PAC-Learning Geometrical Figures

Paul W. Goldberg

Ph.D. Thesis

University of Edinburgh

1992

Abstract

The thesis studies the following problem: Given a set of geometrical figures (such as planar polygons), each one labelled according to whether or not it resembles some "ideal" figure, find a good approximation to that ideal figure which can be used to classify other figures in the same way.

We work within the PAC learning model introduced by Valiant in 1984. Informally, the concepts under consideration are sets of polygons which resemble each other visually. A learning algorithm is given collections of members and non-members of a concept, and its task is to infer a criterion for membership which is consistent with the given examples and which can be used as an accurate classifier of further example polygons.

In order to formalise the notion of a concept, we use metrics which measure the extent to which two polygons differ. A concept is assumed to be the set of polygons which are within some distance of some fixed central polygon. In the thesis we work most extensively with the Hausdorff metric.

Using the Hausdorff metric we obtain NP-completeness results for several variants of the learning problem. In particular we show that it is hard to find a single geometrical figure which is close to the positive examples but not to the negative examples. This result holds under various assumptions about the specific geometrical figures under consideration. It also holds for several metrics other than the Hausdorff metric.

Despite the NP-completeness results mentioned above we have found some encouraging positive results. In particular, we have discovered a general technique for *prediction*. (Prediction is a less demanding learning model than PAC learning. The goal is to find a polynomial-time algorithm which takes as input a sample of labelled examples and is then able to predict the status of further unlabelled examples in polynomial time.) Using our technique we have obtained polynomial-time algorithms for predicting many of the geometrical concept classes studied in the thesis. These algorithms do not classify geometrical figures by measuring their distance from a single "ideal" geometrical figure. Instead, they identify a collection of concepts whose intersection may be used to classify examples reliably.

It is natural to consider the case in which only positive examples are available. In the thesis we show that some but not all of the concept classes may be

predicted from positive examples alone.

We consider prediction to be a useful goal, since it solves the practical problem of classifying unlabelled examples. However in the final section of the thesis we show a theoretical limitation to the effectiveness of this technique. In particular, assuming the existence of trapdoor functions, no polynomial-time algorithm for prediction exists for polygons in the plane which are assumed to be equivalent under classes of isometries that include rotations.

Contents

Chapter One

The Learning Model

In this chapter we introduce the PAC learning model of Valiant. In section 1.1 we give the basic definitions and notation involved, and in section 1.2 we give some examples of learning in this computational model. In section 1.3 we examine the main obstacles encountered in learning and introduce some further concepts, and the main theorems that will be used in later chapters. In 1.4 we then consider variants of the model, results for some of which appear later in this thesis.

1.1. Preliminaries

Background

The PAC (Probably Approximately Correct) model was invented by Valiant [V84, 85] and has given rise to a great deal of research activity in recent years. A detailed introduction and survey of this work can be found in Natarajan's book [N91a]. The purpose of the model is to give a precise computational model of the process of learning without explicit programming. The previous work that it is most strongly related to is inductive inference, of which Angluin and Smith [AS83] provides a survey.

PAC learning is more probabilistic than previous models. In PAC learning, we must accept an element of uncertainty (to be formalised below) that a hypothesis satisfies an error bound. This is intended to enable us to place realistic bounds on the number of observations necessary to achieve this kind of learning. In most earlier work on inductive inference, induction is seen as a limiting process. Much of this work has been based on the aim of "identification in the limit", the paradigm formalised by Gold in [G67]. Here a hypothesis is eventually obtained which is either exact, or guaranteed to satisfy some error bound (as in [W74] for example.)

PAC learning is a complexity-based model. This is connected with the fact that statements about learning may be based solely on the number of examples

seen. Learnability requires the number of examples that need to be seen and the time taken to process them, to be polynomial in parameters of the problem.

Definitions: An *instance domain* is a set (usually denoted $\mathcal{X}$) which is known to a learner, and which consists of all allowable objects which may form the input. A *concept* C is simply a subset of this domain, and objects input to a learning algorithm are classified according to whether or not they fall within this subset, that is exemplify the concept. An *example* is a member e of $\mathcal{X}$ together with the value of the indicator function of C at e (hence classified according to membership/non-membership of C.) A *sample* is a collection of examples all classified according to their membership of a single concept, unknown to the learner.

The objective of a learning algorithm is to receive as input a sample and return a concept which is a good approximation to the unknown concept being used to classify the examples. This unknown concept is called the *target concept*, and the output of a learning algorithm is called the *hypothesis*.

A *concept class* $\mathcal{C}$ is a collection of subsets of $\mathcal{X}$, that is $\mathcal{C} \subseteq 2^{\mathcal{X}}$. A concept being learned is assumed to belong to some fixed concept class, which is known to the learner.

Remarks: In general a concept class $\mathcal{C}$ is a proper subset of $2^{\mathcal{X}}$. If we have $\mathcal{C} = 2^{\mathcal{X}}$ then a sample may have to be more or less exhaustive in order for much to be known about the target concept. It is worth emphasising at this point that learnability (or non-learnability) is a property of a concept class, not an individual target concept, whose learnability will usually depend on what class it is known to belong to. A concept class is learnable if and only if all of its concepts are learnable, a notion formalised below. We motivate these definitions with the following example.

Example 1.1: [BEHW89] Let $\mathcal{X}$ be the set of points in the plane and let $\mathcal{C}$ be the set of axis-parallel rectangles in the plane – that is, a concept in $\mathcal{C}$ is the set of points contained in some axis-parallel rectangle. Then a sample will consist of a finite set of points, the *positive examples* which are contained in some unknown rectangle, and the *negative examples* which lie outside this rectangle. It is easy to find a hypothesis rectangle which contains all the positive and none of the negative examples (just take the smallest one containing the positives) and it turns out that such a hypothesis is a good approximation to the target concept in the sense defined below.

PAC-ness

In general, we cannot expect to learn a concept exactly from a limited number of examples, but must settle for a good approximation. We introduce a notion of measurable error by letting P be a probability distribution function on the domain $\mathcal{X}$ according to which the examples forming a sample are generated. Then for target concept $C \in \mathcal{C}$ and hypothesis $H \in \mathcal{C}$, define the error functions

$$e^+(H) = prob(\text{a random (over p.d.f. } P) \text{ element of } \mathcal{X} \text{ is in } C \setminus H)$$

$$e^-(H) = prob(\text{a random (over p.d.f. } P) \text{ element of } \mathcal{X} \text{ is in } H \setminus C)$$

A learning algorithm aims to limit these errors with a parameter ϵ, which is an upper bound on e^+ and e^-. However since even a large sample may be uninformative or misleading, we introduce a second parameter δ representing uncertainty; δ is the probability that the error of a hypothesis actually exceeds the error bound ϵ. The two parameters can be seen to measure the accuracy with which the hypothesis predicts further examples.

So with probability $1 - \delta$ the error is limited to ϵ. Thus this is called PAC learning, for "probably approximately correct", an acronym introduced by Angluin in [A87].

Definition: A *learning function* for concept class $\mathcal{C}$ with respect to ϵ, δ and a sample size m dependent on ϵ and δ is a mapping from samples of size m to hypotheses, such that for all distributions P and all target concepts in $\mathcal{C}$, with probability at least $1 - \delta$ a hypothesis has error at most ϵ.

Definition: A *learning algorithm* is an algorithm which computes a learning function.

Definition: A concept class is *uniformly learnable* if there exists a learning algorithm which for large enough sample size achieves PAC learning for any $\delta, \epsilon > 0$.

The above definitions ignore efficiency issues, and characterise concept classes according to whether it ever becomes possible to obtain approximations from a large enough sample. We now consider polynomial learnability, with which this thesis is mainly concerned.

Informally, the learnability of a concept class is the ease with which one can obtain a hypothesis which is a good predictor of unclassified examples. We seek

polynomial-time algorithms which compute learning functions, that is, polynomial in parameters of the learning problem. The run-time and sample size required clearly increase as δ and ϵ decrease, and we are interested in algorithms that run in time polynomial in ϵ^{-1} and δ^{-1}. As observed in [V91], a more accurate acronym than "PAC" would be "epac", to include the criterion of efficiency in our algorithms.

For given ϵ, δ, the smallest m for which a learning function exists is called the *sample complexity* of the learning function. This is taken to be equal to ∞ if no finite sample is sufficient. So we want the sample complexity to be polynomial in the parameters of the learning problem. We also want a learning function to be computable from a given sample in polynomial time.

What we have described above is the "functional model" of PAC learning, in which a learning algorithm is viewed as implementing a function from samples to hypotheses. In [HKLW91] it is shown that various variations in how examples are made available to the learner do not affect the set of concept classes that can be learnt in polynomial time. For example the "oracle model" is also widely used. Here the examples are assumed to be made available to a learning algorithm by an oracle *EXAMPLES* which draws them according to P, requiring unit time to draw a single example. Hence PAC-learnability requires a learning algorithm with access to this oracle to halt in polynomial time and output a hypothesis.

Another variant is to assume that two oracles are available, $EXAMPLES^+$ and $EXAMPLES^-$, returning positive/negative examples respectively, according to P restricted to positive/negative examples. (The reason why this apparently more powerful version is equivalent to access to *EXAMPLES* only is that with high probability we may simulate calls to $EXAMPLES^+$ and $EXAMPLES^-$ in polynomial time, using *EXAMPLES*, by waiting for an example of the appropriate type to be returned.)

Definition: The *hypothesis class* of a learning problem (denoted by $\mathcal{H}$) is the set of allowable hypotheses which a learning algorithm may return, $\mathcal{H} \subseteq 2^{\mathcal{X}}$.

Remark: In learning a concept class $\mathcal{C}$, $\mathcal{H}$ is generally taken to be equal to $\mathcal{C}$, and "learnability of $\mathcal{C}$" means that there is an efficient learning algorithm for $\mathcal{C}$ which returns hypotheses in $\mathcal{C}$. However it is sometimes necessary to take a larger hypothesis class for a learning problem to be tractable. Where $\mathcal{H} \neq \mathcal{C}$ we refer to "learnability of $\mathcal{C}$ by $\mathcal{H}$".

1.2. Examples

To motivate the ideas seen so far we consider some more examples of PAC learnability and non-learnability.

Example 1.2: $\mathcal{X} = \mathbb{R}$, $\mathcal{C} = \{[a,b] : a,b \in \mathbb{R}, a \leq b\}$ ie. closed bounded intervals on the real numbers. Hence examples are real numbers for which it is given whether or not they lie in an unknown interval.

A hypothesis could then be generated by taking the closed interval bounded by the highest and the lowest positive examples. In fact any choice of consistent hypothesis (that is, one that contains the positive examples but not the negatives) turns out to be a good one. [BBM90] introduces the term "solid learnability" to describe this phenomenon. A concept class $\mathcal{C}$ is *solidly learnable* by hypothesis space $\mathcal{H}$ if and only if there exists a sample size $m(\epsilon, \delta)$ such that *any* hypothesis in $\mathcal{H}$ which is consistent with an m-sample (ie. a sample of size m) is probably approximately correct (according to ϵ, δ.)

For the distribution-independent model considered here, solid learnability is equivalent to uniform learnability. This is quite a strong property of a concept class, not true of most interesting classes. We will see however that the model can be extended to permit a notion of non-uniform learnability, where a hypothesis has to be chosen from a set of consistent hypotheses which may include bad ones.

Consider for example the following extension of the above concept class:

Example 1.3: $\mathcal{X} = \mathbb{R}$, $\mathcal{C}$ = finite unions of closed intervals in the real line. $\mathcal{C}$ is not uniformly learnable, and not all choices of consistent hypothesis are likely to be accurate. However, a method we would like to accept is to choose a set of intervals of lowest cardinality which is consistent with the sample, that is one interval for each unbroken sequence of positive examples.

This rule reflects the intuition that it is absurd for a hypothesis to contain intervals where no positive examples appeared in the sample, or unnecessary breaks in the intervals. It also incorporates the requirement that a large number of intervals be given more time to learn than a small number. This observation introduces a new parameter to a learning problem (in which an algorithm must be polynomial), namely the size or complexity of the target concept. This issue is discussed in more detail in the next section.

Boolean learning problems are often considered in this framework, where a concept is represented by a boolean formula and consists of the set of all its satisfying assignments, with the domain the set of all vectors of value assignments for its variables. For a class of boolean formulae, a PAC learning algorithm must be polynomial in the number of variables in the target concept.

Example 1.4: A *monomial* consists of a conjunction of unnegated boolean variables. For a fixed number n of boolean variables we may define a concept to be the set of all satisfying assignments to some monomial over n variables, and the concept class $\mathcal{C}_n$ to be the set of all such concepts.

A plausible learning algorithm for the concept class is to take as the hypothesis the conjunction of those variables which are consistently set to true in the positive examples [V84]. This algorithm turns out to satisfy the criteria for polynomial learning.

Example 1.5: DNFs [V85]. Concepts are represented by disjunctive normal form formulae over n boolean variables and consist of the sets of their satisfying assignments. As before, the domain is the set of all vectors of n truth values. Learnability* of this class is a key open problem, which has only been answered for various restricted versions.

Example 1.6: Pattern strings [KP89]. The domain is the set Σ^* of strings over an alphabet Σ. Let Σ' be Σ augmented with a fixed number of variable symbols. Concepts are represented by strings in Σ'^*, and a member of a concept represented by $\sigma \in \Sigma'^*$ is any string in Σ^* obtainable by consistently replacing each variable symbol in σ by a string in Σ^*. Here, learnability requires more time for longer (more complex) patterns.

1.3. Further Issues in PAC learning

We may identify three potential obstacles to learning.

1.) Given representations of an example and a hypothesis it may not be possible to determine in polynomial time whether the example is an element of the hypothesis.

2.) To have enough information for a concept to be learned with enough accuracy may require an infeasibly large sample.

3.) There may exist learning functions with small sample complexity, but any such function may be hard to compute.

* in the "Occam" sense, defined on page 10

In any learning problem in this paradigm, concepts and examples are assumed to be represented as words over some alphabet Σ. For representation schemes $r_{\mathcal{X}} : \Sigma^* \longrightarrow \mathcal{X}$, $r_{\mathcal{H}} : \Sigma^* \longrightarrow \mathcal{H}$ (note that $\mathcal{H} \supseteq \mathcal{C}$) we require the test of $r_{\mathcal{X}}(s_1) \in r_{\mathcal{H}}(s_2)$ to be easy to perform, given $s_1, s_2 \in \Sigma^*$. It should also be feasible to test whether a string in Σ^* represents a valid concept or example. It turns out that for the natural representations of geometrical objects as tuples of their (real-valued) coordinates, (1) is not a problem.

The problem of *polynomial-sample learning*, where one is given a lot of time to learn from a small sample, has been considered (see [V91].) This notion of learning reflects a notion of expensive oracle calls (to *EXAMPLES*) and focuses on the information content of a sample of limited size. Item (2) is the possibility that a concept class is not polynomial-sample learnable. It will be shown that all the classes considered in this thesis satisfy this requirement, but that item (3) is the usual obstacle.

The following theorem of [PV88] relates learning problems to more traditional complexity-theoretic search problems.

Define the *consistent hypothesis problem* for a concept class to be the problem of, given any sample as input, find a hypothesis consistent with it, if one exists.

