*

Iterated logarithm validity

The following result asserts, in particular, that when n_k are equally spaced a stronger version of asymptotic validity, in the spirit of the law of the iterated logarithm, holds.

Theorem 4.6. *Suppose the domain* $\{n_1, n_2, \dots\}$, $n_1 < n_2 < \cdots$, *of a teaching schedule satisfies* $n_k = O(k)$. *Each* $\mathcal{L}$*-taught smoothed conformal predictor* $\Gamma^{\mathcal{L}}$ *satisfies*

$$\left|\frac{\mathrm{Err}_n^\epsilon(\Gamma^{\mathcal{L}}, Q^\infty)}{n} - \epsilon\right| = O\left(\sqrt{\frac{\ln\ln n}{n}}\right) \quad \text{a.s.}$$

and each $\mathcal{L}$*-taught conformal predictor* $\Gamma^{\mathcal{L}}$ *satisfies*

$$\left(\frac{\mathrm{Err}_n^\epsilon(\Gamma^{\mathcal{L}}, Q^\infty)}{n} - \epsilon\right)^+ = O\left(\sqrt{\frac{\ln\ln n}{n}}\right) \quad \text{a.s.}\,,$$

for each $Q \in \mathbf{P}(\mathbf{Z})$ *at each significance level* ϵ.

Efficiency

We will only consider the case of classification, taking the number of multiple predictions as the primary measure of inefficiency.

If Γ is a confidence predictor and ϵ a significance level, we set

$$U^\epsilon(\Gamma) = \left[\liminf_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma)}{n}, \limsup_{n\to\infty} \frac{\mathrm{Mult}_n^\epsilon(\Gamma)}{n}\right].$$

The intervals $U^\epsilon(\Gamma)$ characterize the asymptotical efficiency of Γ; of course, these are random intervals, since they depend on the actual examples output by Reality. It turns out, however, that in the most important case (covering conformal predictors and $\mathcal{L}$-taught conformal predictors, smoothed and deterministic) these intervals are close to being deterministic under the assumption of randomness.

Lemma 4.7. *For each invariant confidence predictor* Γ *(randomized or deterministic), significance level* ϵ, *and probability distribution* Q *on* $\mathbf{Z}$ *there exists an interval* $[a, b] \subseteq (0, 1)$ *such that*

$$U^\epsilon(\Gamma) = [a, b] \quad \text{a.s.}\,,$$

provided the examples and random numbers (if applicable) are generated from Q^∞ *and* $\mathbf{U}^\infty$ *independently.*

Proof. The statement of this lemma is an immediate consequence of the Hewitt–Savage zero-one law (see, e.g., Shiryaev 1996, Theorem IV.1.3). □

We will use the notation $U^\epsilon(\Gamma, Q)$ for the interval whose existence is asserted in the lemma; it characterizes the asymptotical efficiency of Γ at significance level ϵ with examples distributed according to Q.

Theorem 4.8. *Let* Γ *be a (smoothed) conformal predictor and* $\mathcal{L}$ *be a teaching schedule defined on* $N = \{n_1, n_2, \dots\}$, *where* $n_1 < n_2 < \cdots$ *is an increasing sequence. If, for some* $c \in \mathbb{N}$, $n_{k+1} - n_k = c$ *from some* k *on, then* $U^\epsilon(\Gamma^{\mathcal{L}}, Q) = U^\epsilon(\Gamma, Q)$ *for all significance levels* $\epsilon \in (0, 1)$ *and all probability distributions* Q *on* $\mathbf{Z}$.

Theorems 4.4 and 4.8 can be illustrated with the following simple example. Suppose only every mth label is revealed to a conformal predictor, and even this is done with a delay of l, where m and l are positive integer constants. Then (smoothed) conformal predictors will remain asymptotically valid, and their asymptotic rate of multiple predictions will not deteriorate.

4.4 Off-line conformal predictors and semi-off-line ICPs

As we discuss in this section, conformal predictors and ICPs can be applied in the pure off-line mode, but we will then only have a weakened guarantee of validity. The notion of ICP, however, has a natural "semi-off-line" version, which is exactly valid. This section's discussion is independent of the previous section's results (except for a short remark at the end), but it will be clear that they can be fruitfully combined.

Suppose we are given a training set $z_1, \dots, z_l$ of examples $z_i = (x_i, y_i)$ and the problem is to predict the labels y_i, $i = l+1, \dots, l+k$, of the *working examples* $z_{l+1}, \dots, z_{l+k}$. The *off-line conformal predictor* outputs the prediction sets

$$\Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) := \left\{ y \in \mathbf{Y} : \frac{|\{j = 1, 2, \dots, l, i : \alpha_j \ge \alpha_i\}|}{l+1} > \epsilon \right\} \tag{4.19}$$

for each working example z_i, $i = l+1, \dots, l+k$, where the nonconformity scores are computed from a nonconformity measure A:

$$\begin{aligned} \alpha_j &:= A_{l+1}\left(\Lbag z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_l, (x_i, y)\Rbag, z_j\right), \quad j = 1, \dots, l, \\ \alpha_i &:= A_{l+1}\left(\Lbag z_1, \dots, z_l\Rbag, (x_i, y)\right) \end{aligned} \tag{4.20}$$

(cf. (2.18) and (2.19) on p. 26).

In a similar way we can define *off-line ICPs*. For concreteness, we restrict ourselves to the nonconformity measures (4.5) (p. 99). The training set is first split into two parts: the *proper training set*

$$(x_1, y_1, \dots, x_m, y_m) \tag{4.21}$$

of size $m < l$ and the *calibration set*

$$(x_{m+1}, y_{m+1}, \dots, x_l, y_l) \tag{4.22}$$

of size $l - m$. For every working object x_i, $i = l+1, \dots, l+k$, compute the prediction sets

$$\Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) := \left\{ y \in \mathbf{Y} : \frac{|\{j = m+1, \dots, l, i : \alpha_j \ge \alpha_i\}|}{l-m+1} > \epsilon \right\}, \tag{4.23}$$

where the nonconformity scores are defined by

$$\begin{aligned} \alpha_j &:= \Delta\left(y_j, D_{\Lbag(x_1,y_1),\dots,(x_m,y_m)\Rbag}(x_j)\right), \quad j = m+1, \dots, l, \\ \alpha_i &:= \Delta\left(y, D_{\Lbag(x_1,y_1),\dots,(x_m,y_m)\Rbag}(x_i)\right). \end{aligned} \tag{4.24}$$

(Cf. (4.1)–(4.2), p. 99, and (4.5).)

For both conformal predictors and ICPs, it is true that

$$Q^\infty\{(x_1, y_1, x_2, y_2, \dots) : y_i \notin \Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i)\} \le \epsilon \tag{4.25}$$

for every $i = l+1, \dots, l+k$, provided all examples are drawn independently from the distribution Q, but the events in (4.25) are not independent and

$$\frac{|\{i = l+1, \dots, l+k : y_i \notin \Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i)\}|}{k} \tag{4.26}$$

can be significantly above ϵ even when k is very large. (Cf. the description of the "inductivist objection" in §10.2.)

