Pattern recognition and density estimation under the general i.i.d. assumption

Ilia Nouretdinov Volodya Vovk Michael Vyugin
Alex Gammerman

Technical Report

CLRC-TR-01-06

June 2001

Royal Holloway University of London

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

Abstract

Statistical learning theory considers three main problems, pattern recognition, regression and density estimation. When analyzing learning algorithms, the usual assumption made is that the data is generated by an i.i.d. source; sometimes, however, extra assumptions are also made. This paper studies solvability of these three main problems (mainly concentrating on pattern recognition and density estimation) in the "high-dimensional" case, where the patterns in the training and test sets are never repeated. We show that, assuming an i.i.d. data source but without any further assumptions, the problems of pattern recognition and regression can often be solved (and there are practically useful algorithms to solve them). On the other hand, the problem of density estimation, as we formalize it, cannot be solved under the general i.i.d. assumption, and additional assumptions are required.

1 Introduction

We consider the following special case of the standard problem of machine learning: given a training set

$$\Pi = ((x_1, y_1), \ldots, (x_l, y_l)) \tag{1}$$

consisting of labelled instances (x_i, y_i) and given a test unlabelled instance x_{l+1}, predict x_{l+1}'s label y_{l+1}. (This is a special case because we assume that there is only one test instance.) It is assumed that every labelled instance (x_i, y_i), $i = 1, \ldots, l+1$, is generated by some unknown probability distribution P (one and the same for all instances) independently of other instances; we will call this the *i.i.d. assumption* (and we will say *general i.i.d. assumption* to emphasize that no other assumptions are made). Sometimes we will say that x_i are *patterns*; a typical example is where x_i are hand-written digits and y_i are their classifications, $y_i \in \{0, 1, \ldots, 9\}$. Let $\mathbf{X}$ be the set of all possible patterns and $\mathbf{Y}$ be the set of all possible labels; therefore, P is a probability distribution in $\mathbf{X} \times \mathbf{Y}$.

According to Vapnik [10] there are three main problems of statistical learning theory:

- *pattern recognition* (or *classification*), in which the label space $\mathbf{Y}$ is finite (and usually quite small, such as $\{0, 1\}$);
- *regression*, where $\mathbf{Y}$ is the real line;
- *density estimation*, where the task is not just to predict y_{l+1} but to give a probability distribution for different values of y_{l+1}.

The goal of this paper is to draw a first rough boundary between what can and what cannot be done under the general i.i.d. assumption. It turns out that the first two problems can be solved (this is well known) but the third problem cannot (in non-trivial cases).

As usual, the positive results (solvability of pattern recognition and regression problems) do not require any general theory; however, to state the negative result (impossibility of density estimation) in the most natural way we will need the algorithmic notion of randomness (for details, see [5] and Section 2)[1]. The algorithmic theory of randomness asserts that there is a universal statistical test for checking whether a given data sequence can be generated by a given probability distribution; and a sequence $(z_1, \ldots, z_n)$ (in the case that interests us every z_i is a labelled instance, $z_i = (x_i, y_i) \in \mathbf{Z} := \mathbf{X} \times \mathbf{Y}$) is said to be random w.r. to a probability distribution P if the universal test does not detect disagreement between $(z_1, \ldots, z_n)$ and P. We say that a sequence is i.i.d. random (or just random) if it is random w.r. to some product distribution P^n. More accurately, for finite sequences ω the crisp notion of randomness has to be replaced by a function $d_P(\omega)$, called randomness level, measuring the degree to which ω is random w.r. to P: $d_P(\omega) = 1$ means complete randomness and $d_P(\omega) = 0$ means complete lack of randomness; here P is either a probability distribution in $\mathbf{Z}^n$ (where n is the length of ω) or a probability distribution in $\mathbf{Z}$ (in which case $d_P(\omega)$ is defined to be the same thing as $d_{P^n}(\omega)$). These definitions are also applicable to "partial" data sequences such as (Π, x_{l+1}) (where the classification of the last instance is not given).

Let X and Y be random variables taking values in $\mathbf{X}$ and $\mathbf{Y}$ respectively such that (X, Y) is distributed as P. If $\mathbf{Y}$ is finite, we will often use $P(y \mid x)$, where $x \in \mathbf{X}$ and $y \in \mathbf{Y}$, as the shorthand for the probability that $Y = y$ given $X = x$.

In this paper we are mainly interested in the case $\mathbf{Y} = \{0, 1\}$. Our informal statement of the problems of pattern recognition and density estimation is as follows. Suppose the training set Π is i.i.d. random; consider all probability distributions P in $\mathbf{Z}$ such that the training set Π is δ-random w.r. to P:

$$\mathrm{Rand}(\Pi, x_{l+1}, \delta) := \{P : d_P(\Pi, x_{l+1}) > \delta\}\,. \tag{2}$$

Typically δ is a conventional threshold, such as 1%, and $\mathrm{Rand}(\Pi, x_{l+1}, \delta)$ is the set of all probability distributions that can generate Π and x_{l+1} (all other distributions are rejected by the universal statistical test at the level δ). Consider the interval

$$[a, b] := [\inf P\,(1 \mid x_{l+1})\,, \sup P\,(1 \mid x_{l+1})]\,, \tag{3}$$

where P ranges over $\mathrm{Rand}(\Pi, x_{l+1}, \delta)$. The problem of pattern recognition is feasible for the training set Π, test pattern x_{l+1} and significance level δ if

- either b is close to 0, in which case we can confidently predict $y_{l+1} = 0$ (all probability distributions P compatible with the training set agree that $P(1 \mid x_{l+1})$ is close to 0),
- or a is close to 1, in which case we can confidently predict $y_{l+1} = 1$ (all probability distributions P compatible with the training set agree that $P(1 \mid x_{l+1})$ is close to 1).

[1] In principle, it is possible to understand the main result of this paper without algorithmic randomness: the reader can proceed directly to Corollary 15 in Appendix B, and read only the definitions of notions used there.

The problem of density estimation is feasible if a and b are close to each other, i.e., all probability distributions P compatible with the training set have similar $P(1 \mid x_{l+1})$.