Theorem 1.1: [PV88] *Under the assumption $NP \neq RP$, if the consistent hypothesis problem is hard then the associated learning problem is hard.*

Proof (sketch): PAC-learnability requires a concept class to be learnable for any probability distribution being used to generate examples. In particular, we consider the effect of choosing a distribution which is uniform over a set of examples for which it is hard to find a consistent hypothesis. In this case, a learning algorithm becomes a randomised algorithm for the NP-hard consistent hypothesis problem. □

The Vapnik-Chervonenkis Dimension

An important combinatorial tool in the analysis of infinite concept classes is the Vapnik-Chervonenkis dimension of a concept class [HW87], which has its origins in the "growth function" of [VC71]. This provides us with a criterion for learnability as well as an upper bound on the sample size required to learn a concept class of known Vapnik-Chervonenkis dimension. Note that it is an

information-theoretic property of a concept class, independent of complexity. It is defined as follows:

Definition: A subset S of the domain $\mathcal{X}$ is said to be *shattered* by a concept class $\mathcal{C}$ on $\mathcal{X}$ if for every partition of S into two disjoint subsets S_1 and S_2, there exists a concept $C \in \mathcal{C}$ such that $S_1 \subseteq C$ and $S_2 \subseteq \mathcal{H} \setminus C$.

Definition: The *Vapnik-Chervonenkis dimension* of a concept class $\mathcal{C}$ is the cardinality of the largest set shattered by $\mathcal{C}$. (This will be abbreviated to V-C dimension in what follows, and $\dim(\mathcal{C})$ will denote the V-C dimension of $\mathcal{C}$.)

Example 1.7: The set of intervals in the real numbers (example 1.2) has V-C dimension 2, since any subset of a set of two points in $\mathbb{R}$ has an interval which contains it but not the other point(s). For a set of three points however, no interval can contain the end points but not the middle one. Unions of n intervals in $\mathbb{R}$ have V-C dimension $2n$. The class of finite unions of intervals in $\mathbb{R}$ has infinite V-C dimension, but is learnable (as described above) in a weaker sense, which is explained below.

Example 1.8: Let $\mathcal{C}_n$ be halfspaces in $\mathbb{R}^n$, having domain $\mathcal{X}_n$, points in Euclidean n-space. Then a result of [WD80] is that the V-C dimension of $\mathcal{C}_n$ is $n+1$.

Note that any finite concept class $\mathcal{C}$ has V-C dimension $\leq \log|\mathcal{C}|$. Hence a class of concepts which are represented by strings of length $\leq s$ over an alphabet Σ has V-C dimension $\leq s\log(|\Sigma|)$.

[BEHW89] show the equivalence of finite V-C dimension of a concept class to uniform learnability, with the following theorem:

Theorem 1.2: [BEHW89] *Let $\mathcal{C}$ be a non-trivial, well-behaved * concept class.*

i.) There exists a learning function (not necessarily polynomial-time computable) mapping samples to hypotheses in $\mathcal{C}$ if and only if the V-C dimension of $\mathcal{C}$ is finite.

ii.) If the V-C dimension of $\mathcal{C}$ is d, where $d < \infty$ then

a.) for $0 < \epsilon < 1$ and sample size at least

$$\max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{13}{\epsilon}\right)$$

* This is a relatively benign measure-theoretic assumption discussed in an appendix in [BEHW89]

any function mapping such samples to a consistent hypothesis in $\mathcal{C}$ is a learning function for $\mathcal{C}$ (which may not be evaluatable in polynomial time), and

b.) for $0 < \epsilon < \frac{1}{2}$ and sample size less than

$$\max\left(\frac{1-\epsilon}{\epsilon}\ln\frac{1}{\delta}, d(1-2(\epsilon(1-\delta)+\delta))\right)$$

there is no learning function from such samples to any hypothesis class.

In the case of a class $\mathcal{C}$ of boolean formulae such as examples 1.4 and 1.5, the V-C dimension depends on the number n of variables in the target concept. In this case n becomes a parameter of a learning problem for $\mathcal{C}$. The concept class acquires a stratification structure $\mathcal{C} = \{(\mathcal{X}_n, \mathcal{C}_n)\}_{n\geq 1}$. Note that if $\mathcal{C}_n$ is n-variable boolean formulae in some class $\mathcal{C}$ of formulae, then 2^n is an upper bound on the V-C dimension.

n becomes another parameter of the learning problem, and *polynomial learnability with respect to domain dimension* becomes learning with an algorithm whose run-time is polynomial in ϵ, δ and n.

Theorem 1.3 [BEHW89] *Let $\mathcal{C} = \{(\mathcal{X}_n, \mathcal{C}_n)\}_{n\geq 1}$ be a concept class. $\mathcal{C}$ is polynomially learnable if and only if there is a randomised polynomial-time algorithm which takes a sample of $\mathcal{C}_n$ and with some fixed probability returns a consistent hypothesis in $\mathcal{C}_n$, and $\dim(\mathcal{C}_n)$ is polynomial in n.*

As corollaries to theorem 1.3 and this characterisation of learnability, the following classes are not polynomially learnable [KLPV87]:

1.) Disjunctions of two conjunctions

2.) Boolean threshold functions

3.) Boolean formulae in which each variable appears once.

Note that for fixed number n of boolean variables, the above boolean formulae are naturally represented using expressions of length polynomial in n. (This implies polynomial V-C dimension, meaning that the consistent hypothesis problems for the above examples are hard.) Where this is not the case, we may relax the conditions for learnability by allowing a learning algorithm more time for concepts which require longer expressions, that is, let s be the syntactic complexity of a concept (length of formula representing it) and require

for polynomial learnability that an algorithm PAC-learns in time polynomial in $\{s, \epsilon, \delta\}$. We have seen this applied in example 1.3, where more examples are required to learn a union of a large number of intervals. Without allowing the runtime of a learning algorithm to depend on the number of intervals in the target concept, learning would not have been possible. In the following section we give the theoretical justification for this approach.

Occam's Razor

As pointed out in [BEHW87], for a finite hypothesis space of size r, the probability that a hypothesis with error greater than ϵ is consistent with a target concept on a sample of size m, is less than $(1-\epsilon)^m r$. Thus all finite concept classes are uniformly learnable, if not necessarily in polynomial time (Given enough examples, the probability becomes arbitrarily high that only the target concept will be consistent with them.) For infinite hypothesis spaces however, it is often possible to choose consistent hypotheses for which no bounds can be placed on the error.

It is possible to overcome this problem by appealing to the principle of *Occam's Razor* * whereby given more than one explanation for a phenomenon, the simplest should be preferred. In learning theory this translates to finding the simplest hypothesis that is consistent with a sample. Thus we define some hierarchy of complexity on hypotheses by defining a function **size** : $\mathcal{H} \longrightarrow \mathbb{N}$. This is usually the length of the hypothesis in some standard encoding (the length of the word in Σ^* representing the hypothesis.) The problem of finding the simplest (or one of the simplest) hypotheses is an important feature of learnability theory. We give a theorem of Blumer et al. [BEHW87], showing that this is an effective way of obtaining a good hypothesis.

Definition: [BEHW87] Let $\mathcal{C}$ be a concept class with instance domain $\mathcal{X}$. We say that $\mathcal{C}$ is *polynomially Occam-learnable* (with respect to a fixed encoding) if there exists a learning algorithm for $\mathcal{C}$ and a minimum sample size $m(\epsilon, \delta, s)$ polynomial in $\epsilon^{-1}, \delta^{-1}, s$, where s is the complexity of the target concept such that

* "Entities are not to be multiplied beyond necessity" – William of Occam, c1320

a) For all concepts in $\mathcal{C}$ and all probability distributions P on $\mathcal{X}$, given $m(\epsilon, \delta, s)$ independent observations, the algorithm returns a hypothesis which with probability at least $1 - \delta$ has error $\leq \epsilon$.

b) The algorithm produces the hypothesis in time polynomial in the length of the sample (under the given encoding of the observations).

Definition: [BEHW87] An *Occam algorithm* is a learning algorithm which for some fixed parameters $c \geq 1$, $0 \leq \alpha < 1$, finds a hypothesis consistent with all observations, of complexity $\leq n^c m^\alpha$ from a sample of size m of any concept in $\mathcal{C}$ of complexity $\leq n$, and runs in time polynomial in the length of the sample.

In this definition, the complexity of a concept is the length of its representation, where the alphabet Σ must be finite. An Occam algorithm does not necessarily find the simplest consistent hypothesis, but does achieve a kind of data compression, by finding a hypothesis whose complexity is asymptotically smaller than the sample, but "explains" the sample. The following theorem shows that the existence of an Occam algorithm implies polynomial learnability, where $\mathcal{H} = \mathcal{C}$.

Theorem 1.4: [BEHW87] *Given access to oracles $EXAMPLES^+$ and $EXAMPLES^-$, for a target concept of complexity $\leq n$, an Occam algorithm with parameters $c \geq 1$, $0 \leq \alpha < 1$ will produce a hypothesis with error and uncertainty bounds ϵ and δ, in time polynomial in ϵ^{-1}, δ^{-1} and n. The sample size required is*

$$O(ln(\delta^{-1})/\epsilon + (n^c/\epsilon)^{1/(1-\alpha)})$$

The proof of this theorem is a counting argument – there are not enough concepts represented by strings of length bounded by $n^c m^\alpha$ to shatter a large set of examples. Hence the V-C dimension must be low.

In a concept class where Σ contains $\mathbb{R}$ you cannot let **size** be the length of the representation and rely on the V-C dimension to be limited for "small" concepts — it is necessary to appeal to the V-C dimension of $\mathcal{H}$. The following theorem extends the previous result (valid for finite $|\Sigma|$) to the sort of classes of interest in this thesis, in which Σ contains $\mathbb{R}$.

[BEHW89] generalises the notion of an Occam algorithm in [BEHW87] by requiring the set of hypotheses produced by the set of all m-samples consistent

with concepts of size s to have V-C dimension $O(s^c m^\alpha)$. This leads to the following theorem:

Theorem 1.5: [BEHW89] *Let C be a concept class with a given concept complexity measure.*

i.) If there is an Occam algorithm for C then C is polynomially learnable.

ii.) An Occam algorithm which maps m-samples of concepts in C of size s to a set of consistent hypotheses in C of V-C dimension $s^c m^\alpha$ ($c \in \mathbb{N}$, $0 \leq \alpha < 1$) achieves PAC-learning, requiring sample size at most:

$$m = \max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \left(\frac{8s^c}{\epsilon}\log\frac{13}{\epsilon}\right)^{\frac{1}{1-\alpha}}\right)$$

If a hypothesis is drawn from a class of V-C dimension $s^c(\log m)^l$ then the second term in the bound may be replaced by

$$\frac{2^{l+4}s^c}{\epsilon}\left(\log\frac{8(2l+2)^{l+1}s^c}{\epsilon}\right)^{l+1}$$

1.4. Variants of the Learning Model

Valiant's learning model admits many variants, some of which have been mentioned in passing earlier in this chapter. In [HKLW91] it is shown that a lot of these are equivalent, indicating a high degree of robustness of the model to differing assumptions about how data are made available to a learner. We have seen that access to the oracle *EXAMPLES* is equivalent to access to oracles $EXAMPLES^+$ and $EXAMPLES^-$ for positive and negative examples respectively, according to P conditioned on an example of the appropriate type being chosen. (This equivalence is in terms of the concept classes which become learnable.) It also makes no difference whether or not the parameters of the problem (ϵ, δ, n) form part of the input.

There are also an assortment of variations which do affect which classes are learnable, two of which are considered in this thesis. We mention some of these variations before focusing on the ones to be worked with later.

The problem has for example been considered using other oracles, such as $MEMBER(x)$, returning true if the input x is contained in the concept,

false otherwise, or $SUPERSET(S)$, returning true if the input set S contains the concept as a subset, and if not returning an element of the concept not contained in S. These models where the learner can actively ask questions are more powerful in that more classes become learnable, but these are not considered in this thesis. These versions contrast with the model considered here in that computation is carried out during the data-gathering, in order to choose inputs for the oracle calls.

Another different version of the learning model not considered here is learning from known probability distributions. This turns out to be a less demanding form of learnability, sometimes claimed to be too restrictive for real-world applications. For example in [KLPV87] it is shown that μDNF formulae ($DNFs$ where each variable occurs once only) are learnable if the distribution P is known to be uniform.

In this thesis we consider two alternative formulations which change the set of learnable concept classes. These are prediction, and learning from positive examples only.

Prediction

It has been noted that a learning task may become easier if the hypothesis space is increased. At an intuitive level, this alteration allows more flexibility in pronouncing on what distinguishes positive from negative examples. This approach is taken to its logical conclusion in *prediction* [HLW88], so-called because its goal is similar to that of statistical prediction. Here a hypothesis is allowed to be any computable classification scheme for members of $\mathcal{X}$. Polynomial predictability is then defined analogously to polynomial learnability, with the additional requirement that the hypothesis obtained should have a polynomial-time membership test. This requirement is due to the assumption that any such scheme should have practical usefulness for classifying subsequent examples.

Prediction is shown to be robust to varying assumptions in [HKLW91], for example it may be reformulated as the following problem: Given a sample and an unclassified member e of $\mathcal{X}$, classify e with probability of correctness at least $1 - \epsilon$.

Given the non-learnability of most concept classes considered in this thesis, we consider the prediction problem in chapter 6, where it is shown that this relaxation of the demands on a hypothesis can make the problem tractable. [PW90] defines a notion of *prediction-preserving reduction* which allows the

construction of a completeness class of assumedly unpredictable concept classes. We show that one of our geometrical learning problems is intractable in this sense.

Non-predictability results depend on a cryptographic assumption, namely the existence of trapdoor functions, an assumption that implies $P \neq NP$. These are classes of functions for which any such function f is easy to compute, along with its inverse f^{-1}, but it is hard to discover f^{-1} given f. The RSA public-key functions are the best known class of functions believed to have this property.

The connection between trapdoor functions and prediction was made in [KV89]. They show that this implies that the class of general boolean formulae is not predictable, and similarly for regular sets, represented by DFAs.

Learning from Positive Examples only

In this more demanding learning model, only one oracle, namely $EXAMPLES^+$, is available to the learner. This appears to be a natural problem to consider, because it seems to correspond to some real-world learning processes. Evidence that it is more restrictive is that disjunctions cannot be polynomially learned from positive examples for purely information-theoretic reasons. (A complementary result is that conjunctions cannot be learned from negative examples [KLPV87].)

Definition: [N91b] A concept class $\mathcal{C}$ is *minimally consistent* provided that for any finite set S contained in some concept $C \in \mathcal{C}$, there exists a concept containing S which is a subset of any other concept containing S.

Notation: [N91b] For a sample S of positive examples, define $M(S)$ to be the minimal concept containing S, with $M(S)$ undefined if no such concept exists.

[N87] introduces the following notion of dimensionality for learning from positive examples.

Let $\mathcal{C}$ be a concept class with domain $\mathcal{X}$.

Let $d^+(\mathcal{C})$ be the size of the largest subset S of $\mathcal{X}$ with the properties:

1.) $S \subseteq C$, for some $C \in \mathcal{C}$

2.) for all $x \in S$, there exists $C \in \mathcal{C}$ such that $x \notin C$ but $(S - \{x\}) \subseteq C$.