To ensure validity of the off-line ICP, we can modify the application of the ICP constructed from the training set to the working set: after processing each working example (x_i, y_i), $i = l+1, \dots, l+k$, the corresponding nonconformity score α_i should be added to the pool of nonconformity scores used in generating the prediction sets for the following working examples. Formally, redefine

$$\Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) := \left\{ y \in \mathbf{Y} : \frac{|\{j = m+1, \dots, i : \alpha_j \ge \alpha_i\}|}{i-m} > \epsilon \right\}, \tag{4.27}$$

where the nonconformity scores are defined by

$$\begin{aligned} \alpha_j &:= \Delta\left(y_j, D_{\Lbag(x_1,y_1),\dots,(x_m,y_m)\Rbag}(x_j)\right), \quad j = m+1, \dots, i-1, \\ \alpha_i &:= \Delta\left(y, D_{\Lbag(x_1,y_1),\dots,(x_m,y_m)\Rbag}(x_i)\right). \end{aligned} \tag{4.28}$$

Proposition 4.1 (p. 99) says that this modification is conservatively valid, and so (4.26) will not exceed ϵ, up to statistical fluctuations.

Notice that in the case $k \ll (l - m)$ the *semi-off-line ICP* (4.27) differs so little from the off-line ICP that the latter can be expected to be "nearly conservative".

Let us give a formal definition. A confidence predictor Γ is *(δ_n)-conservative*, where $\delta_1, \delta_2, \dots$ is a sequence of nonnegative numbers, if for any exchangeable probability distribution P on $\mathbf{Z}^\infty$ there exists a probability space with two families

$$(\xi_n^{(\epsilon)} : \epsilon \in (0,1), n = 1, 2, \dots), \quad (\eta_n^{(\epsilon)} : \epsilon \in (0,1), n = 1, 2, \dots)$$

of $\{0,1\}$-valued random variables such that:

- for a fixed ϵ, $\xi_1^{(\epsilon)}, \xi_2^{(\epsilon)}, \dots$ is a sequence of independent Bernoulli random variables with parameter ϵ;
- for all n and ϵ, $\eta_n^{(\epsilon-\delta_n)} \le \xi_n^{(\epsilon)}$;
- the joint distribution of $\mathrm{err}_n^\epsilon(\Gamma, P)$, $\epsilon \in (0,1)$, $n = 1, 2, \dots$, coincides with the joint distribution of $\eta_n^{(\epsilon)}$, $\epsilon \in (0,1)$, $n = 1, 2, \dots$.

The definition of conservative validity is a special case corresponding to $\delta_n = 0$, $n = 1, 2, \dots$; we are now interested in the case where δ_n are small (at least for a range of n) positive numbers.

Proposition 4.9. *The confidence predictor*

$$\tilde\Gamma^\epsilon(x_1, y_1, \dots, x_{i-1}, y_{i-1}, x_i) := \begin{cases} \Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) & \text{if } i > l \\ \mathbf{Y} & \text{otherwise}, \end{cases}$$

where $\Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i)$ is defined by (4.23), is (δ_i)-conservative, where

$$\delta_i := \begin{cases} \frac{i-l}{l-m} & \text{if } i > l \\ 0 & \text{otherwise}. \end{cases}$$

Proof. Let $\Gamma^\dagger$ be the smoothed semi-off-line ICP corresponding to Γ, and let $i > l$. Define ξ_i and η_i by the requirements that $\xi_i^{(\epsilon)} = 1$ if and only if $\Gamma^\dagger$ makes a mistake when fed with $x_1, y_1, \ldots, x_i, y_i$, and $\eta_i^{(\epsilon)} = 1$ if and only if Γ makes a mistake when fed with $x_1, y_1, \ldots, x_l, y_l, x_i, y_i$, at the significance level ϵ. To ensure $\eta_i^{(\epsilon-\delta_i)} \le \xi_i^{(\epsilon)}$, it is sufficient to require

$$(1-\epsilon)(i-m) \le (1-\epsilon+\delta_i)(l-m+1) \ ,$$

i.e.,

$$\delta_i \ge (1-\epsilon)\frac{i-l-1}{l-m+1} \ . \qquad \square$$

This proof shows that the fraction of errors made by the off-line ICP at a significance level ϵ on the working set does not exceed $\epsilon + k/(l-m)$, up to statistical fluctuations.

ICPs applied in both off-line and semi-off-line modes are computationally quite efficient. Let us see what the computation time will be if standard algorithms for standard computation tasks are used. In the case of simple predictions, the application of the inductive algorithm D found from the training set of size l to the working set of size k requires time

$$\Theta\left(T_{\text{train}} + kT_{\text{appl}}\right) \ ,$$

where T_{train} is the time required for computing the prediction rule $D_{\lbag z_1,\ldots,z_l \rbag}$ and T_{appl} is the time needed to apply this prediction rule to a new object. The off-line ICP (see (4.23) and (4.24)) requires time

$$\Theta\left(T^\dagger_{\text{train}} + (l-m+k)T^\dagger_{\text{appl}} + (l-m)\log(l-m) + k\log(l-m)\right) \ ,$$

where $T^\dagger_{\text{train}}$ is the time required for computing the prediction rule $D_{\lbag z_1,\ldots,z_m \rbag}$ and $T^\dagger_{\text{appl}}$ is the time needed to apply this prediction rule to a new object (we assume that computing Δ is fast); we allow time $(l-m)\log(l-m)$ for sorting the nonconformity scores obtained from the calibration set (Cormen et al. 2001, Part II) and time $\log(l-m)$ for finding the rank of a working nonconformity score in the set of the calibration nonconformity scores (Cormen et al. 2001, Chaps. 12 and 13). In the case of semi-off-line ICP ((4.27), (4.28)), the required time increases only slightly (for moderately large k) to

$$\Theta\left(T^\dagger_{\text{train}} + (l-m+k)T^\dagger_{\text{appl}} + (l-m)\log(l-m) + k\log(l-m+k)\right) \ .$$

As $k \to \infty$, we have the same asymptotic computation time, $\Theta(k \log k)$, as in §4.1 (cf. (4.6) on p. 100). If, however, our goal is only asymptotic conservativeness as $k \to \infty$, by Theorem 4.4 we can keep only a fraction of $(\ln k)^3$ of the nonconformity scores α_i, $l < i \le l+k$, and so the asymptotic computation time will become $\Theta(k \log\log k)$.

4.5 Mondrian conformal predictors

Our starting point in this section is a natural division of examples into several categories: e.g., different categories can correspond to different labels, or kinds of objects, or just be determined by the ordinal number of the example. As we have already discussed, conformal predictors do not guarantee validity within categories: the fraction of errors can be much larger than the nominal significance level for some categories, if this is compensated by a smaller fraction of errors for other categories. This stronger kind of validity, validity within categories, is the main property of Mondrian conformal predictors (MCPs), constructed in this section. As usual, we will demonstrate validity, in this stronger sense, under the exchangeability assumption; this assumption, however, will be relaxed in Chap. 8.

Validity within categories (or *conditional validity*, as we will say) is especially relevant in the situation of *asymmetric classification*, where errors for different categories of examples have different consequences; in this case we cannot allow low error rates for some categories to compensate excessive error rates for other categories. Because of our interest in asymmetric classification, we will mainly use the language of conformal transducers in our exposition. The standard translation into the language of conformal predictors is straightforward (cf. §2.5), but in the case of asymmetric classification one might prefer to add flexibility to this translation: instead of comparing all p-values with the same threshold ϵ we might take different ϵs for different categories.

At the end of this section we discuss several special cases of MCPs, including conformal predictors and ICPs.