Remark 1 *Our usage of "density estimation" might look strange to some readers, since the term "density" is usually applied to densities w.r. to the Lebesgue measure. We consider, instead, densities with respect to counting measures, such as the counting measure μ on the set $\{0,1\}$ (recall that $\mu\{0\} = \mu\{1\} = 1$). This is sufficient for our purpose of demonstrating impossibility of density estimation: it is clear that we cannot hope to be able to estimate densities w.r. to the Lebesgue measure if we cannot do the much simpler task of estimating densities w.r. to counting measures.*

In this paper we are interested in the "high-dimensional" case, where, in particular, the same pattern never occurs twice among the given patterns $x_1, \ldots, x_{l+1}$ (for example, in a data set of gray-scale 16×16 handwritten digits, we do not expect two images to be *precise* copies of each other). Of course, the training set will usually contain patterns similar, in some sense, to the test pattern, and a natural idea would be to use these similar patterns to estimate the probability that the classification of the test pattern is 1. However, since i.i.d. is our only assumption, similarity of patterns does not imply at all that classifications are likely to coincide or be connected in any way. It might even seem surprising that pattern recognition is possible for real-world high-dimensional data sets under the general i.i.d. assumption. This is, however, demonstrated in [8], and can also be seen from older results of PAC theory, such as Littlestone and Warmuth's Theorem 4.25 (and its corollary Theorem 6.8) in [2]. The latter asserts that

$$b \le \frac{1}{l-d}\left(d \ln \frac{el}{d} + \ln \frac{l}{\delta}\right) \tag{4}$$

if Support Vector Machine (see Vapnik [11]) predicts 0 with d support vectors, and

$$a \ge 1 - \frac{1}{l-d}\left(d \ln \frac{el}{d} + \ln \frac{l}{\delta}\right) \tag{5}$$

if SVM predicts 1 with d support vectors (this is true for a specific statistical test, but will be true for the universal test if δ is replaced with δ/c for some constant c). Littlestone and Warmuth's result can be generalized to regression problems as well: see, e.g., [6] and [2] (Theorems 4.26, 4.28 and 4.30). The results of [8] are stated in a different framework, but in Section 3 we prove the connections between the two frameworks.

Remark 2 *Our only use of the requirement that the data set is "high-dimensional" is that we do not expect repeated patterns in it. Actually the latter condition will often be satisfied in low-dimensional cases as well: e.g., if x_i are real numbers generated according to a continuous distribution. Therefore, in the rest of the paper we will use a more neutral expression "diverse data set".*

The main negative result of this paper is that, for a diverse data set and under the general i.i.d. assumption, density estimation is feasible only in the trivial case where already pattern recognition is feasible: if the interval $[a, b]$ is non-empty, a slight increase in δ will make it contain 0 or 1. (In the algorithmic theory of randomness, δ matters only to within a constant factor, so a slight increase corresponds to multiplying by a universal constant c.) Therefore, non-trivial density estimation, where $[a, b] \subseteq (0, 1)$, can only be possible for a narrow range of δ. (And analysis of the proof will show that one cannot hope to get $[a, b] \subseteq (0, 1)$ even for a narrow range of δ in non-pathological cases: see Corollary 15 on p. 18.) These results elaborate on a remark in [13].

Remark 3 *Since we consider diverse data sets, standard statistical techniques such as the Glivenko–Cantelli theorem (see, e.g., Vapnik [11], Theorem 1.1 on p. 42) are not applicable in this context. The Glivenko–Cantelli theorem considers the case where x_i are absent and y_i take real values; it is equivalent and will be more convenient for us to assume that all x_i are equal to some predefined "null" element of* $\mathbf{X}$*. Therefore, our main assumption, that all x_i's are different, is violated; the Glivenko–Cantelli situation is opposite: all x_i coincide. When all x_i coincide and* $\mathbf{Y}$ *is finite, density estimation is also simple: we can estimate $P\{Y = y \mid X = x_{l+1}\} = P\{Y = y\}$ by the relative frequency of the classification y in the training set (see, e.g., [3]). Extending the Glivenko–Cantelli theorem in the Vapnik–Chervonenkis direction (Vapnik [11], Theorem 4.5 on p. 149) does not help either: since we are interested in estimating conditional probability given $X = x_{l+1}$, the corresponding VC dimension will be infinite or too big to give non-trivial accuracy of estimation.*

2 Mathematical background

We will use the algorithmic notion of randomness only in the case where $\mathbf{Y} = \{0, 1\}$ and $\mathbf{X}$ is a finite set. Our statements are applied to different finite sets $\mathbf{X}$ which are supposed to be subsets of some large countable set $\mathbf{W}$ (big enough to contain all objects of interest to us; e.g., $\mathbf{W}$ might be the set of all finite sequences of ASCII symbols). We assume that $\mathbf{W}$ is a "constructive space", in the sense that all its elements have individual names. Each finite $\mathbf{X} \subset \mathbf{W}$ can be effectively given by listing all its elements; we assume that such a description of $\mathbf{X}$ is given as an input to all algorithms we are going to consider.

Let $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$; denote the set of all probability distributions in a measurable space A as $\mathcal{P}(A)$. Notice that $\mathcal{P}(A)$ is a finite-dimensional space when A is finite (e.g., when $A = \mathbf{Z}^n$). The following are standard definitions of the algorithmic theory of randomness. A real-valued function $f(u)$ is called *semi-computable from above* (resp. *from below*) if there exists a computable function $F(i, u)$, i ranging over the positive integers, such that:

- for all i and u, $F(i, u) \geq F(i + 1, u)$ (resp. $F(i, u) \leq F(i + 1, u)$),
- for all u, $f(u) = \lim_{i\to\infty} F(i, u)$.

A function $t : \cup_n(\mathbf{Z}^n \times \mathcal{P}(\mathbf{Z}^n)) \times \mathbf{W} \to [0,1]$ is called a *conditional statistical test* if, for any n, $P \in \mathcal{P}(\mathbf{Z}^n)$, $w \in \mathbf{W}$ and $\gamma \in [0,1]$:

$$P\{\omega = (z_1, \dots, z_n) : t(\omega, P, w) \le \gamma\} \le \gamma.$$

(It is conventional in the algorithmic theory of randomness to give this definition "on the log-scale", in terms of $-\log t$.) Similarly, a *statistical test* (unconditional) can be defined as a function $t : \cup_n(\mathbf{Z}^n \times \mathcal{P}(\mathbf{Z}^n)) \to [0,1]$ such that, for any n, $P \in \mathcal{P}(\mathbf{Z}^n)$ and $\gamma \in [0,1]$:

$$P\{\omega \in \mathbf{Z}^n : t(\omega, P) \le \gamma\} \le \gamma \tag{6}$$

(in principle, this notion is superfluous, since we can use conditional tests with a fixed condition).