$d^+(\mathcal{C}) = \infty$ if such sets S exists with arbitrarily many elements.

Theorem 1.6: [N91b] *A concept class $\mathcal{C}$ is uniformly learnable from positive examples alone if and only if $\mathcal{C}$ is minimally consistent and $d^+(\mathcal{C})$ is finite.*

For such a class, the hypothesis to take, given sample S of positive examples, is $M(S)$.

Given that this version of the problem is more demanding than learning from positive and negative examples, we combine this approach with prediction, and show that for some concept classes considered in this thesis, a suitable augmentation of the hypothesis class can make learning from positive examples possible, even for some non polynomially learnable classes. As a result, d^+ becomes analogous to the V-C dimension for prediction from positive examples. Learning from positive examples alone may be made possible from knowledge of the underlying probability distribution, but this approach is not considered here.

Chapter Two

The Concept Classes

In this chapter we make precise the notion of a set of similar (mutually resemblant) geometrical objects. Intuitive notions of similarity or resemblance may be formalised using metrics on geometrical objects, so that proximity under the metric indicates visual resemblance. In section 2.1 we describe the framework in which concepts are defined, and summarise the kinds of geometrical objects which are considered. In section 2.2 a variety of possible metrics are defined and analysed. In section 2.3 we conclude with a discussion of other criteria which have been used for resemblance, but are not used in this thesis.

2.1. Definition of a Concept

Before we show how similarity of shape or appearance may be measured, we explain in this section how concept classes are constructed, given a suitable measure. This measure will consist of a function $d : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ denotes the non-negative reals. For elements $e_1, e_2 \in \mathcal{X}$ the value $d(e_1, e_2)$ measures the extent to which they differ. In pattern-matching applications it is desirable that a cost function for geometrical figures should be a metric [M87, HKS91]. However given the subjective nature of visual resemblance, it is not surprising that there exist a wide variety of different metrics that are used for this purpose.

Given a suitable metric d on the set $\mathcal{X}$ of geometrical objects, a concept can be defined as the set of all members of $\mathcal{X}$ lying within some fixed distance r of some fixed element of $\mathcal{X}$. Hence a concept is naturally definable as a sphere in the metric space $\{\mathcal{X}, d\}$. A concept class is a set of such spheres, and in this thesis the set of geometrical objects at the centre of one or more of these spheres will usually be all of $\mathcal{X}$, and for each centre, the radius will either be a constant r or allowed to be any positive real number.

Since a concept is a sphere in this sense we will use the expression "radius of a concept" to mean the radius of that sphere, and if the radius of all concepts in a concept class is fixed at r, we will say that the concept class has radius r.

A concept *defined by* geometrical object c will be used to mean a concept whose centre is the object c, considered as a sphere in the metric space.

Domains of Geometrical Objects

The original formulation considered by the author of the problem of learning geometrical patterns took $\mathcal{X}$ to be the set of polygons in the plane, since a planar polygon is a convenient way of representing some concrete object that might arise in applications [F74]. However, working with simpler geometrical figures allows us to convey results more cleanly, which may then be extended to planar polygons. In particular $\mathcal{X}$ is usually taken to be finite sets of points on the real line or in the plane. Note however that some metrics such as the Fréchet metric (described later), are specific to polygons, and not finite point sets.

For the purpose of learning "shape", rather than a polygon in some particular orientation, we need to work with $\mathcal{X}$ modulo a group of transformations that are considered to preserve shape. (This is the usual definition of "shape" [HKS91].) This notion of shape will be captured by a metric that registers as identical polygons that just differ from each other by being rotated and translated, or dilated if we are not interested in learning size. If a metric does not have this property then it can be modified to have it by defining the distance between two polygons to be their distance apart, minimised over all rotations and translations of one of them. We will use the word *isometry* to refer to a combination of rotations and translations, but not reflections.

In learning, the output of an algorithm, the hypothesis, must take the same form as a concept, that is, it must consist of a member of $\mathcal{X}$ (the centre) together with a particular value for the radius, if that is not fixed. This definition of a concept can be associated with the idea of learning an approximation to an unknown polygon (or other geometrical object), rather than a set of similar polygons. This is a convenient way of viewing the problem, which will be used implicitly later in phrases such as "learning a polygon". However, in all of what follows it is a *set* of polygons that is being obtained, rather than one in particular. And indeed it is sometimes the case that finding an approximating set is easier if it is *not* based on a single polygon.

2.2. Metrics for Resemblance

We restrict our attention to metrics which have been the subject of study in the computational geometry literature. In this section we describe the metrics we use to obtain pattern learning results.

An important feature of any metric which we take as a criterion for geometrical resemblance must be that the metric should be easy to compute. That is, given computational representations of two geometrical objects, it must be possible to compute the distance between those two objects in time polynomial in the combined lengths of their representations. Failure of a metric to satisfy this criterion means that the resulting hypothesis is, in a strong sense, of no more use that the collections of positive and negative examples, since it does not permit classification of subsequent unclassified examples in polynomial time. We refer to the problem of testing whether a given example belongs to a given concept as the *membership test* for the concept class. This is not a potential problem for most learning algorithms, dealing as they do with boolean concepts. Even when geometrical concepts appeared in earlier work [BEHW89], the existence of fast membership tests was easy to establish. Fortunately all the concept classes of interest here also have easy membership tests.

Since an object is a polygon or finite set of points in Euclidean space E or E^2, the natural way to represent one is as the tuple of real numbers giving the coordinates of its vertices in order or the set of points comprising it. We assume here that real numbers occupy unit space, and operations on real numbers take unit time (and are exact.) This consideration is discussed further in the next chapter, on the Vapnik-Chervonenkis dimension.

All of the metrics that we consider are computed from distances between points on the objects on which the metric is used. The usual notion of distance between two points is the Euclidean norm L_2, but the L_∞ (maximum) norm is also used. In many of the results presented here it is sufficient to consider just points on the real line, for which these norms are equivalent.

The Hausdorff Metric

The Hausdorff metric (see e.g. [G83]) is defined as follows:

Let P_1, P_2 be two sets of points in a metric space $\{S, d\}$ (where d is the metric on elements of S). Then the Hausdorff distance between P_1 and P_2 is

$$H(P_1, P_2) = \max\Big\{ \sup_{p_1 \in P_1} \{ \inf_{p_2 \in P_2} \{d(p_1, p_2)\}\}, \sup_{p_2 \in P_2} \{ \inf_{p_1 \in P_1} \{d(p_1, p_2)\}\}\Big\}$$

Here we typically take $S = \mathbb{R}$ or $\mathbb{R}^2$, $d = L_2$ or L_∞. The definition makes sense in a much more general setting.

The reason why the Hausdorff metric reflects geometrical resemblance is that for the distance between two sets of points in S (for example polygons in the plane) to be $\leq r$ we need every point on each set to be within r of some point on the other set.

So if the Hausdorff distance between two polygons is small, every point on each polygon is close to some point on the other polygon. Note however that the interiors of the polygons need not overlap, and their areas may differ greatly.

Alt et al. [ABB91] show that this metric is polynomial-time computable, even when it is minimised over classes of isometries. Rote [R91] gives an optimal $O(n \log n)$ algorithm for computing the Hausdorff metric minimised over translations, for objects consisting of sets of points on the real line. [HKS91] also gives efficient algorithms for the Hausdorff metric on sets of points in $\mathbb{R}^2$ and polygons in $\mathbb{R}^2$, minimised over translations.

The Directed Hausdorff Distance

The Hausdorff metric may be defined in terms of the "directed Hausdorff distance" [HKR91], for which the distance between sets P_1 and P_2 is:

$$h(P_1, P_2) = \sup_{p_1 \in P_1} \left\{ \inf_{p_2 \in P_2} \{d(p_1, p_2)\} \right\},$$

So the Hausdorff distance can be expressed as:

$$H(P_1, P_2) = \max\{h(P_1, P_2), h(P_2, P_1)\}$$

h is not a metric, since it is not symmetric. However it may still be used to define geometric concepts, with two alternative kinds of concept class. A concept may either be (for some fixed $C \in \mathcal{X}, r \in \mathbb{R}^+$) all objects $P \in \mathcal{X}$ such that $h(P, C) \leq r$ or all objects $P \in \mathcal{X}$ such that $h(C, P) \leq r$.

The difference between these concepts and those defined using the Hausdorff metric is that they correspond to proximity in scene analysis: we can say that an object is close to a scene if there is some subset of the points composing that scene which is close to the object in the Hausdorff sense.

The minimax metric

This metric was introduced in [ISI89] for the purpose of matching two finite sequences of points, where the number of points in each sequence is the same, and we must minimise (over translation) the maximum distance between pairs of corresponding points in the sequences. (The metric is intended to relate to the problem of optimal positioning of a component in a larger device, where a given set of short connections have to be made.) So if P_1 is the sequence $(x_1, x_2, ..., x_n)$, $P_2 = (y_1, y_2, ..., y_n)$ then

$$M(P_1, P_2) = \min_I \Big\{ \max_{i=1,...n} \{d(x_i, I(y_i))\} \Big\}$$

where I is a class of isometries, usually taken to be translations.

This is clearly quite restrictive in that both sets of points must contain the same number of points, and the points in each set must each be labelled so as to define a correspondence. It can be seen to be a more demanding notion of resemblance than the Hausdorff metric modulo the isometries I. Observe also that the size or complexity of a hypothesis in the associated concept class must be the same as that of the examples.

We also consider the minimax without minimisation over translations (which we will call the "translation-free minimax metric" although the "mini" is inappropriate.) That is:

$$M(P_1, P_2) = \max_{i=1,...n} \{d(x_i, y_i)\}$$

The Fréchet metric

The *Fréchet metric* was introduced in [F06]. This metric is described in [ABW90], as a means to approximate polygons by simpler ones. It can be expressed intuitively by saying that the distance between polygons P_1 and P_2 is, supposing a man to walk around the perimeter of P_1 and a dog to walk around the perimeter of P_2, the length of the shortest leash that can connect them. No backtracking is allowed. Mathematically this becomes: For polygons P_1 and P_2, if $m_1 : [0,1] \longrightarrow P_1$ and $m_2 : [0,1] \longrightarrow P_2$ are monotonic parametrizations of their perimeters, then the distance between P_1 and P_2 is

$$F(P_1, P_2) = \min_{\text{all } m_1, m_2} \Big\{ \max_{0 \le x \le 1} \{d(m_1(x), m_2(x))\} \Big\}$$

Two polygons which are distance d apart using this metric are *a fortiori* at most distance d apart using the Hausdorff metric. It is hard to construct pairs of polygons which this metric deems to be close but are in some informal sense radically different. This metric is shown to be efficiently computable for immobile chains of line segments in [G91]. In fact we can show that it remains computable in polynomial time for movable polygons, using a combination of the method in [G91] with that of [ABB91]. This appears to be a new result. In the interest of coherence, the full proof is not given here, but will be presented in a later paper.

Theorem: *The Fréchet metric, minimised over reflections and rotations, can be computed in polynomial time.*

Proof (sketch): We use the notion of "critical pairs" of points, in [ABB91]. Given two polygons P_1 and P_2 with a Hausdorff distance of d between them, a critical pair is two points, one on each polygon, such that (1) they are Euclidean distance d apart, and (2) one is the closest point on its polygon to the other. Then in [ABB91] it is observed that in a position that minimises the distance between two polygons, there must be either at least three critical pairs, or two critical pairs with the vectors between each pair being equal and opposite.

While there are more ways in which critical pairs can arise for the Fréchet metric, there are still only a polynomial number of possibilities in which three may arise at the same time. For all combinations of these we may test the Fréchet distance between the two polygons using the algorithm of [G91] for computing the Fréchet distance between fixed polygonal lines. □

2.3. Further Remarks

We have seen that under various notions of resemblance it is possible to recognise quickly the extent to which two objects looks similar. We should note that algorithms used in practice for visual recognition are often not based on metrics, but instead rely on identification of "key features/characteristics" ([BG90], chapter 8.) Expressing a geometrical object in terms of some distinguishing features without regard to their relative position may be related to other learning problems considered in the Valiant framework, where the aim is to identify a small number of relevant attributes from a large number of mainly irrelevant ones. If the relative positions of these features is of interest, then the minimax

metric may capture an aspect of this approach, by reflecting a notion that some individual features of an object are readily distinguishable.

Topology-based metrics have also been proposed, but tend to have properties which are not of interest in computational geometry, and may also be hard or impractical to compute.

Finally, we should note that other notions of geometric resemblance have been proposed, which we have not considered here. For example the area of the symmetric difference of two polygons is another fairly natural measure of similarity which we do not consider in the context of learning.

[ACHKM89] develops the following tool, which reflects another notion of resemblance:

Define the "turn function" of a polygon by scaling it to have unit perimeter, and obtain a function $[0,1] \rightarrow [0,2\pi]$ mapping distance around perimeter starting from an arbitrary point p on the polygon, to angle of inclination of the current edge. This is easy to compute, and (modulo the starting-point p) encodes a polygon modulo isometries and dilatations, so that it reflects "shape" rather that "shape + position".

This function can be used to define a metric-based notion of resemblance. A polygon is associated with a step function, unique modulo horizontal shifts, and the distance between two polygons is the minimum distance between their turn functions, using the L_2 norm. This apparently allows non-spherical concepts to be defined in a natural way, as the set of all polygons with turn functions bounded above and below by two given functions.

Chapter Three

Vapnik-Chervonenkis dimension of the Concept Classes

We have seen in the last chapter that all the concept classes under consideration have a polynomial-time membership test, implying that a hypothesis that is not excessively large will be able to classify unknown objects efficiently. Here we show that in addition it only requires a polynomial-sized sample (polynomial in the size of the target concept) to have enough information to extract a good hypothesis, and this is done by showing that the Vapnik-Chervonenkis dimension of these concept classes grows only polynomially in the size of concepts.

In section 3.1 we consider the issues involved in the representation of these concepts (whose length is used as the **size** function.) We then consider in detail an example of one concept class, namely concepts defined by sets of points in the plane, where members of the domain $\mathcal{X}$ are also sets of points in the plane. We give fairly precise upper and lower bounds on the V-C dimension of this concept class, as functions of both concept size and object size. Then in section 3.2 we give a general result which implies that the V-C dimension is polynomial for all the concept classes under consideration.

3.1. Introduction, and an Example

Representational Issues

Geometrical objects are typically represented using tuples of real numbers which give the coordinates of points defining them. An n-gon in the Euclidean plane, for example, may be represented as the $2n$-tuple of x and y coordinates of its vertices. A concept in a class where the radius of concepts is not fixed is representable as the $(2n+1)$-tuple of the vertices of its central polygon and the radius.

The size of an example or concept will be taken to be the number of real numbers used to represent it. This is based on the uniform cost model for real computation, in which arithmetic operations on reals are considered as

elementary and real numbers occupy unit space. This is the model of computation developed in [BSS89]. The model is observed in [V91] to be generally appropriate for learning from continuous domains. The test for membership of an example e (represented by reals $\langle x_1, x_2, ..., x_{n_1} \rangle$) in a concept C (represented by reals $\langle y_1, y_2, ..., y_{n_2} \rangle$) will consist of some boolean-valued formula taking $\{x_1, ..., x_{n_1}, y_1, ..., y_{n_2}\}$ as arguments, or more generally a program taking as input $\{x_1, ..., x_{n_1}, y_1, ..., y_{n_2}\}$ and returning a boolean value. We require the formula/program to be evaluated in polynomial time.