Mondrian conformal transducers

We are given a division of the Cartesian product $\mathbb{N} \times \mathbf{Z}$ into *categories*: a measurable function

$$\kappa : \mathbb{N} \times \mathbf{Z} \to K$$

maps each pair (n, z) (z is an example and n will be, in our applications, the ordinal number of this example in the data sequence $z_1, z_2, \ldots$) to its category; K is the measurable space (at most countable with the discrete σ-algebra) of all categories. It is required that the elements $\kappa^{-1}(k)$ of each category $k \in K$ form a rectangle $A \times B$, for some $A \subseteq \mathbb{N}$ and $B \subseteq \mathbf{Z}$. Such a function κ will be called a *Mondrian taxonomy.*

Given a Mondrian taxonomy κ, we first define "Mondrian nonconformity measures" and then Mondrian conformal transducers (MCTs).

A *Mondrian nonconformity measure* based on κ is a family of measurable functions $(A_n : n \in \mathbb{N})$ of the type

$$A_n : K^{n-1} \times \left(\mathbf{Z}^{(*)}\right)^K \times K \times \mathbf{Z} \to \overline{\mathbb{R}} \ .$$

The *smoothed Mondrian conformal transducer (smoothed MCT)* determined by the Mondrian nonconformity measure A_n is the randomized confidence transducer producing the p-values

$$\begin{aligned} p_n &= f(x_1,\tau_1,y_1,\dots,x_n,\tau_n,y_n) \\ &:= \frac{|\{i:\kappa_i=\kappa_n \ \&\ \alpha_i>\alpha_n\}| + \tau_n\,|\{i:\kappa_i=\kappa_n \ \&\ \alpha_i=\alpha_n\}|}{|\{i:\kappa_i=\kappa_n\}|}\,, \end{aligned} \tag{4.29}$$

where i ranges over $\{1,\dots,n\}$, $\kappa_i := \kappa(i,z_i)$, $z_i := (x_i,y_i)$, and

$$\begin{aligned} \alpha_i := A_n\bigl(&\kappa_1,\dots,\kappa_{n-1}, \\ &(k\mapsto \lbag z_j : j\in\{1,\dots,i-1,i+1,\dots,n\} \ \&\ \kappa_j=k \rbag),\kappa_n,z_i\bigr) \end{aligned} \tag{4.30}$$

for $i=1,\dots,n$ such that $\kappa_i=\kappa_n$. As usual, the definition of a *Mondrian conformal transducer (MCT)* is obtained by replacing (4.29) with

$$p_n = f(x_1,y_1,\dots,x_n,y_n) := \frac{|\{i:\kappa_i=\kappa_n \ \&\ \alpha_i\ge\alpha_n\}|}{|\{i:\kappa_i=\kappa_n\}|}\,.$$

In general, a *(smoothed) MCT* based on a Mondrian taxonomy κ is the (smoothed) MCT determined by some Mondrian nonconformity measure based on κ.

We say that a randomized confidence transducer f is *category-wise exact w.r.t. to a Mondrian taxonomy* κ if, for all n, the conditional probability distribution of p_n given $\kappa(1,z_1),p_1,\dots,\kappa(n-1,z_{n-1}),p_{n-1},\kappa(n,z_n)$ is uniform on $[0,1]$, where $z_1,z_2,\dots$ are examples generated from an exchangeable distribution on $\mathbf{Z}^\infty$ and $p_1,p_2,\dots$ are the p-values output by f.

Proposition 4.10. *Any smoothed MCT based on a Mondrian taxonomy κ is category-wise exact w.r.t. to κ.*

This proposition generalizes Proposition 4.1 but is a special case of Theorem 8.2 (the finitary version of Theorem 8.1) on p. 193. It implies the category-wise property of conservative validity for MCT, whose p-values are always bounded above by the p-values from the corresponding smoothed MCT.

Using Mondrian conformal transducers for prediction

An example of asymmetric classification is distinguishing between useful messages and spam in the problem of e-mail filtering: classifying a useful message as spam is a more serious error than vice versa. In this case we might want to have different significance levels ϵ_k for different categories k.

Let f be a (smoothed) MCT. Given a set of significance levels ϵ_k, $k\in K$, we can define the prediction set for the label y_n of a new object x_n given old examples $z_1,\dots,z_{n-1}$ as

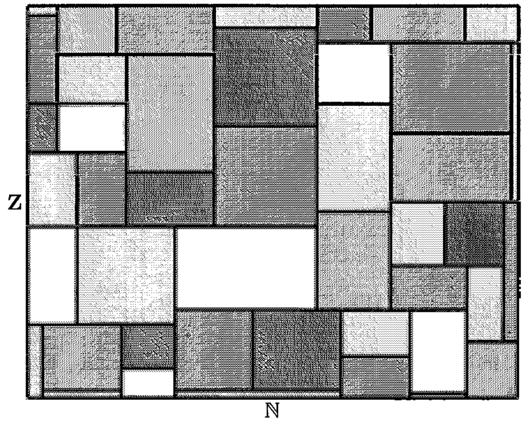

Fig. 4.3. A random Mondrian taxonomy (after Piet Mondrian, 1918)

$$\begin{aligned} &\Gamma^{(\epsilon_k:k\in K)}(z_1,\dots,z_{n-1},x_n) \\ &\qquad := \bigl\{y\in\mathbf{Y} : f(z_1,\dots,z_{n-1},(x_n,y)) > \epsilon_{\kappa(n,(x_n,y))}\bigr\}\,. \end{aligned}$$

Proposition 4.10 now implies that the long-run frequency of errors made by this predictor (*Mondrian conformal predictor*, or *MCP*) on examples of category k does not exceed (approaches, in the case of smoothed transducer) ϵ_k, for each k.

As in the case of conformal prediction, in applications it is usually not wise to fix thresholds ϵ_k, $k\in K$, in advance. One possibility would be to suitably choose three sets of significance levels (ϵ_k^1), (ϵ_k^2), and (ϵ_k^3) such that $\epsilon_k^1\le\epsilon_k^2\le\epsilon_k^3$ for all $k\in K$, and say that $\Gamma^{\epsilon^1}(z_1,\dots,z_{n-1},x_n)$ is a highly confident prediction, $\Gamma^{\epsilon^2}(z_1,\dots,z_{n-1},x_n)$ is a confident prediction, and $\Gamma^{\epsilon^3}(z_1,\dots,z_{n-1},x_n)$ is a casual prediction.

Generality of Mondrian taxonomies

We will next consider several classes of MCTs, involving different taxonomies. In this subsection we consider a natural partial order on the taxonomies, which will clarify the relation between different special cases. (We will use the expression "more general than" for this partial order; it might seem strange here but will be explained in §8.4.) There are many ways to split the rectangle $\mathbb{N}\times\mathbf{Z}$ into smaller rectangles (cf. Fig. 4.3), and it is clearly desirable to impose some order.

We say that a Mondrian taxonomy κ_1 is *more general* than another Mondrian taxonomy κ_2 if, for all pairs (n',z') and (n'',z''),

$$\kappa_1(n',z')=\kappa_1(n'',z'') \Longrightarrow \kappa_2(n',z')=\kappa_2(n'',z'')\,.$$

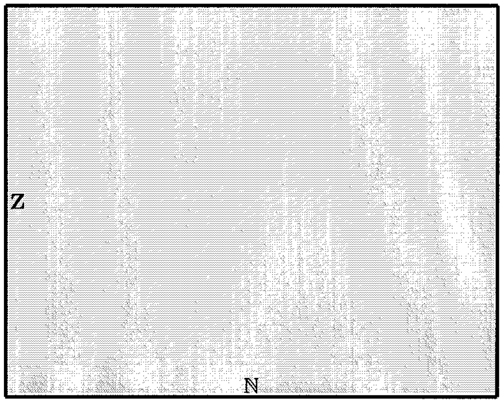

Fig. 4.4. Mondrian taxonomy corresponding to conformal transducers

We will say that κ_1 and κ_2 are *equivalent* if each of them is more general than the other, and we will sometimes identify equivalent Mondrian taxonomies. Identifying equivalent Mondrian taxonomies means that we are only interested in the equivalence relation a given Mondrian taxonomy κ induces ((n', z') and (n'', z'') are *κ-equivalent* if $\kappa(n', z') = \kappa(n'', z'')$) and not in the chosen labels $\kappa(n, z)$ for the equivalence classes.