A conditional statistical test d is called a *universal conditional statistical test* if it is semicomputable from above and, for any other semicomputable from above conditional statistical test t, there exists a constant $c > 0$ such that $d(\omega, P, w) \le ct(\omega, P, w)$ for any n, $\omega \in \mathbf{Z}^n$, $P \in \mathcal{P}(\mathbf{Z}^n)$ and $w \in \mathbf{W}$. A universal conditional test exists by a theorem due to Kolmogorov and Martin-Löf [15, 14]. We fix a universal conditional test d and call $d_P(\omega \mid w) := d(\omega, P, w)$ the *randomness level* of ω w.r. to P when w is known. We also fix some element $w_0 \in W$ (the "null" element) and call $d_P(\omega) := d_P(\omega \mid w_0)$ the (unconditional) *randomness level* of ω w.r. to P. We will usually give our definitions for unconditional level, although sometimes they will be used in the conditional case as well. Notation such as $d_{P^n}(\omega)$ will be sometimes simplified to $d_P(\omega)$; we will also write $t_P(\omega))$ for arbitrary statistical tests (not necessarily universal).

If $\mathcal{Q}$ is a family of probability distributions in $\mathbf{Z}^n$, we define the randomness level of $\omega \in \mathbf{Z}^n$ w.r. to $\mathcal{Q}$ as the supremum

$$d_{\mathcal{Q}}(\omega) := \sup_{P \in \mathcal{Q}} d_P(\omega)$$

(ω is considered random w.r. to a family if it is random w.r. to at least one distribution in the family). According to *Levin's theorem*, when $\mathcal{Q}$ ranges over (constructively) closed families (possibly with different n), $d_{\mathcal{Q}}(\omega)$ is semicomputable from above and is valid in the sense that always

$$\sup_{P \in \mathcal{Q}} P\left\{\omega : d_{\mathcal{Q}}(\omega) \le \gamma\right\} \le \gamma;$$

moreover, it is the maximal, up to a constant factor, function satisfying these properties.

If Π is the training set (1), x_{l+1} is a test pattern and $P \in \mathcal{P}(\mathbf{Z})$, we define the randomness level of Π extended by adding the test pattern as the supremum over all possible classifications of the test pattern:

$$d_P(\Pi, x_{l+1}) := \sup_{y \in \mathbf{Y}} d_P(\Pi, (x_{l+1}, y)).$$

Analogously to Levin's theorem, thus defined function d on the extended sequences in $\mathbf{Z}^l \times \mathbf{X}$, $l = 1, 2, \dots$, is, up to a constant factor, the maximal upper semicomputable function such that the probability of the event $d_P(\Pi, x_{l+1}) \le \gamma$ never exceeds γ, provided the data items (x_i, y_i) are generated from P independently. We will also be using the notation $d_{\mathcal{Q}}(\Pi, x_{l+1})$ for $\sup_{P\in\mathcal{Q}} d_P(\Pi, x_{l+1})$ (we refrain from restating this definition).

We will be especially interested in the randomness level with respect to the family $\mathcal{Q}$ of all probability distributions P^n in $\mathbf{Z}^n$, $P \in \mathcal{P}(\mathbf{Z})$; in this case we will write $d_{\mathrm{iid}}(\cdot)$ instead of $d_{\mathcal{Q}}(\cdot)$.

3 Pattern recognition

As we saw in the introductory section (see p. 3), the problem of pattern recognition is sometimes (when the number of support vectors is small) solvable in the diverse case. In principle, this answers the main question of this paper for pattern recognition, and so this section is to some degree an aside. Here we consider the question of practical applicability of this solution.

A worrying feature of PAC theory is the looseness of the bounds it establishes. Littlestone and Warmuth's bound stated in Section 1 is one of the tightest, but even it usually does not give practically meaningful results. E.g., for the US Postal Service database (described in [11], Section 12.2), the error bound for one out of ten classifiers is something like

$$\frac{1}{l-d}\left(d\ln\frac{el}{d}+\ln\frac{l}{\delta}\right) \approx \frac{1}{7291-274}274\ln\frac{7291e}{274} \approx 0.17,$$

even if we ignore the term $\ln\frac{l}{\delta}$ (274 is the average number of support vectors for polynomials of degree 3, which give the best predictive performance; see Table 12.2 in [11]). Since there are ten classifiers, the bound on the total probability of mistake becomes 1.7, which is not useful. Even if Littlestone and Warmuth's result were applicable to multi-class classifiers, we would still obtain

$$\frac{1}{l-d}\left(d\ln\frac{el}{d}+\ln\frac{l}{\delta}\right) \approx \frac{1}{7291-1677}1677\ln\frac{7291e}{1677} \approx 0.74$$

(1677 is the total number of support vectors for all ten classifiers for polynomial kernels; see [11], Table 12.5), not much better than the trivial bound 1. Other standard theorems known to us give even worse results (see, e.g., [2]).

Papers [4, 8, 13, 9, 6, 7] report practically meaningful results (see below) for pattern recognition and regression. To distinguish the approach of these papers from other approaches (such as PAC and Bayesian) we will call it the "randomness approach"; the attribute "randomness" will similarly be used to indicate that the concept belongs to the randomness approach. The randomness approach has two varieties of results: "universal" randomness results (involving the universal statistical test) and "specific" randomness results (involving specific non-universal statistical tests).

The specific randomness approach is very simple and in the spirit of the classical nonparametric statistics (for a discussion of simple parametric cases, see [1], Section 7.5), in contrast to the sophisticated PAC techniques. Its main idea, in the case of pattern recognition, is as follows: a prediction $\hat{y}_{l+1}$ for

y_{l+1} is considered to be reliable if the alternative classifications $y_{l+1} \neq \hat{y}_{l+1}$ are excluded by some statistical test, which is required to be valid (i.e., is required to satisfy (6)) under any i.i.d. distribution. (In the universal randomness approach, this statistical test is the universal test.) "Excluded" means that the statistical test takes value less than some threshold; we will take 1% as the threshold. As the validity condition (6) (with $\gamma = 0.01$) shows, it is possible that the true classification will be excluded, but the probability of this does not exceed 1%.

The recent paper [7] constructs a statistical test based on the Nearest Neighbours procedure which, for the USPS data set mentioned above, gives reliable predictions for 96% of the test examples (see Table 1 in [7]). In other words, in 96% of all test examples we can exclude all classifications but one at the significance level 1%; using the terminology introduced in [4, 8, 13], we can also say that in 96% of cases the confidence of our prediction is 99% or more. The use of a statistical test based on the Nearest Neighbours is not essential here; similar results can be obtained using, e.g., Support Vector Machine as the underlying algorithm.

We can see that the randomness framework used in [4, 8, 13, 9, 6, 7] is different from that of this paper (see Section 1). However, the following simple theorem shows that there is a close connection (we are again assuming $Y = \{0,1\}$): if $y_{l+1} = 1$ leads to a low randomness level, $P(1 \mid x_{l+1})$ is low, and vice versa (we state this only for the label 1; the statement for 0 is symmetrical).