The following results show that the V-C dimension is polynomial for classes for which this test is an exponential-sized formula in the first-order theory of the reals, with bounded quantification depth. If concepts in a class have such a membership test, this could be interpreted as implying that their representation as tuples of reals is a "natural" encoding.

The Effect of Example Size on the V-C dimension

We need to place some limit on the size or complexity of examples of a target concept, for the following reason. If examples may have unlimited size, the V-C dimension of concepts of bounded size may be infinite. The next section of this chapter analyses the V-C dimension of a typical concept class with this property, namely sets of points in the plane under the Hausdorff metric. It will be shown that for sufficiently large examples, the set of concepts defined by just one point may have infinite V-C dimension.

The concept classes under consideration here may be divided into two categories, depending on whether or not a concept is allowed to contain examples of different complexity to that concept, as can happen in comparing two polygons with different numbers of edges. This feature is not present under the minimax metric for example, since it is only possible to compare sets of points containing the same number of points. In this situation, the concept class $\mathcal{C}$ is split into separate classes $\mathcal{C} = \{\mathcal{X}_n, \mathcal{C}_n\}_{n \in \mathbb{N}}$ where $\mathcal{C}_n$ is the class whose concepts are defined by exactly n points. This is the kind of learning (with respect to domain dimension n) indicated in theorem 1.3. In learning the class $\mathcal{C}_n$ a hypothesis cannot come from any other class $\mathcal{C}_{n'}$, $n' \neq n$. Hence uniform learning for each class $\mathcal{C}_n$ is sought, with an algorithm polynomial in n.

If comparisons between figures of different size are allowed then the class $\mathcal{C}$ is more integrated, and Occam-style learning as indicated in theorem 1.4 and 1.5 (with respect to target concept complexity) is possible. This is a property

of all metrics considered in this thesis other than the minimax metric. The time taken to learn a concept of size s should be polynomial in s.

For most concept classes considered in this thesis, examples as well as concepts may have various sizes. Moreover it is meaningful to test whether a concept contains some example of complexity greater than that concept. In this respect the concept classes differ from other continuous (non-discrete) classes studied in the literature, in which members of the domain are usually points in fixed-dimensional space. Hence it is necessary to combine learning with respect to domain dimension with learning with respect to target concept complexity. Lemma 3.1 will show that an appropriate restriction on example complexity reduces the problem to the Occam-style learning of theorem 1.4.

For polynomial learnability where example complexity is not fixed it is clearly necessary to place some polynomial bound $p(n)$ on the size of examples used to learn a target concept of size n. Otherwise it will take too long to read in the examples before even processing them. The imposition of a polynomial bound also (as we shall see) finesses the problem of dependence of the V-C dimension on example size. The learning problem reduces to the problem of learnability with respect to target concept complexity (Occam-style learning.) However for the time being we will retain explicit parameters for both concept complexity and example complexity.

We will use the following notation, in which the stratification structure is augmented with an additional parameter for example size.

Notation: Let $\mathcal{C}_{n_2}^{n_1}$ denote the concept class of concepts of complexity $\leq n_2$ restricted to examples of complexity $\leq n_1$.

Let $\mathcal{C}_{\infty}^{n_1} = \cup_{n_2 \in \mathbb{N}} \mathcal{C}_{n_2}^{n_1}$ and $\mathcal{C}_{n_2}^{\infty} = \cup_{n_1 \in \mathbb{N}} \mathcal{C}_{n_2}^{n_1}$. Then we will see that $\dim \mathcal{C}_1^{\infty} = \infty$ for sets of points in the plane under the Hausdorff metric, and for the same class, that $\dim \mathcal{C}_{\infty}^{1} = \infty$.

The following lemma shows that given the restriction that example complexity should be polynomial in concept complexity, the criterion for sufficient information for learnability is polynomial V-C dimension for sets of concepts of complexity $\leq n$ restricted to examples of size $\leq n$, (ie. $\dim(\mathcal{C}_n^n)$ is polynomial in n.)

Lemma 3.1: *Let $\mathcal{C}$ be a stratified concept class $\mathcal{C} = \cup_{i,j \in \mathbb{N}} \mathcal{C}_j^i$. Then the following are equivalent:*

i.) The V-C dimension of $\mathcal{C}_n^n$ is polynomial in n.

ii.) For any polynomial p, $\mathcal{C}_n^{p(n)}$ is learnable with polynomial sample complexity.

Proof: *i* $\Rightarrow$ *ii*: For (*ii.*) to hold, we require the V-C dimension of $\mathcal{C}_n^{p(n)}$ to be polynomial in n. (We may assume without loss of generality $p(n) \geq n$.)

$\dim(\mathcal{C}_n^{p(n)}) \leq \dim(\mathcal{C}_{p(n)}^{p(n)})$, which from (*i*) is polynomial in $p(n)$, hence polynomial in n.

ii $\Rightarrow$ *i*: Put $p(n) = n$ and this follows from theorem 1.2. □

Where there is no known limit on example size in terms of concept size, the following criterion may be helpful:

These are equivalent to:

iii.) The number of examples required to learn $C \in \mathcal{C}$ is polynomial in the size of C and the size of the largest example in the sample.

Proof: *i* $\Rightarrow$ *iii*: Let n_1 be the maximum example size, n_2 the size of the target concept.

Then it is sufficient to have enough examples to learn $\mathcal{C}_{n_2}^{n_1}$.

$\dim(\mathcal{C}_{n_2}^{n_1}) \leq \dim(\mathcal{C}_{\max\{n_1,n_2\}}^{\max\{n_1,n_2\}})$, hence is polynomial in $\max\{n_1, n_2\}$

iii $\Rightarrow$ *ii*: obviously. □

So for concept classes of the kind considered here, it is sufficient for us to show that the V-C dimension of the concept class of all those concepts of complexity $\leq n$ over objects of size $\leq n$ is polynomial in n.

Example: V-C dimension of Sets of Points in the Plane under the Hausdorff Metric

The purpose of this example is to show (by explicit construction) some fairly precise bounds on the V-C dimension as a function of concept complexity and example complexity, before showing more general polynomial bounds for a wider family of concept classes. It turns out that the V-C dimension for this example is linear in concept complexity and logarithmic in example complexity. This result can be seen as analogous to that of [H88] for "irrelevant attribute" learning. (This is learning conjunctions of length k over n variables, where $n \gg k$. The sample complexity is logarithmic in n.)

A concept in $\mathcal{C}_{n_2}^{n_1}$ is defined by a set S of $\leq n_2$ points, and consists of all collections of $\leq n_1$ points within r of S under the Hausdorff metric. These are

points in the plane, for which the distance between two individual points is the Euclidean norm. The radius r is a constant for all concepts, and all values of n_1, n_2.

We will start by showing that $\dim(\mathcal{C}_1^\infty) = \infty$, that is, even if concepts are only defined by single points, the V-C dimension can still be arbitrarily large if examples may consist of sufficiently many points.

The following construction shows how for any positive integer n, a set of n objects can be shattered if each object has 3.2^{n-3} points. (This is not the best upper bound on the size of objects required, but arises from a construction that can be conveniently illustrated. $n = 4$ in fig. 3.1; each point on an object is labelled with the number of the object it belongs to. The single points defining the shattering set of concepts are located at the six clusters of four points labelled S.)

An object has one point in each of 3.2^{n-3} lines of n points radiating out from a circle of radius $< r$ (shown in fig. 3.1.) The order of the points may be chosen such that any subset T of $\{1, ..., n\}$ forms the labels on the outermost set of $|T|$ points on one of these lines.

In each of these lines, points must be sufficiently close together that an r-circle may contain all clusters except any one, whose points all lie outside the circle. Hence an r-circle can leave out just the outermost i points in any line, for any $i \leq n$. Since any subset of the n objects is the set of points in a sequence lying outermost in some line, then for any such subset, an r-circle exists which contains all points in the n objects except those in that line which belong to that subset only. Hence the set is shattered.

Observe that in the bit model of the real numbers, the number of significant figures necessary to represent these points must increase as n increases.

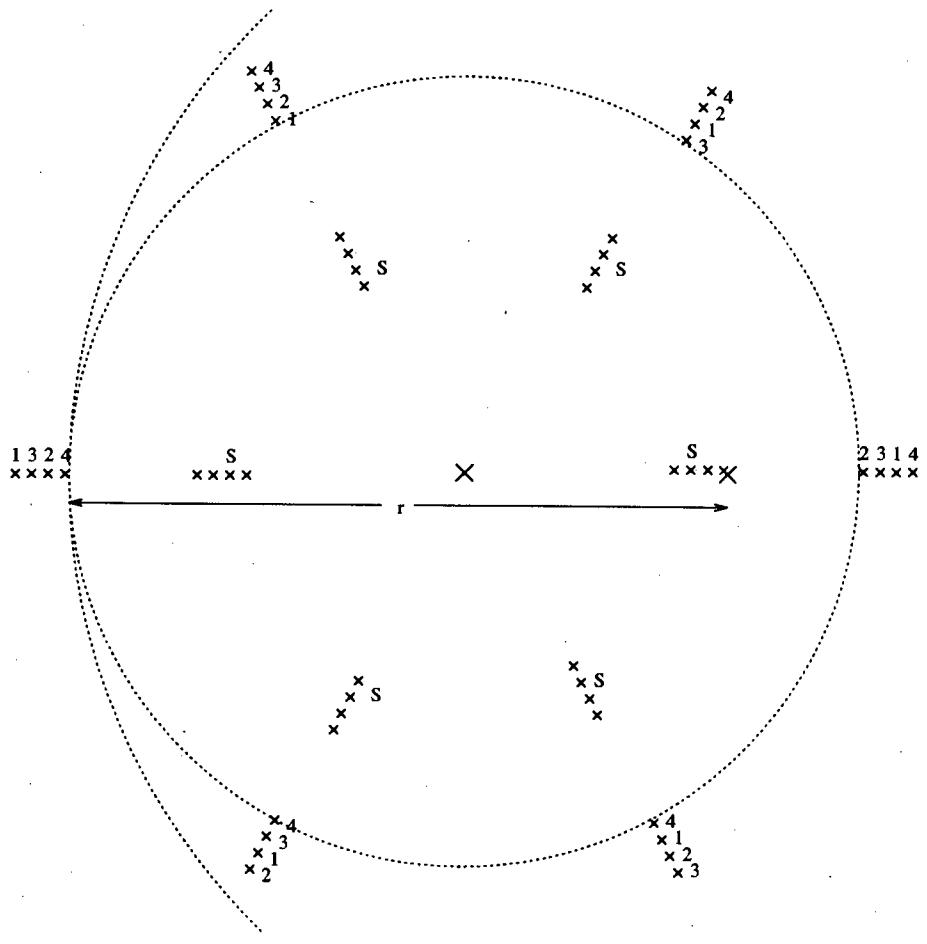

figure 3.1. Construction of n sets of 3.2^{n-3} points shattered by r-balls around single points

This construction shows that $\dim(\mathcal{C}_1^n) = \Omega(\log n)$. The next construction will show that $\dim(\mathcal{C}_n^1) = \Omega(n)$. In particular, for any $n \in \mathbb{N}$ there are n examples (single points) which are shattered by a collection of 2^n concepts each defined by a set of $\leq n$ points.

The n examples are evenly spaced on a circle of radius $< r$. For any subset T of these points, we can construct a set C of $\leq n$ points within Hausdorff distance r of each point in T, and none of the others. This is done by placing a point in C for each point p not in T within r of all points except p. The arrangement of the examples ensures that such a location always exists.

In fig. 3.2, $n = 6$ and the numbered points may be shattered by concepts of radius r defined by subsets of the points labelled S. The point at the centre of the r-circle in the diagram is within r of all points except 1.

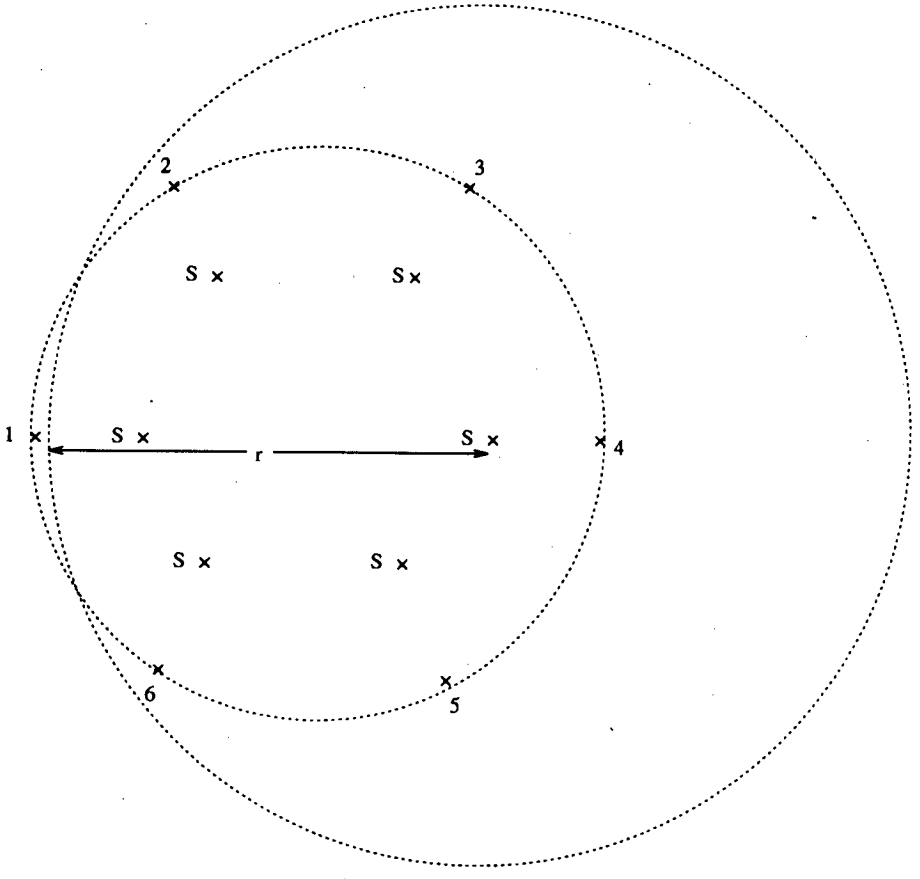

figure 3.2. Construction of n points shattered by 2^n r-balls around sets of n points.

Finally we will show that $\dim(\mathcal{C}_{n_2}^{n_1}) = O(n_2 \log(n_1 n_2))$.