Since we are only interested in taxonomies with at most countable number of categories, the following proposition immediately follows from the standard properties of conditional expectations (see property 2 on p. 280).

Proposition 4.11. *Let a taxonomy κ_1 be more general than a taxonomy κ_2. If a randomized confidence transducer is category-wise exact w.r. to κ_1, it is category-wise exact w.r. to κ_2.*

Conformal transducers

Conformal transducers are MCTs based on the least general (i.e., constant, see Fig. 4.4) Mondrian taxonomy. Proposition 2.4 (p. 27), asserting that smoothed conformal predictors are exact, is a special case of Proposition 4.10.

In the rest of this section we will describe several experimental results for the USPS data set (randomly permuted), using the 1-nearest neighbor ratio (3.1) (p. 54) as the nonconformity measure. We start from results demonstrating the lack of conditional validity for conformal predictors. The USPS data set is reasonably balanced in the proportion of examples labeled by different digits; for less well-balanced data sets the lack conditional validity of non-Mondrian conformal predictors is often even more pronounced.

Figure 4.5 (plotting Err_n, ϵn, Mult_n, and Emp_n against n for the confidence level 95%; the plots for Err_n, Mult_n, and Emp_n are almost indistinguishable from the analogous plots for the deterministic conformal predictor) shows that the smoothed conformal predictor is valid "on average" on the USPS data set.

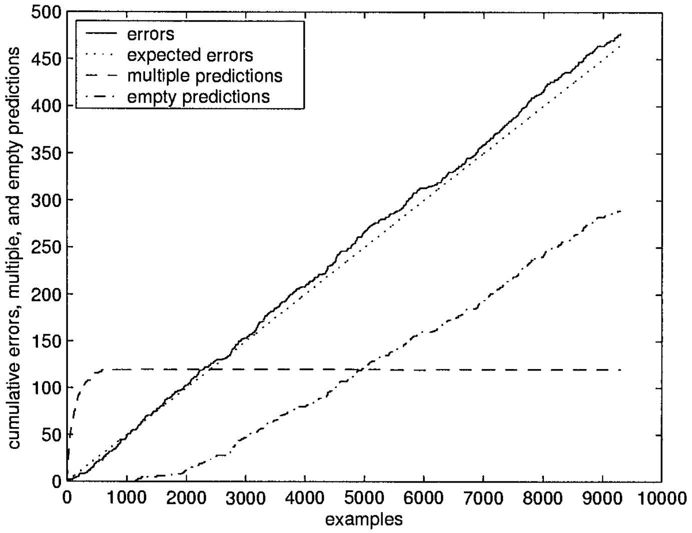

Fig. 4.5. The performance of the smoothed conformal predictor on the USPS data set at the 95% confidence level

Figure 4.6 gives similar plots, but only taking into account the predictions made for the examples labeled "5". It shows that the smoothed conformal predictor is not valid at the 95% confidence level on those examples, giving 11.7% of errors. Since the error rate of 5% is achieved on average, the error rate for some digits is better than 5%; for example, it is below 1% for the examples labeled "0".

Inductive conformal transducers

Inductive conformal transducers, which output the p-values

$$\frac{|\{j : \alpha_j \ge \alpha_n\}|}{n - m_k}$$

(deterministic case) or

$$\frac{|\{j : \alpha_j > \alpha_n\}| + \tau_n |\{j : \alpha_j = \alpha_n\}|}{n - m_k}$$

(smoothed case), where

$$\alpha_j := A_{m_k+1}\left(\lbag (x_1, y_1), \dots, (x_{m_k}, y_{m_k}) \rbag, (x_j, y_j)\right), \quad j = m_k + 1, \dots, n$$

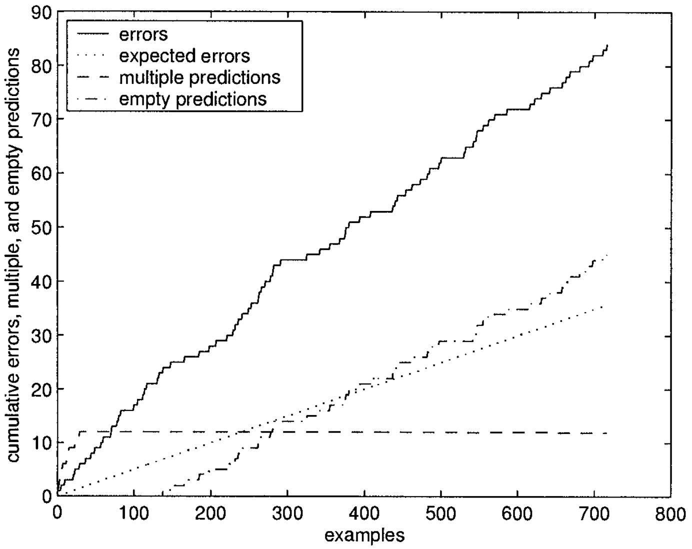

Fig. 4.6. The performance of the smoothed conformal predictor on the USPS data set for the examples labeled "5" at the 95% confidence level

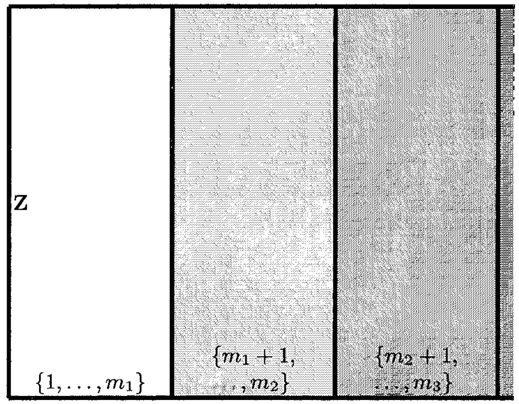

Fig. 4.7. Mondrian taxonomy corresponding to inductive conformal transducers

(cf. (4.2) and (4.3), p. 99), are also a special case of MCTs. The corresponding taxonomy is shown in Fig. 4.7. The result of §4.1 that ICPs are valid is a special case of Proposition 4.10. Similarly to conformal predictors, ICPs sometimes violate the property of label-wise validity.

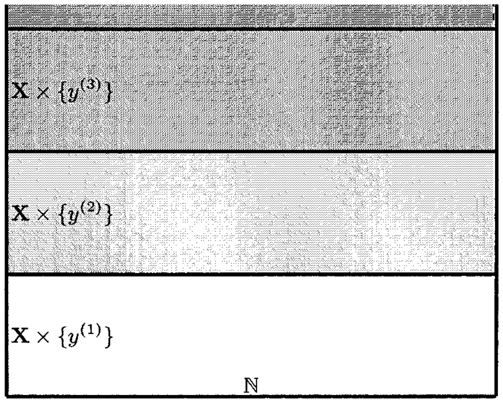
$\mathbf{X} \times \{y^{(3)}\}$
$\mathbf{X} \times \{y^{(2)}\}$
$\mathbf{X} \times \{y^{(1)}\}$
$\mathbb{N}$

Fig. 4.8. Label-conditional Mondrian taxonomy

Label-conditional Mondrian conformal transducers

An important special case is where the category of an example is determined by its label. The corresponding taxonomy is shown in Fig. 4.8, where it is assumed that $\mathbf{Y} = \{y^{(1)}, \dots, y^{(L)}\}$.