Theorem 4 *et* Π *be a training set of size* l *and* x_{l+1} *be a test pattern. There exists a constant* $c > 0$ *such that:*

- *For any constant* $\epsilon > 0$ *and probability distribution* $P \in \mathcal{P}(\mathbf{Z})$*,*

$$d_P(\Pi, (x_{l+1}, 1) \mid \epsilon) \leq \epsilon \Longrightarrow P(1 \mid x_{l+1}) \leq \frac{c\epsilon}{d_P(\Pi, x_{l+1} \mid \epsilon)}. \tag{7}$$

- *For any constant* $\epsilon > 0$ *and probability distribution* $P \in \mathcal{P}(\mathbf{Z})$*,*

$$P(1 \mid x_{l+1}) \leq \epsilon \Longrightarrow d_P(\Pi, (x_{l+1}, 1) \mid \epsilon) \leq c\epsilon d_P(\Pi, x_{l+1} \mid \epsilon). \tag{8}$$

Proof E quation (7) can be rewritten as

$$d_P(\Pi, (x_{l+1}, 1) \mid \epsilon) \leq \epsilon \Longrightarrow d_P(\Pi, x_{l+1} \mid \epsilon) \leq c\epsilon / P(1 \mid x_{l+1});$$

therefore, it is sufficient to show that the function equal to $\epsilon / P(1 \mid x_{l+1})$ if $d_P(\Pi, (x_{l+1}, 1) \mid \epsilon) \leq \epsilon$ and to 1 otherwise is a statistical test. Its semicomputability from above obviously follows from semicomputability from above of $d_P(\Pi, (x_{l+1}, 1) \mid \epsilon)$; the validity follows from

$$P\left\{(\Pi, x_{l+1}) : d_P(\Pi, (x_{l+1}, 1) \mid \epsilon) \leq \epsilon \;\&\; \epsilon / P(1 \mid x_{l+1}) \leq \gamma\right\} \leq \gamma.$$

Equation (8) follows from the fact that the function equal to $\epsilon d_P(\Pi, x_{l+1} \mid \epsilon)$ if $P(1 \mid x_{l+1}) \leq \epsilon$ and $y_{l+1} = 1$ and equal to 1 otherwise is a statistical test in Z^{l+1}.

Theorem 5 *Let Π be a training set of size l and x_{l+1} be a test pattern. There exists a constant $c > 0$ such that for any constant $\epsilon > 0$, $j \in J$ and probability distribution $P \in \mathcal{P}(\mathbf{Z})$,*

$$d_P(\Pi, (x_{l+1}, j) \mid \epsilon) \leq \epsilon \Longrightarrow P(j \mid x_{l+1}) \leq \frac{c\epsilon}{d_P(\Pi, x_{l+1} \mid \epsilon)}. \tag{9}$$

Proof L et

$$J = \{j \mid d_P(\Pi, (x_{l+1}, j) \mid \epsilon) \leq \epsilon\};$$

then it is sufficient to prove that

$$P\{y_{l+1} \in J \mid x_{l+1}\} \leq \frac{c\epsilon}{d_P(\Pi, x_{l+1} \mid \epsilon)}.$$

As it is equivalent to

$$d_P(\Pi, x_{l+1} \mid \epsilon) \leq \frac{c\epsilon}{P\{y_{l+1} \in J \mid x_{l+1}\}},$$

it is enough to prove that

$$\frac{\epsilon}{P\{y_{l+1} \in J \mid x_{l+1}\}}$$

is a test. Indeed, it is semicomputable from above because

$$P\{y_{l+1} \in J \mid x_{l+1}\} = P\{d_P(\Pi, (x_{l+1}, y_{l+1}) \mid \epsilon) \leq \epsilon\}$$

is semicomputable from below.

It remains to prove the property

$$P\{\frac{\epsilon}{P\{y_{l+1} \in J \mid x_{l+1}\}} \leq \gamma\} \leq \gamma;$$

suppose it is not true.

In this case we have

$$P\{(\Pi, x_{l+1}) : P\{y_{l+1} \in J \mid x_{l+1}\} \geq \frac{\epsilon}{\gamma}\} > \gamma;$$

which implies

$$P\{d_P(\Pi, (x_{l+1}, y_{l+1})) \mid \epsilon) \leq \epsilon\} > \frac{\epsilon}{\gamma}\gamma = \epsilon$$

and leads to contradiction with the fact that d_P is a test.

Let us see what this theorem means in practice. First we need to restate the theorem in terms of specific tests. Its proof shows that for any statistical test t (we will take as t essentially the statistical test constructed in [7]) there exists a statistical test T such that we have the analogue of (9) with $c = 1$:

$$t_P(\Pi, (x_{l+1}, 1) \mid \epsilon) \leq \epsilon \Longrightarrow P(1 \mid x_{l+1}) \leq \frac{\epsilon}{T_P(\Pi, x_{l+1} \mid \epsilon)};$$

analogously,

$$t_P(\Pi, (x_{l+1}, 0) \mid \epsilon) \leq \epsilon \Longrightarrow P(0 \mid x_{l+1}) \leq \frac{\epsilon}{T_P(\Pi, x_{l+1} \mid \epsilon)}.$$

This can be easily extended to more than 2 labels. Let us take $\epsilon = 0.01$ and the test (which does not depend on P and ϵ at all) constructed in [7] as $t(\omega) = t_P(\omega \mid w)$. As we already mentioned, for the 96% of the USPS test examples we have either

$$t(\Pi, (x_{l+1}, j)) \leq 0.01$$

for all $j \in \{0, \ldots, 9\}$ but one. Then for any probability distribution satisfying

$$T_P(\Pi, x_{l+1} \mid 0.01) > 0.2$$

(this will hold with probability 80% or more if P is the true distribution generating the data) we will have

$$P(j \mid x_{l+1}) < 0.01/0.2 = 0.05$$

for all j but one. We can see that the P-probability of mistake will be bounded by 0.45; hardly a useful result, but still better than 1.7 found above.

Our interpretation of the calculation in the previous paragraph is that the (ϵ, δ) framework might be too ambitious for practical problems. We believe that reporting confidence and credibility (as defined in [8] and [13]) is much more informative.