Let $v = \dim(\mathcal{C}_{n_2}^{n_1})$. So v examples of size n_1 (sets of n_1 points) may be shattered by 2^v concepts defined by up to n_2 points. The r-circles around each point in these examples divide the plane into $O\big((vn_1)^2\big)$ distinct regions.* Within each region, two concept points are equivalent in terms of the implications they

* A justification for this claim is given on page 81.

have on membership of any example. Hence there are $\leq \sum_{i=0}^{n_2} \binom{(vn_1)^2}{i} = O((vn_1)^{2n_2})$ distinct concepts available, of which 2^v must shatter the v examples. Hence $2^v \leq (vn_1)^{2n_2}$, i.e. $v \leq 2n_2 \log(vn_1)$, so

$$v \leq 2n_2 \log v + 2n_2 \log n_1$$

$$v \leq n_1 \Longrightarrow v = O(n_2 \log n_1)$$

$$v > n_1 \Longrightarrow v = O(n_2 \log v) \Longrightarrow v = O(n_2 \log n_2)$$

Combining these:

$$v = O\big(n_2(\log n_1 + \log n_2)\big) = O(n_2 \log(n_1 n_2))$$

3.2. A Family of Concept Classes with Polynomial V-C dimension

We will use the following notation: an example $e = \langle x_1, ..., x_n \rangle$, concept $C = \langle y_1, ..., y_n \rangle$, $\mathcal{C}_n$ is defined by a boolean-valued formula Φ_n with free variables $\{x_1, ..., x_n, y_1, ..., y_n\}$, such that $\Phi_n(x_1, ..., x_n, y_1, ..., y_n)$ is true if and only if $e \in C$.

There is no obvious precise criterion for saying what is and is not a "natural encoding" for concept classes of the sort under consideration. One possibility is to express such a criterion in terms of the allowable form of the test for an example e to belong to a concept C. For our purposes this will be taken to mean that Φ_n is expressible as a formula in the first-order theory of the real numbers. Furthermore we require for our proofs Φ_n to have only a constant depth of quantifier alternation, at most exponential length and an exponential bound on the degree of polynomials it contains. Under these constraints the associated concept class $\mathcal{C}_n$ has polynomial V-C dimension. It is an open question to what extent these constraints can be relaxed while retaining the property that the V-C dimension must be polynomial. These conditions rule out such tactics as:

1.) Encoding arbitrarily many real values in a single real number by interleaving the digits in their decimal expansions.

2.) Obtaining a concept class of high V-C dimension whose concepts and examples are parametrised by only one real number, by embedding the real line in a higher dimensional space.

We further show that the concept classes of geometrical objects considered in this thesis satisfy these conditions. Hence by theorem 1.2 there is enough information in a sample of size polynomial in $\epsilon^{-1}, \delta^{-1}$, and n for a good hypothesis to be extracted. In this section we will deal with concept classes with examples constrained to be polynomial in target concept complexity, which lemma 3.1 shows to be equivalent to concept classes parametrised by just target concept complexity. Consequently we will refer to concept classes $\mathcal{C}_n$ parametrised by the target complexity only.

Observation 3.2: We may note, to begin with, that if real numbers are represented using some fixed finite number d of bits of precision, then a simple counting argument can be used to show that for any membership test Φ_n, the V-C dimension of the concept class under consideration is polynomially bounded. A concept in $\mathcal{C}_n$ only needs dn bits in its representation, where n is the number of real numbers in the representation of a concept. Hence there are only 2^{dn} concepts and the crude upper bound of dn for the V-C dimension can be used.

Note that for "real numbers" of fixed precision tricks (1) and (2) above for unnatural encodings would anyway not work. Neither would the constructions in the previous section, since they require real numbers represented in this way to have non-constant bit complexity, in fact $n \log n$ in the first construction. We now show that the V-C dimension is polynomially bounded, even without this restriction.

Our main result is the following:

Theorem 3.3: *Let $\{\mathcal{C}_n : n \in \mathbb{N}\}$ be a set of concept classes indexed by the natural numbers having the property that concepts and examples in $\mathcal{C}_n$ are represented by n real values. Suppose the membership test of a given example in a given concept can be expressed as a formula Φ_n in the first-order theory of the real numbers with $2n$ free variables representing a concept and example in $\mathcal{C}_n$ and fixed quantification depth, whose polynomials have degree exponential in n and the length of Φ_n is exponential in n.*

Then the V-C dimension of $\mathcal{C}_n$ is polynomial in n.

The proof uses a quantifier elimination scheme of Renegar [R92] to reduce Φ_n to a boolean formula Ψ_n whose atomic predicates are polynomials in the free variables of Φ_n. Then we use an upper bound implied by Milnor in [M64] (of which a simpler proof is given in [R92]) on the number of consistent sign

assignments to a set of multivariate polynomials. Before proving the theorem we state these two results.

Using the notation of [R92], a formula in the first-order theory of the reals has the general form:

$$(Q_1 x^{[1]} \in \mathbb{R}^{n_1}) \ldots (Q_\omega x^{[\omega]} \in \mathbb{R}^{n_\omega}) P(y, x^{[1]}, \ldots, x^{[\omega]})$$

where the Q_i are quantifiers, $x^{[i]}$ is a vector of n_i real quantified variables, and $y = (y_1, \ldots, y_l)$ is a vector of real free variables.

P consists of a boolean formula $\mathbb{P}$ having m atomic predicates consisting of polynomial equalities or inequalities of degree bounded by d (whose variables are in y or the $x^{[i]}$.)

Theorem 3.4:[R92] *There is a quantifier-elimination procedure which requires only* $(md)^{2^{O(\omega)} l \Pi_k n_k}$ *operations and* $(md)^{O(l+\Sigma_k n_k)}$ *calls to* $\mathbb{P}$.

The algorithm constructs a quantifier-free formula of the form

$$\bigvee_{i=1}^{I} \bigwedge_{j=1}^{J_i} (h_{ij}(y) \, \Delta_{ij} \, 0),$$

where

$$I \leq (md)^{2^{O(\omega)} l \Pi_k n_k},$$

$$J_i \leq (md)^{2^{O(\omega)} \Pi_k n_k},$$

the degree of h_{ij} *is at most* $(md)^{2^{O(\omega)} \Pi_k n_k}$, *and* Δ_{ij} *represents one of the symbols* $\{<, \leq, =, \neq, \geq, >\}$

The theorem of [M64, R92] has been used in other works in complexity theory to establish upper and lower bounds. Its statement requires the following definition:

Definition: Let $\{g_1, \ldots, g_m\}$ be a finite set of m polynomials in n real variables. The *connected sign partition* of g_i, denoted $CSP\{g_i\}$ is the set of maximal connected components of $\mathbb{R}^n$ having the property that for any pair $\mathbf{x}$ and $\mathbf{y}$ of points in one component in $\mathbb{R}^n$, the sign of $g_i(\mathbf{x})$ is the same as the sign of $g_i(\mathbf{y})$ for $i = 1, \ldots, m$. (The sign of a polynomial is set to $1, 0$, or -1 depending on whether it is positive, zero or negative.)

Theorem 3.5: [M64, R92] *The number of elements of* $CSP\{g_i\}$ *has the upper bound* $(md)^{O(n)}$.

Proof of theorem 3.3: We use the quantifier elimination scheme of [R92] to reduce Φ_n to the simpler quantifier-free form. In this information-theoretic setting, it is the form of the quantifier-free formula rather than the time taken to construct it which is of importance.

The bounds on I and J_i in theorem 3.4 show that the number of polynomials acting as atomic predicates is superexponential in the depth of quantifier alternation (which we require to be constant for a concept class) and exponential in the number of quantified variables and the number of free variables.

Hence this procedure leads to an exponential blowup in the size of the formula Φ_n used to classify examples according to a concept. Let Ψ_n be the resulting formula, a DNF formula whose predicates are polynomial equalities and inequalities. Ψ_n has free variables $\{x_1, ..., x_n\}$ representing an example and $\{y_1, ..., y_n\}$ representing a concept.

Let $\mathcal{C}_n$ have V-C dimension $v(n)$. Let $\{e_1, ..., e_{v(n)}\}$ be a shattered set of examples. For each e_i let $\Psi_n(e_i)$ denote the formula in $\{y_1, ..., y_n\}$ obtained by substituting the values e_i has for $\{x_1, ..., x_n\}$ in Ψ_n.

Let $s(n)$ be the length of Φ_n. Hence Φ_n contains $O(s(n))$ polynomials. Let $d(n)$ be their degree. Let ω be the number of quantifiers in Φ_n, for all n.

The formula $\Psi_n(e_i)$ contains $(sd)^{O(n^\omega)}$ polynomials. They have degree $(sd)^{O(n^\omega)}$. Let S be the union over $i = 1, ..., v(n)$ of all polynomials contained in $\Psi_n(e_i)$. So $|S| \leq v(sd)^{n^{O(1)}}$.

For $\{e_1, ..., e_{v(n)}\}$ to be shattered, all the polynomials contained in S must be able to take $2^{v(n)}$ different sign assignments.

From theorem 3.5, the number of sign assignments is bounded above by $(md)^{O(n)}$ where m is the number of polynomials, d is their degree, and n is the number of variables. This upper bound is:

$$\left(v(sd)^{O(n^\omega)}.(sd)^{O(n^\omega)}\right)^{O(n)}$$

$$= \left(v(sd)^{O(n^\omega)}\right)^{O(n)}$$

Hence,

$$2^v \leq \left(v(sd)^{O(n^\omega)}\right)^{O(n)}$$

$$v \leq O(n)\log\left(v(sd)^{O(n^\omega)}\right)$$

$$v \leq O(n)\left(\log v + O(n^\omega)\log(sd)\right)$$

So $v(n)$ must be polynomially bounded in n. □

Observations 3.6: 1.) Since the above upper bound for the V-C dimension is only logarithmic in the length of Φ_n and the degree of the polynomials, these may be exponential. The V-C dimension is at least linear in the number of free variables defining a concept, corresponding with observation 3.2. The upper bound is exponential in the quantification depth, and it is an open question whether this dependence can be reduced.

If the formula is given another parameter for example complexity ($\Phi^{n_1}_{n_2}$ defines $C^{n_1}_{n_2}$) then the V-C dimension will only be logarithmic in example complexity.

2.) A corollary of theorem 3.3 is that if Φ_n is a class of polynomial-sized arithmetic circuits whose output is a boolean function of the signs of the values at the nodes, then the V-C dimension is also polynomial. This is because the degree of the polynomials which it may calculate, and their length when expressed using standard arithmetic operators, are exponentially bounded.

3.) Theorem 3.3 is proved for concepts and examples represented by exactly n real values. It can still take account of examples of varying sizes since geometric objects may be "padded out" with extra points coincident with other points, without affecting their distance from other objects.

We will conclude by applying this result to some examples of geometrical concept classes described in the last chapter. We will start with the example already considered, of point sets under the Hausdorff metric.

Corollary 3.7: *Let C_n be the concept class whose concepts are sets of arrangements of n points which are within some fixed distance r from some fixed arrangement of n points, under the Hausdorff metric.*

Then C_n has V-C dimension polynomial in n.

Proof: Represent an n-point set P by the $2n$-tuple of the coordinates of its points in some order.

$$P = \langle x_1, y_1, x_2, y_2, ..., x_n, y_n \rangle$$

where x_i, y_i are the coordinates of the ith point of P.

Represent a concept C in $\mathcal{C}_n$ by the $(2n+1)$-tuple of the coordinates of the points of the central set* in some order, followed by the radius r.

$$C = \langle c_1, d_1, c_2, d_2, ..., c_n, d_n, r \rangle$$

where c_i, d_i are the coordinates of the ith vertex of the central polygon.

Then $P \in C$ is expressible as:

$$\bigwedge_{i=1}^{n} \left\{ \bigvee_{j=1}^{n} \{ d((x_i, y_i), (c_j, d_j)) \leq r \} \right\} \wedge \bigwedge_{i=1}^{n} \left\{ \bigvee_{j=1}^{n} \{ d((c_i, d_i), (x_j, y_j)) \leq r \} \right\}$$

where $d(x, y)$ is the Euclidean distance between x and y. This is expressible as a polynomial of degree 2 in the coordinates of x and y. When this expression is expanded out in the desired form it has length $O(n^2)$, polynomial in n, and is in fact quantifier-free. □

Some of the concept classes defined in the last chapter involved minimisation over classes of linear translations. The natural way of expressing the test of two objects being close together under such conditions involves quantified variables, that is, there exists some position for which they are close.

Corollary 3.8: *Define $\mathcal{C}_n$ as before, but with the distance between two sets of n points taken to be the Hausdorff distance minimised over all isometrical linear transformations of the point sets.*

Then $\mathcal{C}_n$ has V-C dimension polynomial in n.

Proof: As before, represent an n-point set P by the $2n$-tuple of the coordinates of its points in some order.

$$P = \langle x_1, y_1, x_2, y_2, ..., x_n, y_n \rangle$$

where x_i, y_i are the coordinates of the ith point of P.

Represent a concept C in $\mathcal{C}_n$ by the $2n+1$-tuple of the coordinates of the points of the central set in some order, followed by the radius r.

$$C = \langle c_1, d_1, c_2, d_2, ..., c_n, d_n, r \rangle$$

* ie. the set of points which define the geometrical object at the centre of C

where c_i, d_i are the coordinates of the ith vertex of the central point set.

Then $P \in C$ holds provided that there exists some isometry t such that $t(P)$ is within Hausdorff distance r of the centre of C. This is expressible as:

$$\exists(x,y,s,c)\Big\{(s^2+c^2=1)$$
$$\wedge \bigwedge_{i=1}^{n}\Big\{\bigvee_{j=1}^{n}\big\{d\big((x_ic-y_is+x, x_is+y_ic+y),(c_j,d_j)\big) \le r\big\}\Big\}$$
$$\wedge \bigwedge_{i=1}^{n}\Big\{\bigvee_{j=1}^{n}\big\{d\big((c_i,d_i),(x_jc-y_js+x, x_js+y_jc+y)\big) \le r\big\}\Big\}\Big\}$$

where s and c are the sine and cosine of the angle through which P is rotated before being translated by (x,y).

When this expression is expanded out in the desired form it has length $O(n^2)$, with just one quantifier and four quantified variables. □

Corollary 3.9: *Let $\mathcal{C}_n$ be the concept class whose concepts are sets of n-gons which are within some fixed distance r from some fixed polygon, under the Hausdorff metric.*

Then $\mathcal{C}_n$ has V-C dimension polynomial in n.

Proof: Represent an n-gon P by the $2n$-tuple of its vertices in clockwise order from some arbitrary starting vertex.

$$P = \langle x_1, y_1, x_2, y_2, ..., x_n, y_n\rangle$$

where x_i, y_i are the coordinates of the ith vertex of P.

Represent a concept C in $\mathcal{C}_n$ by the $2n+1$-tuple of the vertices of its central polygon in clockwise order from some arbitrary starting vertex, followed by the radius r.

$$C = \langle c_1, d_1, c_2, d_2, ..., c_n, d_n, r\rangle$$

where c_i, d_i are the coordinates of the ith vertex of the central polygon.

Then the following formula states that the polygon P is in concept C.

$$\exists(x,y,s,c)\Big\{(s^2+c^2=1)$$

$$\wedge\forall(p_x,p_y)\exists(c_x,c_y)\big\{(p_x,p_y) \text{ is on polygon } P$$

$$\Rightarrow \big((c_x, c_y) \text{ is on central polygon of } C \wedge d((p_x, p_y), (c_x, c_y)) \leq r\big)\big\}$$

$$\wedge \forall (c_x, c_y) \exists (p_x, p_y) \big\{(c_x, c_y) \text{ is on polygon } C$$

$$\Rightarrow \big((p_x, p_y) \text{is on central polygon of } P \wedge d((p_x, p_y), (c_x, c_y)) \leq r\big)\big\}\Big\}$$

The statement that a point with given coordinates is on a given n-gon means that the point is a convex combination of two consecutive vertices of the n-gon, which is a boolean combination of linear inequalities in the coordinates, of length $O(n)$. When expanded out, this formula is of length polynomial in n and it has a constant number of quantified variables, so it is of the required form. □

Chapter Four

Learnability and Non-learnability Results

We have seen a variety of metrics on geometrical figures which are easy to compute and which express notions of similarity of shape. In general the associated concept classes have polynomial V-C dimension, which means that it is sufficient for learning to find a hypothesis which is consistent with the given sample, and the complexity of the hypothesis should be polynomial in the complexity of the target concept and examples, and sublinear in the sample size.