Our experiments will be restricted to the "symmetric" case, where the same significance level (5%) is used for all categories. Figure 4.9 demonstrates empirically the category-wise validity of MCPs. In contrast to Fig. 4.6, the label-conditional MCP gives 5.3% of errors when the significance level is set to 5% for the label "5". Figures 4.6 and 4.9 show that the correction in the number of errors results in an increased frequency of multiple predictions; there is also a decrease in the number of empty predictions.

Attribute-conditional Mondrian conformal transducers

The conditionality principle (Cox 1958b; Cox and Hinkley 1974, §2.3) is often illustrated using the following simple example (slightly modified) due to Cox (1958b). Suppose we have two instruments for measuring an unknown bit; at each trial one instrument is used once, and the instrument to use is chosen at random (tossing a fair coin). Instrument 1 is more accurate, with the probability of mistake equal to 1%, whereas the probability of mistake for instrument 2 is 5%. Formally, each object is a pair $x = (i, b)$, where $i \in \{1, 2\}$ is the instrument used and $b \in \{0, 1\}$ is the result of the measurement; the label $y \in \{0, 1\}$ is the true bit.

It is intuitively clear that at confidence level 99.5% the optimal valid confidence predictor (cf. the description of the Bayes confidence predictor in §3.4) will predict objects $(1, \dots)$ with singular predictions and will not predict objects $(2, \dots)$ at all (in the sense that its predictions will be the set $\{0, 1\}$ of

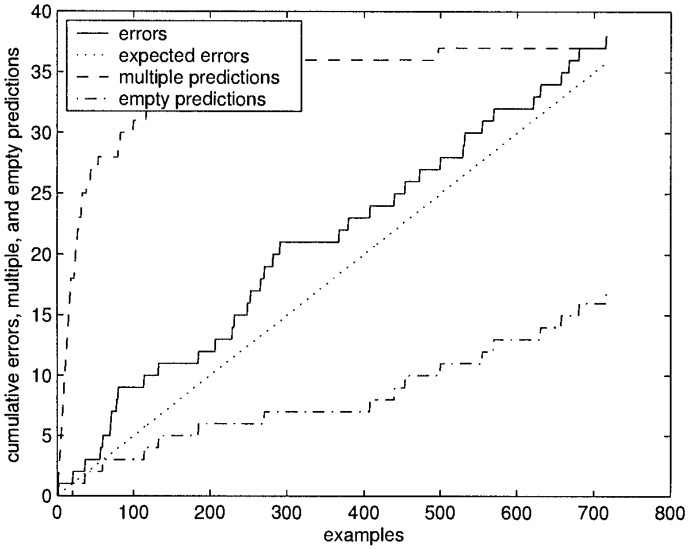

Fig. 4.9. The performance of the label-conditional MCP (based on the taxonomy $\kappa(n,(x,y)) = y$) on the USPS data set for the examples labeled as "5" at the 95% confidence level

all labels). At confidence level 97% the optimal valid confidence predictor will asymptotically predict all objects with singular predictions.

In both cases conditional validity is problematic (as argued by Cox); it does not prevent, however, the predictions from being valid on average. But the situation becomes even worse if we want to have two different significance levels for objects $(1,\dots)$ and $(2,\dots)$: if we take 0.5% for $(1,\dots)$ and 3% for $(2,\dots)$, any validity is lost.

The taxonomy for Cox's example is shown in Fig. 4.10, where "Instrument 1" stands for the set of examples $((1,\dots),\dots)$ and "Instrument 2" stands for the set of examples $((2,\dots),\dots)$.

In our experiments with different data sets we have not seen as gross failures in the conformal predictor's attribute-wise validity as those in the label-wise validity. The USPS data set does not have any natural attributes to condition on, since all attributes in it are of the same nature (the brightness level of a pixel) and continuous, but even for the data sets that do have natural attributes to condition on the conformal predictor's conditional performance was reasonable.

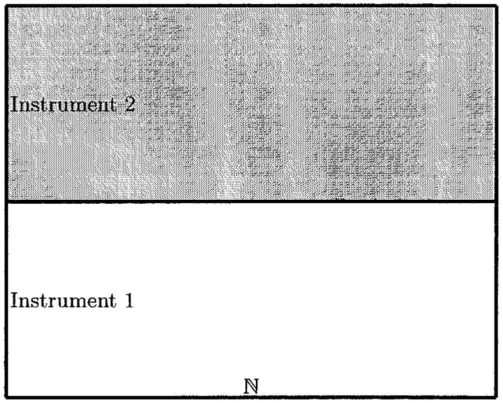

Fig. 4.10. Cox's example

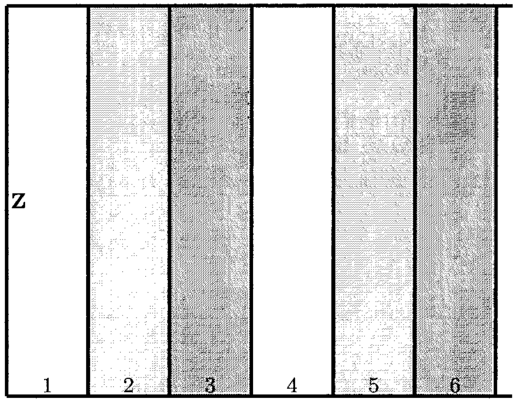

Fig. 4.11. Slow teacher

Slow teacher

It is interesting that MCPs can be used to deal with the problem of slow teacher considered in §4.3 above. The delay l is assumed to be constant. Define $\kappa(n,z) := n \bmod (l+1)$ (this is illustrated in Fig. 4.11 for $l = 2$) and take a nonconformity measure A_n (see (4.30) on p. 115) that depends on its arguments only via $\lbag z_j : j \in \{1,\dots,i-1,i+1,\dots,n\}\ \&\ \kappa_j = \kappa_n \rbag$ and (κ_n, z_i). The corresponding smoothed MCP only needs a slow teacher with lag l, and Proposition 4.10 implies that it is not only asymptotically valid, but is valid in the sense that its errors are independent Bernoulli random variables with the right parameter. Of course, this predictor can only be used where there is a surfeit of examples.

4.6 Proofs

Proof of Theorem 4.2, I: $n_k/n_{k-1} \to 1$ is sufficient

We start from a simple general lemma about martingale differences.