Remark 6 *At the beginning of this section we discussed the weakness, from the practical point of view, of the error bound in Littlestone and Warmuth's theorem. In this remark we will briefly discuss similar weaknesses of other results of PAC theory and statistical learning theory. The confidence measures provided by the randomness approach can always be extracted from the given data set alone. This is also true about Littlestone and Warmuth's theorem and most results reported in [2], but not true about many other well-known results. For example, Theorem 10.7 in [11] (p. 415) gives the error bound*

$$ER(\alpha_l) \leq \frac{E \min\left(\mathcal{K}_{l+1}, \left(\frac{\mathcal{D}_{l+1}}{\rho_{l+1}}\right)^2\right)}{l+1};$$

this bound might be tighter than that in Littlestone and Warmuth's theorem, but there is no way to see from the data set (Π, x_{l+1}) *at hand what the expected value in the right-hand side is going to be. All other theorems in [11] and [12] giving error bounds for the Support Vector Machine are of this type. Another common problem preventing one from evaluating theoretical error bounds is using unspecified constants and imposing conditions on the underlying probability distribution that are impossible to verify given only the data set* (Π, x_{l+1}) *(see [2] for numerous examples).*

4 Density estimation

The following theorem shows that non-trivial estimates of $P(1 \mid x_{l+1})$ are impossible: there will always exist a probability distribution under which (Π, x_{l+1}) is as random as under any other distribution ($d_P(\Pi, x_{l+1}) \geq d_{\text{iid}}(\Pi, x_{l+1})/c$) and for which $P(1 \mid x_{l+1})$ is trivial (0 or 1).

Theorem 7 *There exists a constant $c > 0$ which satisfies the following. If Π is a training set without repeating patterns and x_{l+1} is a test pattern different from all training patterns, there exists a probability distribution $P \in \mathcal{P}(\mathbf{Z})$ such that*

$$d_P(\Pi, x_{l+1}) \geq d_{\text{iid}}(\Pi, x_{l+1})/c \ \&\ P(1 \mid x_{l+1}) \in \{0, 1\}.$$

Proof sketch Let Q be a probability distribution in $\mathbf{Z}$ such that (Π, x_{l+1}) is random w.r.t. to Q. We represent the distribution Q as a mixture of the distributions P which are "simple" for Q (in the sense that that the marginal distributions in $\mathbf{X}$ for P and Q coincide and $P(1 \mid x)$ is always 0 or 1). Intuitively, our construction corresponds to the following picture: to generate a sample $(x_1, y_1), \ldots, (x_{l+1}, y_{l+1})$ (assuming no two x_i coincide) according to the distribution Q, we can first choose randomly $f : X \to \{0, 1\}$ according to the conditional distributions of Q (i.e., we set $f(x) := 1$ with probability $Q(1 \mid x)$, independently for all $x \in X$), and then compute $y_i = f(x_i)$ from x_i generated from the marginal distribution of Q. With every such f we associate the distribution P with the same marginal as Q and with $P(1 \mid x) = f(x)$ for all $x \in X$. The randomness of (Π, x_{l+1}) w.r.t. to Q then implies the randomness of (Π, x_{l+1}) w.r.t. to one of the simple P. There are some technical complications connected with the fact that this argument only works for the samples where x_i are not repeated, whereas randomness tests have to be defined for all samples. For details, see Appendix A.

In the previous section we discussed practical implications of Theorem 5. Since Theorem 7 is a negative result, we cannot discuss its practical implication at the same level of concreteness. It shows, however, that attempts to design general density estimation algorithms are doomed unless the patterns are at least occasionally repeated (the extreme case being Glivenko–Cantelli) or we have some prior information about the density $P(j \mid x_{l+1})$ (say, a smooth, in some sense, dependence on x_{l+1}).

As usual for "universal" results (i.e., results using the universal randomness test), Theorem 7 involves an undefined constant, c. This makes its direct application problematic: imagine that, say, $c = 100$. The proof of Theorem 7 shows, however, that, roughly speaking, c can be taken equal to 1. For details, see Appendix B (especially Corollary 15 on p. 18).

5 Regression

The main subjects of this paper are pattern recognition and density estimation, but in this section we briefly discuss regression, stating an analogue of

Theorem 5. Because of the looseness of PAC bounds, we will again deduce the feasibility of regression for diverse data sets from the existence of randomness regression procedures that work in the diverse case; our main reference in this section will be the recent paper [6]. (Besides their looseness, the PAC results tend to make extra assumptions about the i.i.d. distribution generating the data: e.g., all theorems in Section 4.5 "Generalization for Regression" of [2] assume the boundedness of support for the generating distribution. Such assumptions cannot be verified by looking at the data set; cf. Remark 6 above.)

Let F be a *region estimator*, in the sense that F is a function defined on the sequences

$$(\Pi, x_{l+1}) = ((x_1, y_1), \ldots, (x_l, y_l), x_{l+1}) \in (\mathbf{X} \times \mathbf{Y})^l \times \mathbf{X},$$

for all $l = 1, 2, \ldots$, which takes values in the Borel subsets of the real line:

$$F(\Pi, x_{l+1}).$$

We have two natural measures of quality of region estimators. First we can say that a region estimator F is (ϵ, δ)-*valid*, where ϵ and δ are (small) positive constants, if, for any l, any Π (as per (1)), any $x_{l+1} \in X$ and any $P \in \mathrm{Rand}(\Pi, x_{l+1}, \delta)$ (see (2)),

$$P(F(\Pi, x_{l+1})) \geq 1 - \epsilon.$$

This is the notion we are interested in in this paper. The usual randomness definition (used in [6]) is, however, as follows: we say that F is a $(1-\epsilon)$-region estimator, where ϵ is a (small) positive constant, if, for any l and any $P \in \mathcal{P}(\mathbf{Z})$,

$$P^{l+1}\left\{(\Pi, x_{l+1}, y_{l+1}) : y_{l+1} \in F(\Pi, x_{l+1})\right\} \geq 1 - \epsilon$$

(Π is defined by (1)).

Paper [6] constructs $(1-\epsilon)$-region estimators that work well for artificial and standard benchmark data sets in the situation where the patterns x_i are never repeated. Therefore, to demonstrate the feasibility of regression in the framework of this paper ((ϵ, δ)-validity), we only need to establish the connection between these two notions of quality for region estimators. The following theorem establishes this connection in the needed direction. Since we defined randomness level only in the case $\mathbf{Y} = \{0, 1\}$, we state it as a theorem of existence of a statistical test; notice that we do not require $\mathbf{Y}$ is a real line.