In this chapter we show that for most versions of this problem, the task of finding such a consistent hypothesis, which we call the Consistent Hypothesis Problem (CHP) is NP-complete, and that it often does not help to allow a hypothesis in the concept class to have greater complexity than the target concept (for Occam learning.) The completeness results do in fact hold even if no *a priori* limit is placed on the hypothesis complexity. Most of these proofs are done using a reduction from CNF. In the first section we give the results for a one-dimensional version of the problem, in which objects are sets of points on the real line, and we then show how these results extend to two-dimensional objects. Note that the diagrams in this and the next chapter are explained in the index on page 98.

Before we proceed to show how to construct instances of the CHP from CNF formulae, we will show that the reductions do not depend on exact real arithmetic, which justifies the construction of such problem instances in terms of exact real numbers. The following observation shows that the reductions are robust under some error tolerance ϵ.

Theorem 4.1: *In a reduction from an NP-complete problem to an instance of a CHP involving real values, there exists some error ϵ such that all the real values may be perturbed by up to ϵ while still preserving the encoding of the NP-complete problem.*

Proof: We assume that the spheres are closed, so that a concept is all objects $\leq r$ from some fixed $c \in \mathcal{X}$.

Let $\epsilon = \frac{1}{2}[min_{n \in NEG}\{d(n,c)\} - r]$

Then putting $r = r + \epsilon$, all positive examples are still positive, and all negative examples are still negative.

A similar argument holds for open spheres. □

4.1. Non-learnability under Variants of the Hausdorff Metric

We start by showing the basic result that sets of points on the line which are similar under the Hausdorff metric are hard to learn. Here the problem is to learn a concept class of given radius, whose objects are immobile. This result will then be extended via reductions from this learning problem to the problem of learning movable objects (under translations along the line) and for finding a hypothesis of unrestricted radius.

Theorem 4.2: *A concept class of fixed radius r defined by the Hausdorff metric on immobile sets of points on the real line has an NP-complete CHP.*

Proof: By reduction from CNF, the problem of finding a satisfying assignment to a CNF formula.

Let Φ be a CNF formula over variables $\{v_1, ..., v_n\}$ with k clauses. We reduce the problem of finding a satisfying assignment for Φ to an instance of the above CHP with $2n + k + 2$ examples each with up to $n + 1$ points.

The general idea is that $2n + 2$ examples will force any consistent hypothesis to be of a form that encodes the values of n boolean variables, and then each clause in the formula is translated into an example which forces the encoding of the variables to satisfy that clause.

Include one positive example with n points spaced more than $2r$ apart. (see e_1 in fig. 4.1, where the spacing is $3r$). Then include a second positive example consisting of the first one shifted r to the right. (e_2 in fig. 4.1).

Between them these examples constrain the points composing a consistent hypothesis to lie within the r-intervals bounded above and below by corresponding points in e_1 and e_2. There must be at least one point in each interval. A hypothesis must take the form of h_1 in fig. 4.1.

A further n negative examples each with $n + 1$ points constrain the points in each interval to lie on one side of it. To apply this further constraint on the

first (leftmost) interval in h_1, two points in a negative example are placed just outside this interval on each side of it, and one point is placed within each of the other $n-1$ intervals. (example e_3 in fig. 4.1.)

The only way in which this example can be more than r away from a hypothesis is for the point(s) in the first interval to lie at one end of it. For then the point on example e_3 by the other end will be more than r away from the nearest point on the hypothesis, making the example negative, as required. The hypothesis now takes the form of h_2, with containment of a point in the hypothesis by the resulting pair of subintervals created governed by an exclusive-or relationship.

n examples, e_3—e_{n+2} in fig. 4.1, treat all the intervals in this way. Now the positions of points in each interval pair (ie. on right/left of centre) naturally encode the value of a boolean variable. (see h_3 in fig. 4.1)

In the ith interval pair interpret a point in the left subinterval as an assignment true to variable v_i and a point on the right as the assignment false. A further k negative examples encode the formula itself, and force the positions of points in intervals in a consistent hypothesis to encode a satisfying assignment.

Each clause becomes one (negative) example. If the clause contains v_i, place a point slightly to the right of the ith interval, and if it contains $\neg v_i$ slightly to the left of the ith interval. If it does not contain v_i place a point in the middle of the ith interval. This forces the assignments represented by a consistent hypothesis to satisfy that clause, and so a hypothesis consistent with all of them would satisfy their conjunction. An example e_{n+3} of the object (a negative example) corresponding to the clause $v_1 \vee v_2 \vee \neg v_3$ is given in fig. 4.1.

□

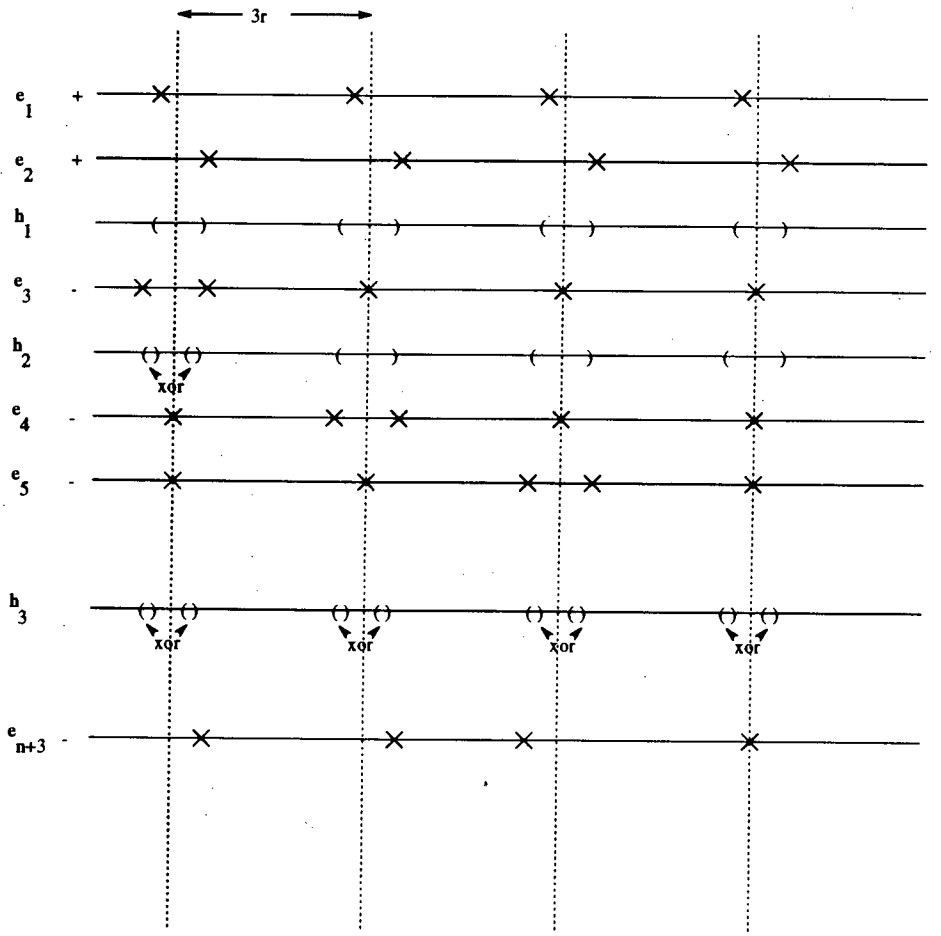

figure 4.1.

The basic idea for extending this result to movable objects is that in the case of learning such objects we can append some recognisable feature to each object so that their positions are constrained with instances of this feature being aligned, thus fixing their position relative to each other. This general method can be used to extend hardness results in other related learning problems.

Corollary 4.3: *Since the consistent hypothesis problem is hard for "fixed" examples (concepts defined by the standard Hausdorff metric), it is hard for movable examples (concepts defined using the minimum Hausdorff distance).*

Proof: We reduce the CHP in the fixed case to the CHP for translatable examples. An instance of the problem (with objects having up to n points) consists of a set $\{e_1, e_2, ..., e_m\}$ of positive and negative examples, with each e_i naturally represented by an n-tuple of real numbers:

$$e_i = \langle x_{i,1}, x_{i,2}, ..., x_{i,n} \rangle$$

We can assume that these points are in order of increasing size, and that where an object has fewer than n points it is padded out using coincident points (leaving its geometrical form unaffected).

Let m be the smallest number occurring in any of these n-tuples. Let r be the radius of the concept class.

Assume w.l.o.g. e_1 is a positive example. If there are no positive examples then the consistent hypothesis problem is trivial.

Replace the first example e_1 by two (positive) examples e'_0 and e'_1 where:

$$e'_0 = \langle m - 10r, x_{1,1}, x_{1,2}, ..., x_{1,n} \rangle$$

$$e'_1 = \langle m - 12r, m - 8r, x_{1,1}, x_{1,2}, ..., x_{1,n} \rangle$$

Replace other examples e_i by

$$e'_i = \langle m - 12r, m - 8r, x_{i,1}, x_{i,2}, ..., x_{i,n} \rangle$$

where e'_i is positive if and only if e_i is.

It is claimed that for this new problem, any consistent hypothesis must be centred around an object which takes the form $\langle m - 11r, m - 9r, C \rangle$ where $\langle C \rangle$ is the centre of a consistent hypothesis for the original (fixed object) problem.

For $i > 1$, any displacement of example e'_i relative to e'_0 will cause its Hausdorff distance from e'_0 to exceed $2r$.

Since a hypothesis is a sphere of diameter $2r$ all positive examples must be fixed relative to e'_0. A negative example must have Hausdorff distance $> r$ from the centre of the hypothesis in all positions along the real line. This is guaranteed by the construction if the example is shifted relative to e'_0, so it only constrains the form of the hypothesis when it is not displaced relative to e'_0.

The hypothesis is consequently determined by examples without translations, so its form is the same as that for the original problem. □

Theorem 4.2 can also be extended to the case where instead of finding a hypothesis of given radius we do not restrict the radius of a consistent hypothesis being sought.

This implies that the larger class of spheres of any radius centred at sets of points in the line under the Hausdorff metric is hard to learn. It is shown by appending a "radius-fixing" mechanism to a positive example, so that a consistent hypothesis is forced to have some given radius, thus reducing the problem to that of learning a concept class of given radius.

Corollary 4.4: *It is hard to find a consistent hypothesis of any radius (still using the standard Hausdorff metric).*

Proof: We reduce the problem of finding a hypothesis of fixed radius to that of finding a hypothesis of arbitrary radius as follows. Let e_1 be a positive example in an instance of the first problem, and let r be the (given) radius. (Note that the problem is trivial if there are no positive examples.)

Replace e_1 by $e'_1, e'_2, e'_3, e'_4, e'_5$, each of which is generated by appending to e_1 extra points strictly more than $2r + \epsilon$ (where ϵ is a positive real number, which will be an upper bound on the extent to which the radius of a consistent hypothesis can differ from r) to the right of the rightmost point of all examples. e'_1 and e'_2 are positive examples, the rest are negative.

Let p be a position (expressed as a real value) more than $3r + \epsilon$ to the right of any example in the fixed-radius problem. e'_1 has one extra point at $p - 1$ and e'_2 has one extra point at $p + 1$. e'_3 has extra points at $p - 1$ and $p - (1 + \epsilon)$ and e'_4 has extra points at $p + 1$ and $p + (1 + \epsilon)$. e'_5 has extra points at $p - 1$, $p - (1 + \epsilon)$, $p + 1$, $p + (1 + \epsilon)$. (See fig. 4.2.)

The effect of the examples is to force a consistent hypothesis to have radius within ϵ of r and to have a point situated within ϵ of p.

All other examples in the instance of the original problem also have an extra point appended at position p. Hence a consistent hypothesis of radius r can now be obtained for the original problem by removing the additional point at position p.

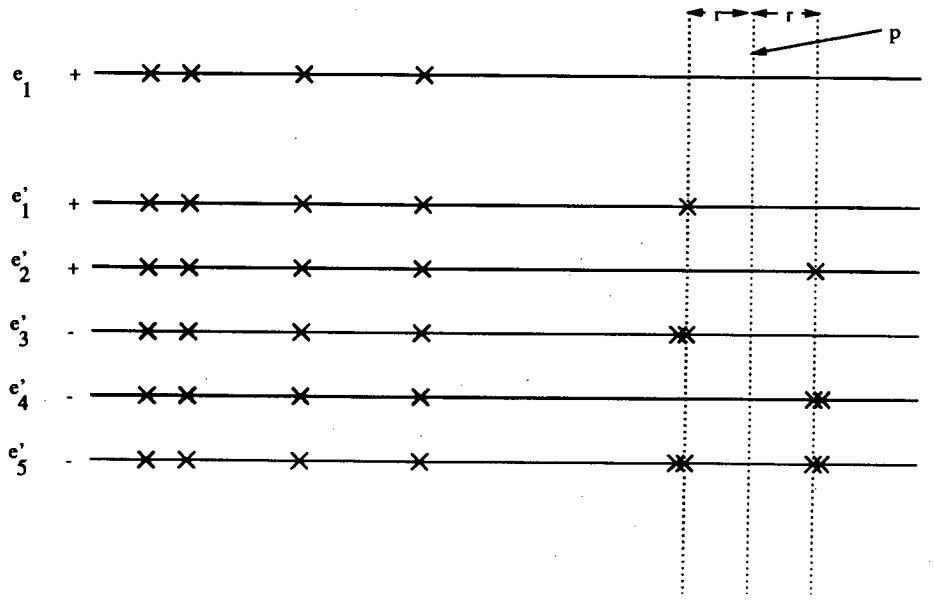

figure 4.2.

Why this works:

e'_1 and e'_2 between them imply that the radius is at least r.

e'_1 and e'_3 between them imply that there is a point on the hypothesis to the right of position $p - r$ whose distance from that position lies between the radius minus ϵ and the radius.

e'_2 and e'_4 between them imply that there is a point on the hypothesis to the left of position $p + r$ whose distance from that position lies between the radius minus ϵ and the radius.

e'_5 implies that the distance between one of the points on e'_5 and the closest point implied by the above examples exceeds the radius, which is only possible if the radius is less than $r + \epsilon$.

Hence there is a point or points within ϵ of p and the radius is within ϵ of r. □

Using a combination of these two techniques we obtain similarly:

Corollary 4.5: *It is hard to find a consistent hypothesis of any radius under the Hausdorff metric modulo translations.*

4.2. Learnability under the Directed Hausdorff distance

Recall that the directed Hausdorff distance is defined to be

$$h(P_1, P_2) = \sup_{p_1 \in P_1} \{ \inf_{p_2 \in P_2} \{ d(p_1, p_2) \} \}$$

where d is some distance norm on individual points. P_1 is assumed to be a component of a scene, and P_2 is assumed to be a scene.