Lemma 4.12. *If $\xi_1, \xi_2, \dots$ is a martingale difference w.r. to σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \dots$ and $w_1, w_2, \dots$ is a sequence of positive numbers such that, for all $i = 1, 2, \dots$,*

$$\mathbb{E}(\xi_i^2 \mid \mathcal{F}_{i-1}) \le w_i^2 \,,$$

then

$$\mathbb{E}\left(\left(\frac{\xi_1 + \dots + \xi_n}{w_1 + \dots + w_n}\right)^2\right) \le \frac{w_1^2 + \dots + w_n^2}{(w_1 + \dots + w_n)^2} \,.$$

Proof. Since elements of a martingale difference sequence are uncorrelated, we have

$$\mathbb{E}\left((\xi_1 + \dots + \xi_n)^2\right) = \sum_{1 \le i \le n} \mathbb{E}(\xi_i^2) + 2 \sum_{1 \le i < j \le n} \mathbb{E}(\xi_i \xi_j) \le \sum_{1 \le i \le n} w_i^2 \,. \qquad \square$$

Fix a significance level ϵ and a power probability distribution Q^∞ on $\mathbf{Z}^\infty$ generating the examples $z_i = (x_i, y_i)$; the $\mathcal{L}$-taught smooth conformal predictor $\Gamma^{\mathcal{L}}$ is fed with the examples z_i and random numbers $\tau_i \in [0,1]$. The error sequence and predictable error sequence of $\Gamma^{\mathcal{L}}$ will be denoted

$$e_n := \mathrm{err}_n^\epsilon(\Gamma^{\mathcal{L}}) = \begin{cases} 1 & \text{if } y_n \notin \Gamma^{\mathcal{L},\epsilon}(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) \\ 0 & \text{otherwise} \end{cases}$$

and

$$d_n := \overline{\mathrm{err}}_n^\epsilon(\Gamma^{\mathcal{L}}) = (Q \times \mathbf{U})\Big\{(x, y, \tau) \in \mathbf{Z} \times [0,1] : \\ y \notin \Gamma^{\mathcal{L},\epsilon}(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)\Big\} \,.$$

Along with the original predictor $\Gamma^{\mathcal{L}}$ we also consider the *ghost predictor*, which is Γ fed with the examples

$$z'_1 = (x'_1, y'_1) := z_{\mathcal{L}(n_1)}, z'_2 = (x'_2, y'_2) := z_{\mathcal{L}(n_2)}, \dots$$

and random numbers $\tau'_1, \tau'_2, \dots$ (independent from each other and from the sequences z_i and τ_i). The ghost predictor is given all labels and each label is given without delay. Notice that its input sequence $z_{\mathcal{L}(n_1)}, z_{\mathcal{L}(n_2)}, \dots$ is also distributed according to Q^∞. The error and predictable error sequences of the ghost predictor are

$$e'_n := \mathrm{err}_n^\epsilon(\Gamma, (z'_1, z'_2, \dots)) \\ = \begin{cases} 1 & \text{if } y'_n \notin \Gamma^\epsilon(x'_1, \tau'_1, y'_1, \dots, x'_{n-1}, \tau'_{n-1}, y'_{n-1}, x'_n, \tau'_n) \\ 0 & \text{otherwise} \end{cases}$$

and

$$d'_n := \overline{\mathrm{err}}_n^\epsilon(\Gamma, (z'_1, z'_2, \dots)) = (Q \times \mathbf{U})\Big\{(x, y, \tau) \in \mathbf{Z} \times [0,1] : \\ y \notin \Gamma^\epsilon(x'_1, \tau'_1, y'_1, \dots, x'_{n-1}, \tau'_{n-1}, y'_{n-1}, x, \tau)\Big\} \,.$$

It is clear that, for each k, d_n is the same for all $n = n_{k-1} + 1, \dots, n_k$, their common value being

$$d_{n_k} = d'_k \,. \tag{4.31}$$

Corollary 4.13. *For each k,*

$$\mathbb{E}\left(\left(\frac{(e'_1 - \epsilon)n_1 + (e'_2 - \epsilon)(n_2 - n_1) + \dots + (e'_k - \epsilon)(n_k - n_{k-1})}{n_k}\right)^2\right) \\ \le \frac{n_1^2 + (n_2 - n_1)^2 + \dots + (n_k - n_{k-1})^2}{n_k^2} \,.$$

Proof. It is sufficient to apply Lemma 4.12 to $w_i := n_i - n_{i-1}$ (n_0 is understood to be 0 in this section), the independent zero-mean (by Proposition 2.4 on p. 27) random variables $\xi_i := (e'_i - \epsilon)w_i$, and the σ-algebras $\mathcal{F}_i$ generated by $\xi_1, \dots, \xi_i$. □

Corollary 4.14. *For each k,*

$$\mathbb{E}\left(\left(\frac{(e'_1 - d'_1)n_1 + (e'_2 - d'_2)(n_2 - n_1) + \dots + (e'_k - d'_k)(n_k - n_{k-1})}{n_k}\right)^2\right) \\ \le \frac{n_1^2 + (n_2 - n_1)^2 + \dots + (n_k - n_{k-1})^2}{n_k^2} \,.$$

Proof. Use Lemma 4.12 for $w_i := n_i - n_{i-1}$, $\xi_i := (e'_i - d'_i)w_i$, and the σ-algebras $\mathcal{F}_i$ generated by $z'_1, \dots, z'_i$ and $\tau'_1, \dots, \tau'_i$. □

Corollary 4.15. *For each k,*

$$\mathbb{E}\left(\frac{(e_1 - d_1) + (e_2 - d_2) + \dots + (e_{n_k} - d_{n_k})}{n_k}\right)^2 \le \frac{1}{n_k} \,.$$

Proof. Apply Lemma 4.12 to $w_i := 1$, $\xi_i := e_i - d_i$, and the σ-algebras $\mathcal{F}_i$ generated by $z_1, \dots, z_i$ and $\tau_1, \dots, \tau_i$. □

Lemma 4.16. *If* $\lim_{k\to\infty}(n_k/n_{k-1}) = 1$ *for some strictly increasing sequence of positive integers* $n_1, n_2, \ldots$, *then*

$$\lim_{k\to\infty} \frac{n_1^2 + (n_2 - n_1)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2} = 0 .$$

Proof. For any $\delta > 0$, there exists a K such that $\frac{n_k - n_{k-1}}{n_{k-1}} < \delta$ for any $k > K$. Therefore,

$$\begin{aligned}
&\frac{n_1^2 + (n_2 - n_1)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2} \\
&\le \frac{n_K^2}{n_k^2} + \frac{(n_{K+1} - n_K)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2} \\
&\le \frac{n_K^2}{n_k^2} + \frac{n_{K+1} - n_K}{n_K}\frac{n_{K+1} - n_K}{n_k} + \frac{n_{K+2} - n_{K+1}}{n_{K+1}}\frac{n_{K+2} - n_{K+1}}{n_k} + \cdots \\
&+ \frac{n_k - n_{k-1}}{n_{k-1}}\frac{n_k - n_{k-1}}{n_k} \le \frac{n_K^2}{n_k^2} + \delta\frac{(n_{K+1} - n_K) + \cdots + (n_k - n_{k-1})}{n_k} \le 2\delta
\end{aligned}$$

from some k on. □

Now it is easy to finish the proof of the first part of the theorem. In combination with Chebyshev's inequality and Lemma 4.16, Corollary 4.13 implies that

$$\frac{(e'_1 - \epsilon)n_1 + (e'_2 - \epsilon)(n_2 - n_1) + \cdots + (e'_k - \epsilon)(n_k - n_{k-1})}{n_k} \to 0$$

in probability; using the notation $k(i) := \min\{k : n_k \ge i\} = s(i) + 1$, we can rewrite this as

$$\frac{1}{n_k}\sum_{i=1}^{n_k}\left(e'_{k(i)} - \epsilon\right) \to 0 . \qquad (4.32)$$

Similarly, (4.31) and Corollary 4.14 imply

$$\frac{1}{n_k}\sum_{i=1}^{n_k}\left(e'_{k(i)} - d'_{k(i)}\right) = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(e'_{k(i)} - d_i\right) \to 0 , \qquad (4.33)$$

and Corollary 4.15 implies

$$\frac{1}{n_k}\sum_{i=1}^{n_k}(e_i - d_i) \to 0 \qquad (4.34)$$

(all convergences are in probability). Combining (4.32)–(4.34), we obtain

$$\frac{1}{n_k}\sum_{i=1}^{n_k}(e_i - \epsilon) \to 0 ; \qquad (4.35)$$

the condition $n_k/n_{k-1} \to 1$ allows us to replace n_k with n in (4.35).