Theorem 8 *There exists a conditional statistical test t which satisfies the following. For any $\epsilon > 0$, any computable $(1-\epsilon)$-region estimator F, any l, any training set Π of size l, any test pattern x_{l+1}, and any probability distribution $P \in \mathcal{P}(\mathbf{Z})$,*

$$P(F(\Pi, x_{l+1}) \mid x_{l+1}) \geq 1 - \frac{\epsilon}{t_P(\Pi, x_{l+1} \mid \epsilon)}. \tag{10}$$

Proof T he inequality in (10) can be rewritten as

$$P(F^c(\Pi, x_{l+1}) \mid x_{l+1}) \leq \frac{\epsilon}{t_P(\Pi, x_{l+1} \mid \epsilon)}$$

(E^c stands for the complement of E) and, therefore, as

$$t_P(\Pi, x_{l+1} \mid \epsilon) \leq \frac{\epsilon}{P(F^c(\Pi, x_{l+1}) \mid x_{l+1})}.$$

We can see that it is sufficient to show that the function in the right-hand side of this inequality is a statistical test. Since we assumed computability, we only need to prove

$$P\left\{\frac{\epsilon}{P(F^c(\Pi, x_{l+1}) \mid x_{l+1})} \leq \gamma\right\} \leq \gamma.$$

If this inequality were violated,

$$P\left\{P(F^c(\Pi, x_{l+1}) \mid x_{l+1}) \geq \frac{\epsilon}{\gamma}\right\} > \gamma,$$

we would have

$$P^{l+1}\left\{(\Pi, x_{l+1}, y_{l+1}) : y_{l+1} \notin F(\Pi, x_{l+1})\right\} > \epsilon,$$

which contradicts F being a $(1-\epsilon)$-region estimator.

The notion of (ϵ, δ)-validity of region estimators can be adapted to non-universal tests: "(ϵ, δ)-valid w.r. to a statistical test t" is defined in the same way as "(ϵ, δ)-valid" but using t instead of the randomness level d. Theorem 8 shows that if ϵ and δ are fixed positive constant and F is a $(1-\epsilon)$-region estimator, it will be an $(\epsilon/\delta, \delta)$-valid w.r. to $t(\cdot \mid \epsilon)$ region estimator. In particular, it will be a $(\sqrt{\epsilon}, \sqrt{\epsilon})$-valid region estimator.

Remark 9 *The problem of regression, as understood in this section, is to predict real y_{l+1} (giving, e.g., a predictive region $F(\Pi, x_{l+1})$, hopefully containing y_{l+1}). There is another understanding of the problem of regression, however. According to Vapnik [11] (p. 26), the problem of* regression estimation *is to estimate the* regression function

$$r(x) := \mathbf{E}(Y \mid X = x),$$

which maps $x \in \mathbf{X}$ to the expected value (w.r. to the true distribution P generating the pairs (x_i, y_i)) of Y conditional on $X = x$; in our current context, the problem of regression estimation is to estimate the value $r(x_{l+1})$ of the regression function at the test pattern. We can see that the problem of regression estimation is simpler than the problem of density estimation (where we are required to estimate the full conditional distribution of Y given $X = x_{l+1}$ rather than its mean value). However, in the special case $\mathbf{Y} = \{0, 1\}$ considered in Section 4 the difference disappears; therefore, our results show that the problem of density estimation is infeasible in the diverse case even for binary y_i, unless $r(x_{l+1})$ is close to 0 or 1.

6 Open problems

The following are open problems naturally arising from the results of this paper:

- ⋄ Prove impossibility of density estimation in the case where the training set can contain repeating patterns but there are no patterns identical to the test one.
- ⋄ It is clear that density estimation is possible when the training set contains many patterns identical to the test one (cf. Remark 3 above). It would be interesting to prove a general result containing this positive result and Theorem 7 as special cases.
- ⋄ As discussed in Section 3, it is important (if possible) to find practically meaningful PAC error bounds for diverse data sets (ideally, approaching the quality of the measures of confidence offered by the randomness approach). This might require abandoning the most appealing feature of the usual PAC bounds: logarithmic dependence on δ (as in (4) and (5)).

7 Conclusion

In this paper we argued that, in the "diverse" case where the patterns x_i are never repeated in the given data set:

- ⋄ The problems of pattern recognition and regression can be feasible (moreover, they are often feasible in practice).
- ⋄ The problem of density estimation is never feasible even in the case $\mathbf{Y} = \{0, 1\}$ (see Section 4 and Corollary 15).

Acknowledgements

This work was partially supported by EPSRC through grants GR/M14937 ("Predictive complexity: recursion-theoretic variants") and GR/M16856 ("Comparison of Support Vector Machine and Minimum Message Length methods for induction and prediction"). We are grateful to the Programme Committee for numerous suggestions which greatly helped us to improve the presentation.

References

[1] David R. Cox and David V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[2] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Methods.* Cambridge, Cambridge University Press, 2000.

[3] Alex Gammerman and A. R. Thatcher. Bayesian diagnostic probabilities without assuming independence of symptoms. *Yearbook of Medical Informatics*, pp. 323–330, 1992.

[4] Alex Gammerman, Vladimir Vapnik, and Volodya Vovk. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 148–156, San Francisco, CA, 1998. Morgan Kaufmann.

[5] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer, New York, 2nd edition, 1997.

[6] Ilia Nouretdinov, Tom Melluish, and Volodya Vovk. Ridge regression confidence machine. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[7] Kostas Proedrou, Ilia Nouretdinov, Volodya Vovk, and Alex Gammerman. Transductive confidence machines for pattern recognition. Technical Report CLRC-TR-01-02, Computer Learning Research Centre, Royal Holloway, University of London, 2001. Available from the CLRC web site, `http://www.clrc.rhul.ac.uk`.

[8] Craig Saunders, Alex Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 722–726, 1999.

[9] Craig Saunders, Alex Gammerman, and Volodya Vovk. Computationally efficient transductive machines. In *Proceedings of ALT'00*, 2000.

[10] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

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

[12] Vladimir N. Vapnik and Olivier Chapelle. Bounds on error expectation for Support Vector Machines. In Alex J. Smola, Peter Bartlett, B. Schölkopf, and C. Schuurmans, editors, *Advances in Large Margin Classifiers*, pp. 5–26. MIT Press, 1999.

[13] Volodya Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the 16th International Conference on Machine Learning*, pp. 444–453, 1999.

[14] Volodya Vovk and Vladimir V. V'yugin. On the empirical validity of the Bayesian method. *Journal of Royal Statistical Society* B, 55:253–266, 1993.

[15] Alexander K. Zvonkin and Leonid A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25:83–124, 1970.

8 Appendix A. Proof of Theorem 7

First we give several definitions. Any probability distribution P in $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ (remember that $\mathbf{Y} = \{0, 1\}$) defines a probability distribution $\mathbf{m}_P$ in the set $\mathbf{X}$ and a function $\mathbf{f}_P : \mathbf{X} \to [0, 1]$ in the following way: for any $E \subseteq \mathbf{X}$,

$$\mathbf{m}_P(E) := P\{(x, y) : x \in E, y \in \mathbf{Y}\}$$

(i.e., $\mathbf{m}_P$ is the marginal distribution of P) and

$$\mathbf{f}_P(x) := P(1 \mid x) := \frac{P\{(x, 1)\}}{P\{(x, 0)\} + P\{(x, 1)\}}$$

(i.e., $\mathbf{f}_P(x)$ is the conditional probability that $Y = 1$ given $X = x$; when $P\{(x, 0)\} + P\{(x, 1)\} = 0$, we can set, for example, $\mathbf{f}_P(x) := 1/2$).