Then a concept is defined in one of two ways, depending on which way this asymmetrical distance function is directed. A concept under the directed Hausdorff metric will usually be taken to mean $\{S : h(T, S) \leq r\}$. Hence a concept is the set of all scenes containing component T. Note that under this definition a component of a scene may have greater complexity (number of points comprising it) than the scene itself. We must make sure that there is an Occam style $n^c m^\alpha$ bound on the complexity of a hypothesis scene.

Under this new weaker notion of similarity we have a positive learning result, namely that a concept class of this form, for sets of points in the real line, is Occam learnable. However the problem is hard if the metric is reversed, ie. once again the CHP is NP-complete for learning concepts of the form $\{S : d^{\leftarrow}(S, C) \leq r\}$ (ie. learning scenes from components).

The algorithm that follows uses a basic greedy set cover procedure in generating points for the hypothesis that account for as many as possible of the negative examples, repeating until all negative examples have been explained (that is, the hypothesis is consistent with them all). These points are found subject to the constraint that they are consistent with all positive examples. To show that not too many hypotheses are needed (ie. the Occam's Razor criterion is satisfied) the following technical lemma will be useful.

Lemma 4.6: Occam Hypothesis from Greedy Set Cover.

For some $c \in \mathbb{N}$, $0 \leq \alpha < 1$

$$log_{1+p(n)^{-r}} m = O(n^c m^\alpha)$$

where p *is a positive valued polynomial.*

Proof: This is equivalent to:

$$log_{1+p(n)^{-r}} m \leq k n^c m^\alpha$$

for sufficiently large n, m, some constant k. Equivalently:

$$m \leq \left(1 + p(n)^{-1}\right)^{kn^c m^\alpha}$$

We now note that:

$$(1 + p(n)^{-1})^{kn^c m^\alpha} = 1 + kn^c m^\alpha p(n)^{-1} + \frac{kn^c m^\alpha (kn^c m^\alpha - 1)}{2} p(n)^{-2}$$
$$+ \frac{kn^c m^\alpha (kn^c m^\alpha - 1)(kn^c m^\alpha - 2)}{6} p(n)^{-3} + \ldots$$

For $c \geq \deg(p), \alpha = \frac{1}{2}$, there exists k such that for sufficiently m, n, the third term in the expansion exceeds m.

Note that α can be made arbitrarily small by considering higher degree terms. This allows the following observation to be made, which will be used later:

$$\left(log_{1+p_1(n)^{-1}} m\right).\left(log_{1+p_2(n)^{-1}} m\right) = O(n^c m^\alpha)$$

So a hypothesis whose size is the product of two such functions is also an Occam hypothesis, and the value of α can still be arbitrarily small. □

We claim that the following algorithm is an Occam algorithm for learning (in one dimension) components from scenes. (That is, if a concept is defined by a set S of points, then point sets in $\mathcal{X}$ belong to the concept if and only if some subset of those points is close to S under the Hausdorff metric.)

Algorithm: FIND-COMPONENT

Let $POS = \{e_1, ..., e_m\}$
Let $NEG = \{e'_1, ..., e'_m\}$
For $1 \leq i \leq m$ let $e_i = \langle x_{(i,1)}, x_{(i,2)}, ...x_{(i,n)} \rangle$
Let $S_i = \{x \in \mathbb{R} : x$ is within r of some point in $e_i\}$
Let $S'_i = \{x \in \mathbb{R} : x$ is within r of some point in $e'_i\}$
{*S_i is a union of at most n intervals.*}
$I := \cap_{i=1}^{m} S_i$
{*I is a union of $O(mn)$ intervals, and contains the points comprising the centre of the target concept* }
Let $N := \{\mathbb{R} - S'_i : 1 \leq i \leq m\}$
{*Greedy set cover for negative examples:* }
repeat
find a point $p \in I$ lying in the maximum possible number of members of N
add that point to centre of hypothesis
$N := N -$ members of N containing p
until $N = \varnothing$

Proof of properties of the algorithm:

Any consistent hypothesis must lie within the r-ball around each positive example. Hence it must lie within their intersection I.

Also a consistent hypothesis must not lie within r-balls around any of the negative examples. Hence there must be at least one point in the hypothesis in each member of N. These are necessary and sufficient conditions for a consistent hypothesis.

The algorithm clearly terminates and produces a hypothesis that satisfies these criteria, hence it is consistent. We need to check that it is an Occam algorithm, ie. poly-time and producing a hypothesis of size (number of points) of order $n^c m^\alpha$ $0 \leq \alpha < 1$, $c \in \mathbb{N}$, where m is the size of the sample and n the size of the target concept.

Let T be the set of n points defining the target concept. Since there are n points in I (namely the set T) that account for every negative example, the point chosen at each iteration of the loop will reduce the size of N by at least $|N|/n$. Hence if we increase the sample size by a factor of $(n+1)/n$ we increase the size of the hypothesis by a constant. So the size of the hypothesis $= O(log_{\frac{(n+1)}{n}} m)$, which by lemma 4.6 implies that it is an Occam hypothesis.

Time taken by the algorithm:

Within each execution of the main loop p can be found in polynomial time by exhaustive search of intervals between boundary point of members of N'. Hence the whole algorithm is polynomial time, hence it is an Occam algorithm.

(Note that the runtime will still be polynomial in more than one dimension, but the order of growth will be higher.)

Theorem 4.7: *The CHP is hard when the directed Hausdorff metric is reversed.*

Proof: In this alternative viewpoint, concepts are scenes and positive examples must resemble components of them under the Hausdorff metric. This implies that the superimposition of all positive examples in a sample must be within radius r of a component of any consistent hypothesis under the Hausdorff metric. Consequently only one positive example should be necessary in the following reduction.

We can reduce CNF to the consistent hypothesis problem as follows. A CNF formula with k clauses over n variables becomes an instance of the CHP with $n + k + 1$ examples each defined by up to n points.

Take a single positive example, which has one point for each variable, spaced more than $2r$ apart (e_1 in fig. 4.3, which has spacing of $3r$ as in fig. 4.1.) Then a hypothesis must contain a point in each interval of size $2r$ with the points of e_1 at their centres (bracketed intervals in h_1).

Then include a negative example consisting of just two points $r/2$ to the right and left of the position of the leftmost point of e_1. For this example to be negative we require all points in the leftmost interval in the hypothesis to lie to one side of it, so that the hypothesis does not contain this example. $n - 1$ similar examples applied to the other intervals divide them each into two subintervals whose containment of points in the hypothesis is governed by the same exclusive-or relationship.

Having encoded a set of n boolean values we can encode a disjunction using a negative example in much the same way as previously. The negative example e in fig. 4.3 encodes the disjunction $v_1 \vee v_2 \vee \neg v_3$, as before.

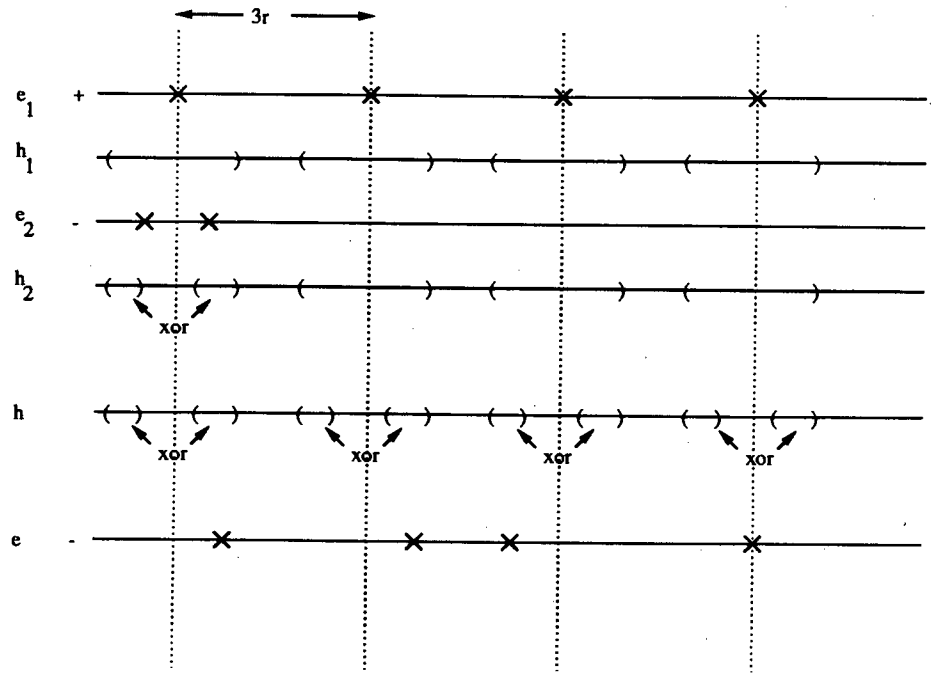

figure 4.3.

Theorem 4.8: *The CHP is hard when the directed Hausdorff distance is minimised over translations.*

Proof: Again, encode CNF as the consistent hypothesis problem. A conjunction of k clauses over n variables will become an instance of the CHP with $2n + k + 4$ examples defined by up to $4n$ points.

For an n-variable CNF formula, we begin the encoding with two positive examples (e_1 and e_2 in fig. 4.4) which (in conjunction with subsequent examples) will have only one position relative to each other for which a consistent hypothesis can exist.

Let S be a tuple of n points spaced out at intervals of at least $7r/2$. Let $m = \min_{s \in S} s$, $M = \max_{s \in S} s$. Let x be a real number such that $m > x + 9r$. Define e_1 and e_2 by positive examples

$$e_1 = \langle x, x + 7r, S \rangle$$

$$e_2 = \langle x + 2r, x + 5r, S \rangle$$

(Note that all tuples of numbers in this translation-minimised version of the problem are equivalent modulo addition of a constant to each number in the tuple.)

Assuming that there must be points in a hypothesis located at $x + r$ and $x + 6r$, then e_1 and e_2 are constrained to be immobile relative to each other. We have that all points in a consistent hypothesis must be located at $x + r$, $x + 6r$, or in the n $2r$-intervals centred at the points of S.

A further $n + 2$ negative examples actually force each of these $n + 2$ disjoint locations to be occupied. These are the $n + 2(n + 1)$-tuples formed by removing one point from $\langle x, x + 7r, S \rangle$ ($e_3 - e_{n+2}$ in fig. 4.4). Note that e_3 and e_4 force a consistent hypothesis to include points within r of the leftmost points of e_1, which must consequently be situated at $x + r$ and $x + 6r$, since these are the only positions which make it consistent with e_2. These points cause e_1 and e_2 to be locked together, since there is only one relative position (above) in which they satisfy this hypothesis.

The hypothesis takes the form h_1 in fig. 4.4.

A further n positive examples are used to split each $2r$-interval into two exclusive-or subintervals.

These are of the form $\langle x + r, x + 6r, T \rangle$ where $\max_{s \in T} s = M + (5r/2)$, $\min_{s \in T} = m - (5r/2)$, and all points in $[m, M]$ are within r of some point in T, with the exception of one of the $2r$-intervals containing a hypothesis point. Each end of that interval is $3r/2$ from the nearest point in the example. (Hence a $5r$-interval is centred around the hypothesis interval.)

These divide each interval into exclusive-or subintervals, and are the only ones to use the ability of objects to move along the line. They can move up to r in either direction, and in fact must be moved at least $r/2$ in one direction to be consistent with the form of the hypothesis. If the example is moved to the right (resp. left) then the point(s) in the singled-out interval must be on the left (resp. right) $r/2$ of that interval.

This results in a hypothesis of the form h_2 (fig. 4.4) ie. a set of n pairs of intervals for which containment of a hypothesis point is governed by an exclusive-or relationship, thus encoding the values of boolean variables. Hence as before a negative example can naturally encode a disjunction of variables and their negations. For example e_{n+4} (fig. 4.4) encodes $v_1 \vee v_2 \vee \neg v_3$.

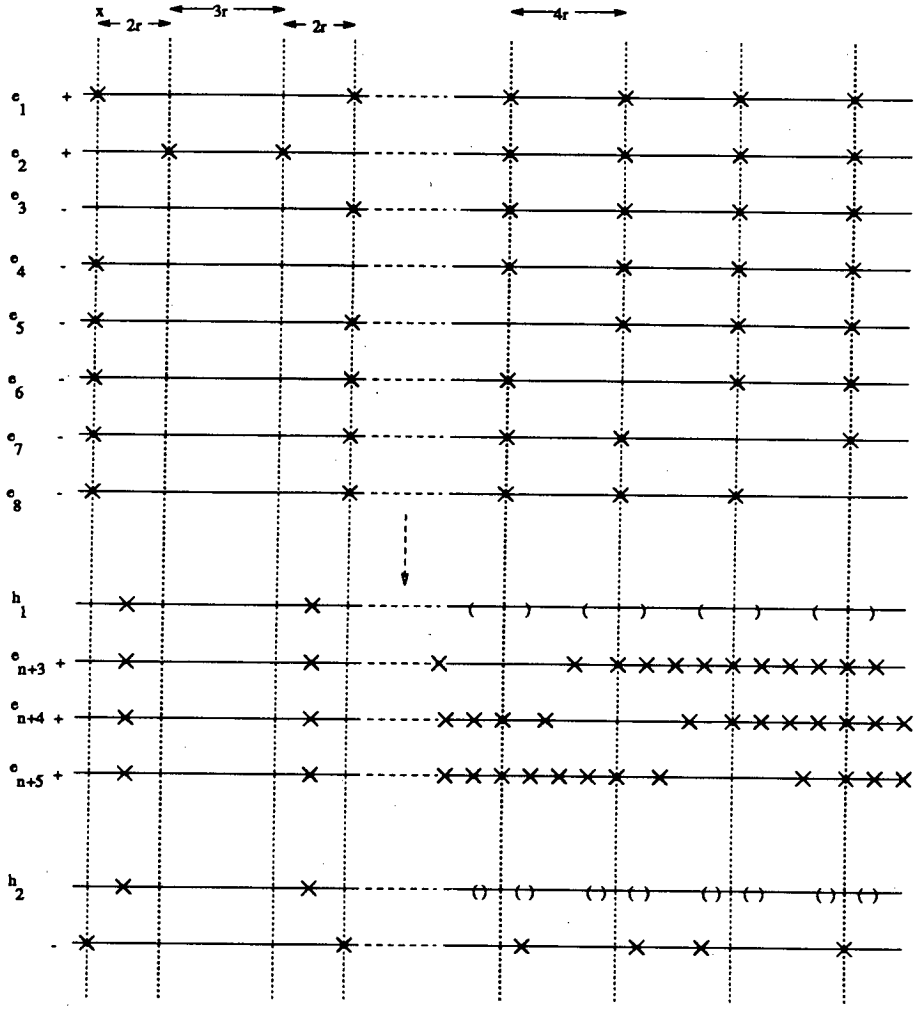

figure 4.4.

4.3. Non-learnability under the Minimax Metric

We conclude our collection of intractability results with NP-completeness proofs for concepts defined using the minimax metric. Under this metric it is impossible to compare two sets of points of different size. This simplifies the learning situation, since we are unable to appeal to an Occam algorithm, but must find a hypothesis with the same number of points as the examples (and the target concept).