Proof of Theorem 4.2, II: $n_k/n_{k-1} \to 1$ is necessary

Fix $\epsilon := 5\%$. As a first step, we construct the example space $\mathbf{Z}$, the probability distribution Q on $\mathbf{Z}$ and a smoothed conformal predictor for which d'_k deviate consistently from ϵ. Let $\mathbf{X} = \{0\}$, $\mathbf{Y} = \{0, 1\}$, so that z_i is, essentially, always 0 or 1. The probability distribution Q is uniform on $\mathbf{Z}$: $Q\{0\} = Q\{1\} = 1/2$. The nonconformity measure is

$$\alpha_i = A\left(\Lbag\zeta_1, \ldots, \zeta_{i-1}, \zeta_{i+1}, \ldots, \zeta_k\Rbag, \zeta_i\right) := \begin{cases} \zeta_i & \text{if } \zeta_1 + \cdots + \zeta_k \text{ is even} \\ 1 - \zeta_i & \text{if } \zeta_1 + \cdots + \zeta_k \text{ is odd} . \end{cases}$$

It follows from the central limit theorem that

$$\frac{|\{i = 1, \ldots, k : z'_i = 1\}|}{k} \in (0.4, 0.6) \qquad (4.36)$$

with probability at least 99% for k large enough. We will show that d'_k deviates significantly from ϵ with probability at least 99% for sufficiently large k. Let $\alpha_i := A(\Lbag z'_1, \ldots, z'_{i-1}, z'_{i+1}, \ldots, z'_k\Rbag, z'_i)$ with z'_k is interpreted as y (corresponding to (x, y) in the previous subsection). There are two possibilities:

- If $z'_1 + \cdots + z'_{k-1}$ is odd, then

  $$\begin{aligned}
  z'_k = 1 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is even} \Longrightarrow \alpha_k = z'_k = 1 \\
  z'_k = 0 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is odd} \Longrightarrow \alpha_k = 1 - z'_k = 1 .
  \end{aligned}$$

  In both cases we have $\alpha_k = 1$ and, therefore, outside an event of probability at most 1%,

  $$\begin{aligned}
  d'_k &= (Q \times \mathbf{U})\left\{(y, \tau) : \tau\,|\{i = 1, \ldots, k : \alpha_i = 1\}| \le k\epsilon\right\} \\
  &= \int_{\mathbf{Y}} 1 \wedge \frac{k\epsilon}{|\{i = 1, \ldots, k : \alpha_i = 1\}|} Q(\mathrm{d}y) \ge \frac{k\epsilon}{0.7k} = \frac{10}{7}\epsilon .
  \end{aligned}$$

- If $z'_1 + \cdots + z'_{k-1}$ is even, then

  $$\begin{aligned}
  z'_k = 1 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is odd} \Longrightarrow \alpha_k = 1 - z'_k = 0 \\
  z'_k = 0 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is even} \Longrightarrow \alpha_k = z'_k = 0 .
  \end{aligned}$$

  In both cases $\alpha_k = 0$ and, therefore, outside an even of probability at most 1%,

  $$\begin{aligned}
  d'_k = (Q \times \mathbf{U})\Big\{&(y, \tau) : \\
  &|\{i = 1, \ldots, k : \alpha_i = 1\}| + \tau\,|\{i = 1, \ldots, k : \alpha_i = 0\}| \le k\epsilon\Big\} \\
  &\le (Q \times \mathbf{U})\left\{(y, \tau) : 0.3k \le k\epsilon\right\} = 0 .
  \end{aligned}$$

To summarize, for large enough k,

$$|d'_k - \epsilon| = |d_{n_k} - \epsilon| > \epsilon/3 \tag{4.37}$$

with probability at least 99% (cf. (4.31); we write 99% rather than 98% since the two exceptional events of probability 1% coincide: both are the complement of (4.36)).

Suppose that

$$\frac{1}{n}\sum_{i=1}^{n} e_i \to \epsilon \tag{4.38}$$

in probability; we will deduce that $n_k/n_{k-1} \to 1$. By (4.34) (remember that Corollary 4.15 and, therefore, (4.34) do not depend on the condition $n_k/n_{k-1} \to 1$) and (4.38) we have

$$\frac{1}{n_k}\sum_{i=1}^{n_k} d_i \to \epsilon\,;$$

we can rewrite this in the form

$$\sum_{i=1}^{n_k} d_i = n_k(\epsilon + o(1))$$

(all $o(1)$ are in probability). This equality implies

$$\sum_{k=0}^{K} d_{n_k}\,(n_k - n_{k-1}) = n_K(\epsilon + o(1))$$

and

$$\sum_{k=0}^{K-1} d_{n_k}\,(n_k - n_{k-1}) = n_{K-1}(\epsilon + o(1))\,;$$

subtracting the last equality from the penultimate one we obtain

$$d_{n_K}\,(n_K - n_{K-1}) = (n_K - n_{K-1})\epsilon + o(n_K)\,,$$

i.e.,

$$(d_{n_K} - \epsilon)\,(n_K - n_{K-1}) = o(n_K)\,.$$

In combination with (4.37), this implies $n_K - n_{K-1} = o(n_K)$, i.e., $n_K/n_{K-1} \to 1$ as $K \to \infty$.

Proof of Theorem 4.4

This proof is similar to the proof of Theorem 4.2. (The definition of $\Gamma^{\mathcal{L}}$ being asymptotically exact involves the assumption of exchangeability rather than randomness; however, since we assumed that $\mathbf{Z}$ is a Borel space, de Finetti's theorem, stated in §A.5, shows that these assumptions are equivalent in our current context.) Instead of Corollaries 4.13, 4.14, and 4.15 we now have:

Corollary 4.17. *As $k \to \infty$,*

$$\frac{(e'_1 - \epsilon)n_1 + (e'_2 - \epsilon)(n_2 - n_1) + \cdots + (e'_k - \epsilon)(n_k - n_{k-1})}{n_k} \to 0 \quad a.s.$$

Proof. It is sufficient to apply Kolmogorov's strong law of large numbers (stated in §A.6) to the independent zero-mean random variables $\xi_i = (e'_i - \epsilon)(n_i - n_{i-1})$. Condition (A.8) (p. 286) follows from

$$\sum_{i=1}^{\infty} \frac{(n_i - n_{i-1})^2}{n_i^2} < \infty\,,$$

which is equivalent to (4.18). □

Corollary 4.18. *As $k \to \infty$,*

$$\frac{(e'_1 - d'_1)n_1 + (e'_2 - d'_2)(n_2 - n_1) + \cdots + (e'_k - d'_k)(n_k - n_{k-1})}{n_k} \to 0 \quad a.s.$$

Proof. Apply the martingale strong law of large numbers (§A.6) to the martingale difference $\xi_i = (e'_i - d'_i)(n_i - n_{i-1})$ w.r. to the σ-algebras $\mathcal{F}_i$ generated by $z'_1, \ldots, z'_i$ and $\tau'_1, \ldots, \tau'_i$. □

Corollary 4.19. *As $k \to \infty$,*

$$\frac{(e_1 - d_1) + (e_2 - d_2) + \cdots + (e_{n_k} - d_{n_k})}{n_k} \to 0 \quad a.s.$$

Proof. Apply the martingale strong law of large numbers to the martingale difference $\xi_i = e_i - d_i$ w.r. to the σ-algebras $\mathcal{F}_i$ generated by $z_1, \ldots, z_i$ and $\tau_1, \ldots, \tau_i$. □

Corollary 4.17 can be rewritten as (4.32), Corollary 4.18 as (4.33), and Corollary 4.19 as (4.34); all convergences are now almost certain. Combining (4.32)–(4.34), we obtain (4.35). It remains to replace n_k with n, as before.