Conversely, any probability distribution m in $\mathbf{X}$ and function $f : \mathbf{X} \to [0, 1]$ define a probability distribution $\mathbf{P}_{m,f}$ in $\mathbf{Z}$ with the marginal distribution m and the conditional probabilities f:

$$\mathbf{m}_{\mathbf{P}_{m,f}} = m, \quad m\{x\} \neq 0 \Longrightarrow \mathbf{f}_{\mathbf{P}_{m,f}}(x) = f(x).$$

Consider any probability distribution P in $\mathbf{Z}$. We will say that a probability distribution R in $\mathbf{Z}$ is *simple for* P if $\mathbf{m}_R = \mathbf{m}_P$ and $\mathbf{f}_R(x) \in \{0, 1\}$ for any $x \in \mathbf{X}$ such that $\mathbf{m}_R\{x\} > 0$. We say that R is *simple* if it is simple for some P (i.e., if $\mathbf{f}_R(x) \in \{0, 1\}$ for any $x \in \mathbf{X}$ such that $\mathbf{m}_R\{x\} > 0$).

For any $n = 1, 2, \ldots$, define

$$\overline{\mathbf{Z}^n} := \{((x_1, y_1), \ldots, (x_n, y_n)) \in \mathbf{Z}^n : \forall i, j : x_i = x_j \Rightarrow i = j\}$$

(all data sequences with no repeated patterns are in $\overline{\mathbf{Z}^n}$).

For any probability distribution P in $\mathbf{Z}$ we define a measure (not necessarily a probability distribution) $\overline{P^n}$ on $\mathbf{Z}^n$ as the restriction of P^n to $\overline{\mathbf{Z}^n}$:

$$\overline{P^n}\{\omega\} := \begin{cases} P^n\{\omega\} & \text{if } \omega \in \overline{\mathbf{Z}^n}, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that for a simple $R \in \mathcal{P}(\mathbf{Z})$ we always have $\overline{R^n} = R^n$. More generally, for any $P \in \mathcal{P}(\mathbf{Z}^n)$ we define $\overline{P}$ to be the restriction of P to $\overline{\mathbf{Z}^n}$.

We will derive the statement of the theorem from the following lemma.

Lemma 10 *There exists a constant $c > 0$ such that for any n, any $\omega \in \overline{\mathbf{Z}^n}$ and any probability distribution $P \in \mathcal{P}(\mathbf{Z})$, there exists another probability distribution $R \in \mathcal{P}(\mathbf{Z})$ which is simple for P and satisfies*

$$d_R(\omega) \geq d_P(\omega)/c.$$

To prove this result, we need another two lemmas. In the first of these, we use randomness level w.r.t. to a measure which is not necessarily a probability distribution; this generalized definition is analogous to the definition for probability distributions.

Lemma 11 *There exists a constant $c > 0$ such that, for any n, $\omega \in \overline{\mathbf{Z}^n}$ and $P \in \mathcal{P}(\mathbf{Z})$,*

$$d_{\overline{P^n}}(\omega) \geq d_{P^n}(\omega)/c.$$

Proof W e will prove the statement of the lemma for all $Q \in \mathcal{P}(\mathbf{Z}^n)$, not just for the powers P^n. Define $t_Q(\omega) = d_{\overline{Q}}(\omega)$ if $\omega \in \overline{\mathbf{Z}^n}$ and $t_Q(\omega) = 1$ otherwise. It is easy to check that t is a statistical test, and by the definition of the universal test we have, for $\omega \in \overline{\mathbf{Z}^n}$,

$$d_{\overline{Q}}(\omega) = t_Q(\omega) \geq d_Q(\omega)/c.$$

Lemma 12 *For any n, $\omega \in \mathbf{Z}^n$ and $P \in \mathcal{P}(\mathbf{Z})$,*

$$\overline{P^n}(\omega) = \sum_{g:X \to \{0,1\}} \overline{\mathbf{P}^n_{\mathbf{m}_P,g}}(\omega)\pi_P(g),$$

where

$$\pi_P(g) = \prod_{x \in X} \mathbf{f}_P(x)^{[g(x)]} \qquad \textit{and} \qquad a^{[b]} = \begin{cases} a & \text{if } b = 1, \\ 1 - a & \text{if } b = 0. \end{cases}$$

Notice that $\sum_{g:X\to\{0,1\}} \pi_P(g) = 1$.

Proof L et $\omega = ((x_1, y_1), \ldots, (x_n, y_n))$; without loss of generality we assume $\omega \in \overline{\mathbf{Z}^n}$. We have

$$\overline{\mathbf{P}^n_{\mathbf{m}_P,g}}(\omega) = \mathbf{m}_P(x_1)\ldots\mathbf{m}_P(x_n) y_1^{[g(x_1)]}\ldots y_n^{[g(x_n)]};$$

therefore,

$$\begin{aligned}
\sum_g \overline{\mathbf{P}^n_{\mathbf{m}_P,g}}(\omega)\pi_P(g) &= \mathbf{m}_P(x_1)\ldots\mathbf{m}_P(x_n) \sum_g y_1^{[g(x_1)]}\ldots y_n^{[g(x_n)]} \prod_{x\in X} \mathbf{f}_P(x)^{[g(x)]} \\
&= \mathbf{m}_P(x_1)\ldots\mathbf{m}_P(x_n) \sum_{g:g(x_i)=y_i,\forall i} \prod_{x \in X} \mathbf{f}_P(x)^{[g(x)]} = \mathbf{m}_P(x_1)\ldots\mathbf{m}_P(x_n) \\
&\times \left(\sum_{g:X\setminus\{x_1,\ldots,x_n\}\to\{0,1\}} \prod_{x\in X\setminus\{x_1,\ldots,x_n\}} \mathbf{f}_P(x)^{[g(x)]} \prod_{i=1}^n \mathbf{f}_P(x_i)^{[y_i]} \right) \\
&= \mathbf{m}_P(x_1)\ldots\mathbf{m}_P(x_n) \prod_{i=1}^n \mathbf{f}_P(x_i)^{[y_i]} = \overline{P^n}(\omega).
\end{aligned}$$

We will derive the statement of Lemma 10 from Lemmas 11 and 12.