Proof of Theorem 4.8

We will only consider the case where Γ is a smoothed conformal predictor (the proof for deterministic Γ is almost identical: just ignore all random numbers τ).

Fix a significance level ϵ and define

$$\overline{\mathrm{mult}}_n(\Gamma^{\mathcal{L}}) = (Q \times \mathbf{U})\Big\{(x,\tau) \in \mathbf{X} \times [0,1] : \\ \big|\Gamma^{\mathcal{L},\epsilon}(x_1, \tau_1, y_1, \ldots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)\big| > 1\Big\}\,,$$

$$\overline{\mathrm{mult}}_k(\mathit{\Gamma}) = (Q \times \mathbf{U})\Big\{(x,\tau) \in \mathbf{X} \times [0,1] : \\ \big|\mathit{\Gamma}^\epsilon(x'_1,\tau'_1,y'_1,\dots,x'_{k-1},\tau'_{k-1},y'_{k-1},x,\tau)\big| > 1\Big\} ,$$

$$\overline{\mathrm{Mult}}_n(\mathit{\Gamma}^{\mathcal{L}}) = \sum_{i=1}^{n} \overline{\mathrm{mult}}_i(\mathit{\Gamma}^{\mathcal{L}}) , \quad \overline{\mathrm{Mult}}_k(\mathit{\Gamma}) = \sum_{i=1}^{k} \overline{\mathrm{mult}}_i(\mathit{\Gamma}) .$$

Since $\mathrm{Mult}^\epsilon_n(\mathit{\Gamma}^{\mathcal{L}}) - \overline{\mathrm{Mult}}_n(\mathit{\Gamma}^{\mathcal{L}})$ is a martingale and

$$|\,\mathrm{mult}^\epsilon_n(\mathit{\Gamma}^{\mathcal{L}}) - \overline{\mathrm{mult}}_n(\mathit{\Gamma}^{\mathcal{L}})| \le 1 ,$$

the martingale strong law of large numbers (see §A.6) implies that

$$\lim_{n\to\infty} \frac{\mathrm{Mult}^\epsilon_n(\mathit{\Gamma}^{\mathcal{L}}) - \overline{\mathrm{Mult}}_n(\mathit{\Gamma}^{\mathcal{L}})}{n} = 0 \quad \text{a.s.} \tag{4.39}$$

Analogously,

$$\lim_{k\to\infty} \frac{\mathrm{Mult}^\epsilon_k(\mathit{\Gamma},(z'_1,z'_2,\dots)) - \overline{\mathrm{Mult}}_k(\mathit{\Gamma})}{k} = 0 \quad \text{a.s.} \tag{4.40}$$

By (4.39) and (4.40), we can replace Mult^ϵ with $\overline{\mathrm{Mult}}$ in the definitions of $U^\epsilon(\mathit{\Gamma}^{\mathcal{L}},Q)$ and $U^\epsilon(\mathit{\Gamma},Q)$.

It is clear that

$$\overline{\mathrm{mult}}_n(\mathit{\Gamma}^{\mathcal{L}}) = \overline{\mathrm{mult}}_{k(n)}(\mathit{\Gamma})$$

for all n. Combining this with $k(n) = n/c + O(1)$, we obtain

$$\sum_{i=1}^{n} \overline{\mathrm{mult}}_i(\mathit{\Gamma}^{\mathcal{L}}) = c \sum_{i=1}^{\lfloor n/c \rfloor} \overline{\mathrm{mult}}_i(\mathit{\Gamma}) + O(1) ,$$

and so $\overline{\mathrm{Mult}}_n(\mathit{\Gamma}^{\mathcal{L}}) = c\,\overline{\mathrm{Mult}}_{\lfloor n/c \rfloor}(\mathit{\Gamma}) + o(n)$. The statement of the theorem immediately follows.

4.7 Bibliographical remarks

Computationally efficient hedged prediction

To cope with the relative computational inefficiency of conformal predictors, inductive conformal predictors were introduced in Papadopoulos et al. 2002a and Papadopoulos et al. 2002b in the off-line setting and in Vovk 2002b in the on-line setting. Before the appearance of inductive conformal predictors, several other possibilities had been studied, such as "competitive transduction" (Saunders 2000) and "transduction with hashing" (Saunders et al. 2000; Saunders 2000).

Specific learning algorithms and nonconformity measures

Equation (4.11) is sometimes known as the Sherman–Morrison formula (it can be checked easily by multiplying the right-hand side by $K + uv'$ on the left and simplifying); for details, see Henderson and Searle 1981.

The bootstrap was proposed by Efron (1979). For recent reviews, see Efron 2003 and other articles in the same issue of *Statistical Science.* There are two main varieties of regression bootstrap: "bootsrapping residuals" and "bootsrapping cases". (For details, see Montgomery et al. 2001, pp. 509–510, or Draper and Smith 1998, pp. 285–286.) We only gave an example of using the first of these procedures, following Davison and Hinkley 1997 (Algorithm 6.4), the original idea being due to Stine (1985).

Decision trees are reviewed, besides Mitchell 1997 (Chap. 3), in Ripley 1996 (Chap. 7); the latter contains many pointers to the relevant literature. The C4.5 algorithm was introduced in Quinlan 1993.

In our description of hedged prediction based on boosting we followed Proedrou 2003; both definitions (4.14) and (4.15) of conformity scores are due to him. The first boosting algorithm was proposed by Schapire (1990); AdaBoost is due to Freund and Schapire (1997).

Neural networks are popular in both classification and regression; good references are Bishop 1995 and Ripley 1996. The current wave of interest was mainly initiated by Rumelhart and McClelland (1986).

Cox 1970 and more recent Hosmer and Lemeshow 2000 are useful sources for logistic regression. Cox 1958a may be the first publication describing logistic regression, although Jerome Cornfield might have used it several years before 1958 (Reid 1994, p. 448).

In this book we have given examples of nonconformity measures based on least squares, ridge regression, logistic regression, nearest neighbors, support vector machines, decision trees, boosting, bootstrap, and neural networks. The number of known machine learning algorithms is huge, however, and potentially any of them can be used as a source of nonconformity measures; in particular, we did not touch the important class of genetic algorithms (see, e.g., Mitchell 1996). For a very readable introduction to machine learning algorithms, see Mitchell 1997, and for recent developments see the proceedings of the numerous machine learning conferences, such as NIPS, ICML, UAI, COLT, and ALT.

Weak teachers

The characterization of lazy teachers for which conformal prediction is valid in probability was obtained by Ilia Nouretdinov. The general notion of a teaching schedule and the device of a "ghost predictor" is due to Daniil Ryabko (Ryabko et al. 2003). Nouretdinov's result (our Theorem 4.2), generalized in light of Ryabko et al. 2003, appeared in Nouretdinov and Vovk 2003. Theorems 4.4 and 4.8 are from Ryabko et al. 2003.

Mondrian conformal predictors

For further information, see Vovk et al. 2003a.