Proof of Lemma 10. Define

$$t_{\overline{P^n}}(\omega) = \sup_{R \text{ is simple for } P} d_{\overline{R^n}}(\omega). \tag{11}$$

We will show that this function is a statistical test in the following sense: the validity condition (6) is satisfied for measures of the form $\overline{P^n}$ and t is semicomputable from above inside its domain. First notice that it satisfies the condition

$$\overline{R^n}\{\omega = (z_1, \dots, z_n) : t_{\overline{P^n}}(\omega) \leq \gamma\} \leq \gamma$$

for any R which is simple for P. Using Lemma 12, we obtain that it also satisfies the required validity condition

$$\overline{P^n}\{\omega = (z_1, \dots, z_n) : t_{\overline{P^n}}(\omega) \leq \gamma\} \leq \gamma,$$

since $\overline{P^n}$ is a mixture of $\overline{R^n}$. Semicomputability from above follows from the fact that for fixed $\overline{P^n}$ and ω the supremum in (11) involves only finitely many R. It can be shown (a simple argument omitted) that $t_{\overline{P^n}}$ can be extended to a statistical test t_Q (where Q are not necessarily of the form $\overline{P^n}$).

Hence, $t_{\overline{P^n}}(\omega) \geq d_{\overline{P^n}}(\omega)/c_1$ for some constant $c_1 > 0$, and there exists such R (on which the supremum is reached) that

$$d_{R^n}(\omega) = d_{\overline{R^n}}(\omega) \geq d_{\overline{P^n}}(\omega)/c_1,$$

and, by Lemma 11, we have

$$d_{\overline{P^n}}(\omega) \geq d_{P^n}(\omega)/c_2.$$

The statement of the theorem can be easily derived from Lemma 10. Indeed, there exists a measure $P \in \mathcal{P}(\mathbf{Z})$ such that $d_P(\Pi, x_{l+1}) = d_{\mathrm{iid}}(\Pi, x_{l+1})$; without loss of generality we assume $d_{\mathrm{iid}}(\Pi, x_{l+1}) > 0$. Then by Lemma 10 there exists a measure R which is simple for P and such that $d_R(\Pi, x_{l+1}) \geq d_{\mathrm{iid}}(\Pi, x_{l+1})/c$ for some constant $c > 0$. Because the distribution R is simple for P, we have $R(1 \mid x_{l+1}) \in \{0, 1\}$ (the case $\mathbf{m}_P\{x_{l+1}\} = 0$ is excluded by $d_{\mathrm{iid}}(\Pi, x_{l+1}) > 0$).

A Appendix B. Specific form of Theorem 7

In this appendix we will see that the following specific (i.e., not involving the universal test) version of Theorem 7 holds.

Theorem 13 *For any statistical test t there exists a statistical test T such that the following holds. If Π is a training set without repeating patterns and x_{l+1} is a test pattern different from all training patterns, then for any probability distribution $P \in \mathcal{P}(\mathbf{Z})$ there exists a probability distribution $R \in \mathcal{P}(\mathbf{Z})$ which is simple for P and satisfies*

$$t_R(\Pi, x_{l+1}) \geq T_P(\Pi, x_{l+1}).$$

The proof immediately follows from the following specific version of Lemma 10.

Lemma 14 *For any statistical test t there exists a statistical test T such that the following holds. For any n, any $\omega \in \overline{\mathbf{Z}^n}$ and any probability distribution $P \in \mathcal{P}(\mathbf{Z})$, there exists a probability distribution $R \in \mathcal{P}(\mathbf{Z})$ which is simple for P and satisfies*

$$t_R(\omega) \geq T_P(\omega).$$

Proof T his proof follows the proof of Lemma 10. Define

$$t'_{\overline{P^n}}(\omega) = \sup_{R \text{ is simple for } P} t_{\overline{R^n}}(\omega).$$

We already know that this is a statistical test (the universality of d was not used in the proof of Lemma 10). There exists such R (on which the supremum is reached) that

$$t_{R^n}(\omega) = t_{\overline{R^n}}(\omega) = t'_{\overline{P^n}}(\omega).$$

It remains to define $T_P(\omega)$ as $t'_p(\omega)$ if $\omega \in \overline{\mathbf{Z}^n}$ and as 1 otherwise.

To see the relevance of Theorem 13 to the practical problem of density estimation, define (analogously to (2))

$$\mathrm{Rand}_t(\Pi, x_{l+1}, \delta) := \{P : t_P(\Pi, x_{l+1}) > \delta\},$$

where t is any (not necessarily the universal) statistical test; this is the set of probability distributions that the statistical test t deems compatible with the data set (Π, x_{l+1}). We also define

$$\mathrm{DE}_t(\Pi, x_{l+1}, \delta) := \left[\inf_{P \in \mathrm{Rand}_t(\Pi, x_{l+1}, \delta)} P(1 \mid x_{l+1}), \sup_{P \in \mathrm{Rand}_t(\Pi, x_{l+1}, \delta)} P(1 \mid x_{l+1})\right]$$

(set to $\emptyset$ if $\mathrm{Rand}_t(\Pi, x_{l+1}, \delta) = \emptyset$); this is the interval (denoted $[a, b]$ in Section 1) for possible values of the conditional probability that $y_{l+1} = 1$. One performs non-trivial density estimation when $\mathrm{DE}_t(\Pi, x_{l+1}, \delta)$ is an interval that does not contain 0 and 1. Theorem 13 shows that non-trivial density estimation is impossible, in the following sense.

Corollary 15 *For any statistical test t there exists a statistical test T such that the following holds. If Π is a training set without repeating patterns and x_{l+1} is a test pattern different from all training patterns, then*

$$\mathrm{DE}_t(\Pi, x_{l+1}, \delta) \subseteq (0, 1) \Longrightarrow \mathrm{DE}_T(\Pi, x_{l+1}, \delta) = \emptyset.$$

Proof G iven t choose T as in lemma 14 Suppose $DE_t(\Pi, x_{l+1}, b) \subseteq (0, 1)$ but $DE_T(\Pi, x_{l+1}, \delta) \neq \emptyset$. Choose $P \in Rand_T(\Pi, x_{l+1}, \delta)$. By lemma 14 there is an R simple for P that belongs $Rand_t(\Pi, x_{l+1}, \delta)$; therefore, $DE_t(\Pi, x_{l+1}, \delta)$ will contain 0 or 1.

The trivial case $P\{X = x_{l+1}\} = 0$ has to be considered separately: we require that T should be 0 on such (Π, x_{l+1}).

Therefore, we can do non-trivial density estimation at significance level δ only in trivial cases where all i.i.d. distributions can be excluded at the level δ.