We continue to restrict the learning problem to sets of points on the real line. The CHP for concepts of this form remains NP-hard. We have:

Theorem 4.9: *The CHP for concepts of n points on the real line given by the minimax metric is NP-complete.*

Proof: We will show in theorem 4.10 that the learning problem is NP-complete, even if objects are assumed to be immobile. This assumption is not in the original spirit of the minimax metric, where the "mini" refers to minimisation over translations, but the result reveals the full extent of the intractability of this learning problem. By employing a similar construction to that used in corollary 4.3, the result follows. □

Theorem 4.10: *The CHP for concepts of n points on the real line given by the minimax metric for immobile objects (this should now perhaps be called the "max" metric) is NP-complete.*

Proof: We reduce CNF to the consistent hypothesis problem. A CNF formula with k clauses and n variables is encoded using $2 + 2n + k$ examples each having $2n$ points on the real line.

Include positive examples e_1 and e_2, where e_2 differs from e_1 in being shifted r to the right. (Since points within an object must be distinguished from each other, they are numbered in fig. 4.5. Note that while the points in the examples are depicted as being $2r$ apart, there is no need for any separation between them since they are automatically distinguished, and non-corresponding points are assumed not to interact.)

This constrains each point forming the centre of a consistent hypothesis to lie in a fixed r-interval.

Include two negative examples e_3, e_4, for which points 3 to n are situated in the corresponding intervals for points in the consistent hypothesis. In e_3 point 1 lies $r/4$ to the right of its corresponding interval and point 2 lies $r/4$ to the left. In e_4 point 1 lies $r/4$ to the left of its corresponding interval and point 2 lies $r/4$ to the right.

This forces the consistent hypothesis points in intervals 1 and 2 to lie either both in the left $r/4$-subinterval or both in the right $r/4$-subinterval.

$2(n-1)$ similar examples are used to constrain the other $n-1$ pairs of consecutive points $\{(2i-1, 2i) \ : \ 2 \leq i \leq n\}$ in the same way.

Now each pair of consecutive points $(2i-1, 2i)$ may encode a boolean value, the only difference between this construction and previous ones being that two points are needed for each boolean variable.

e_{2n+3} gives the encoding of the clause $v_1 \vee v_2 \vee \neg v_3$, interpreting points $(2i-1, 2i)$ as encoding $v_i = \mathit{true}$ if they lie in the left subintervals, and $v_i = \mathit{false}$ if they lie in the right subintervals. □

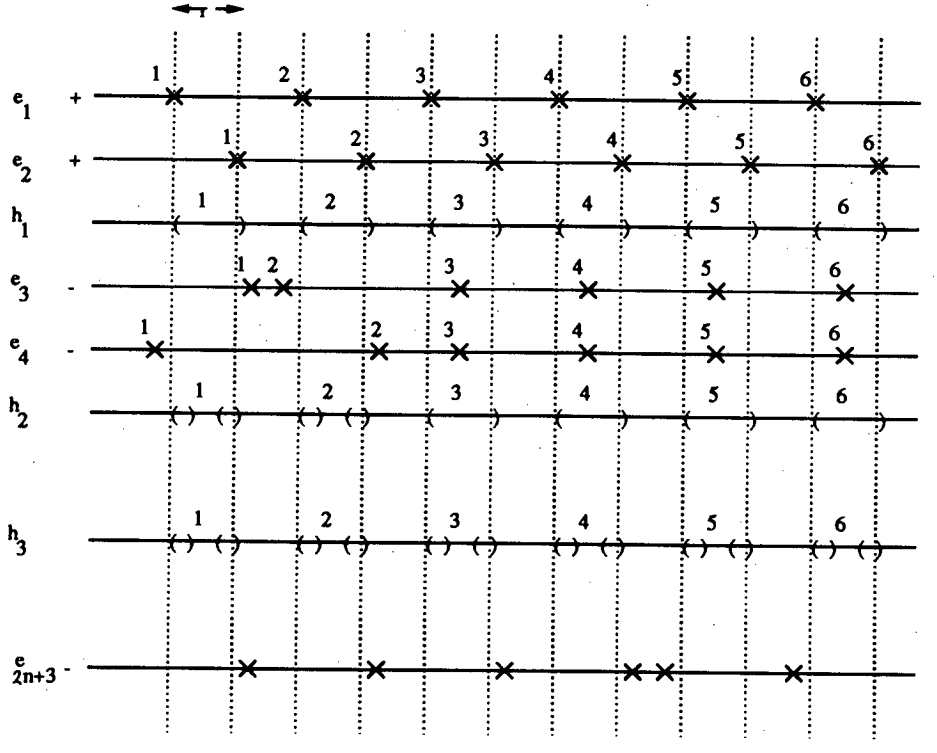

figure 4.5.

Observations: We have seen how hardness results for learning concepts of fixed radius r defined by some metric can be extended to the cases where the radius is arbitrary, and where the metric is taken modulo a group of transformations. The techniques in corollaries 4.3 and 4.4 are probably fairly generally applicable.

The algorithm for learning components from scenes can be extended to the case where the radius is unknown. It is just necessary to test for a consistent hypothesis for a polynomial-sized collection of prospective values for the radius.

4.4. Extension of learning results to Two Dimensions

The foregoing results have shown learnability/non-learnability for classes of concepts defined by sets of points in the real line. We will continue by using these as the foundation for further learnability/non-learnability results for the kind of learning problems which motivated this study, namely the problem of learning planar polygons. It turns out that the one-dimensional learning problems constitute a good basis for the two-dimensional case. Under the same assumptions regarding which metric or distance function is appropriate to represent resemblance, the difficulty of learning a geometrical object is usually the same in both cases.

The L_∞ norm is used throughout. The proofs do not work for the L_2 (Euclidean) norm, although learning is likely to be at least as hard under the L_2 norm as under the L_1 or L_∞ norms. Indeed we will see later that predicting sets of points in the plane from positive examples under the minimax metric is easy under the L_∞ norm but impossible under the L_2 norm for information-theoretic reasons.

We begin by extending the non-learnability of point sets on the real line under the Hausdorff metric, to polygons in the plane. While it may appear obvious that planar polygons are at least as hard to learn as sets of points on the line, in the case of polygons we have to deal with the connectedness of the objects. However it remains possible to exhibit a fairly natural reduction from the one-dimensional problem to the polygon-learning problem, so long as the distance between two points in the plane is measured using the L_∞ norm as opposed to the Euclidean distance. These two metrics are of course equivalent for sets of points on the line.

Sets of Similar Polygons are hard to Learn

Theorem 4.11: *The CHP for sets of similar polygons under the Hausdorff metric (with L_∞ norm) is NP-complete.*

Proof: via a reduction from the CHP for sets of points on the line.

Given sets POS and NEG of m configurations of n points in the real line, we need to transform them into sets POS' and NEG' of planar polygons such that given a hypothesis polygon H which is close (under the Hausdorff metric) to elements of POS' and NEG', we can transform H back into a set h of points on the line which is close to elements of POS but not NEG. Throughout, the radius of concept classes will be fixed as r.

We can transform a set $\langle x_1, ..., x_n \rangle$ of points in $\mathbb{R}$ into a connected arrangement of lines consisting of a horizontal "base line" with vertical lines attached at the positions given by $\{x_1, ..., x_n\}$. This set of lines itself is not a polygon but can be transformed into a simple closed polygon which is arbitrarily close to it in the Hausdorff sense. In the reduction, all elements of POS and NEG are transformed into this kind of arrangement of lines in the plane.

Choose $m, M \in \mathbb{R}, m < M$, such that all members of POS and NEG lie in the interval (m, M). Then we first define three members of POS' which act as a template for hypothesis polygons. They are the rectangle $\big((m-r,-r),(M+r,-r),(M+r,r),(m-r,r)\big)$ and the lines $\big((m+r,r),(M-r,r)\big)$ and $\big((m+r,r/2),(M-r,r/2)\big)$. (See fig. 4.6(1).)

These two lines can be treated as polygons, since they are arbitrarily close to simple closed polygons such as the rectangle $\big((m+r,r),(M-r,r),(M-r,r+\epsilon),(m+r,r+\epsilon)\big)$ for the first of the two lines.

Between them these examples force any consistent hypothesis to incorporate the "base line" $\big((m,0),(M,0)\big)$ and any other parts of the polygon to lie in the rectangle $\big((m,0),(M,0),(M,3r/2),(m,3r/2)\big)$. This is the shaded region of fig. 4.6(2).

A member $\langle x_1, ..., x_n \rangle$ of POS is mapped to the set of lines

$$\{\big((m,0),(M,0)\big), \big((x_1,0),(x_1,9r/4)\big), ..., \big((x_n,0),(x_n,9r/4)\big)\}$$

which is connected, hence has simple closed polygons arbitrarily close to it. Each such figure is included in POS' (fig 4.6(3).)

A member $\langle x_1, ..., x_n \rangle$ of NEG is mapped to the set of lines

$$\{\big((m,0),(M,0)\big), \big((x_1,0),(x_1,5r/4)\big), ..., \big((x_n,0),(x_n,5r/4)\big)\}$$

which is also connected, and each such figure is included in NEG' (fig. 4.6(4).)

This completes the transformation from $\{POS, NEG\}$ to $\{POS', NEG'\}$. We now show how to extract a consistent hypothesis for $\{POS, NEG\}$ given a polygon H consistent with $\{POS', NEG'\}$.

Let H^- be the set of points in H which have y-coordinate greater than r. Clearly H^- is easy to compute and is a set of lines no larger than the set of lines comprising H.

Now project H^- onto the real line, so obtaining the set S of all points which are x-coordinates of some point in H^-. S is a set of intervals on the real line which we claim is close to elements of POS but not elements of NEG. Replace each interval by its endpoints, and if the length of the interval is greater than $2r$, intermediate points spaced r apart. Let h be the resulting set of points. We claim that h defines a consistent hypothesis for POS and NEG.

To show that h defines a consistent hypothesis, note first that this follows if the set S of points in the real line is close to members of POS but not NEG, under the Hausdorff metric. We continue by proving that fact.

1.) S is within r of members of POS:

Let $e \in POS$, with corresponding polygon e' in POS'. A point $s \in S$ corresponds to a point (s, y) on H, for some $y > r$. Hence for $e' \in POS'$, there must be a point on e' which is within r of (s, y). This point must have y-coordinate strictly greater than zero, hence it must lie on a vertical line, thus corresponding to a point in e. So $s \in S$ is close to some point in e. Points in e have a corresponding point in e' with y-coordinate $9r/4$. There must be point in H within r of each of these points in e', having y-coordinate $\geq 5r/4$, hence forming a point in S.

2.) S is not within r of members of NEG:

Suppose in fact S is within r of $e \in NEG$. Let e' be the corresponding member of NEG'. Then it follows that H is within r of e', since all points of H with y-coordinate $\leq r$ are within r of those parts of e' with y-coordinate $\leq r$. Points in H with y-coordinate $> r$ must be close to those parts of e' with y-coordinate $> r$. Hence they are $\leq r$ apart under the Hausdorff metric, contradicting the negativity of e'.

We have shown that if there is a polynomial-time algorithm for finding a hypothesis polygon consistent with a given sample of polygons, then it can be used to construct a polynomial-time algorithm for the CHP for sets of points on the line. This completes the reduction. □

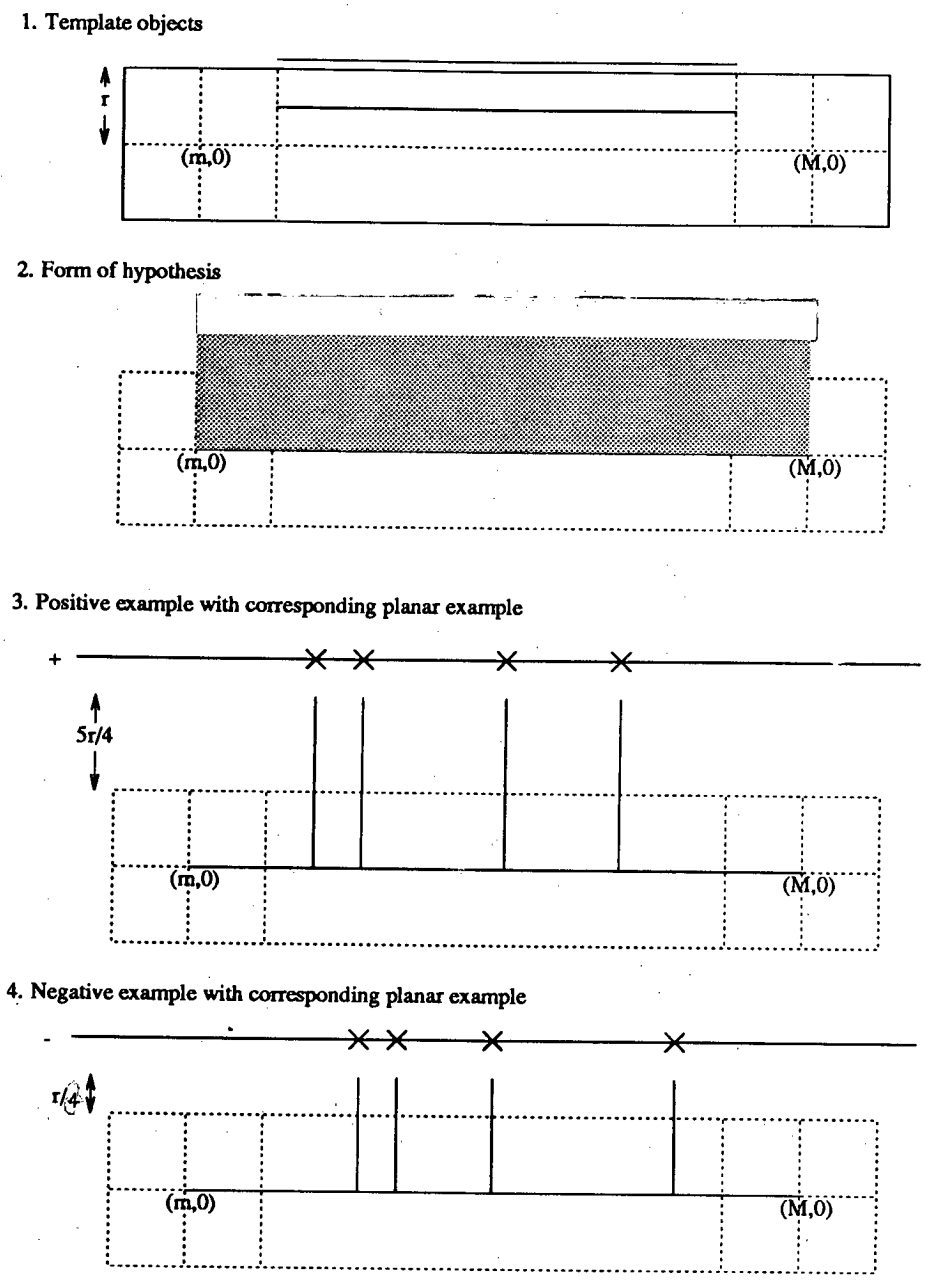

figure 4.6.

To show that it is hard to learn sets of points in the plane, as opposed to polygons, we can essentially use the trivial reduction of embedding the line in the plane, provided that the L_∞ norm is used to measure the distance between

points. Under the L_∞ norm any point in a hypothesis which does not lie in the embedded line can be moved to the nearest point on that line without affecting the consistency of any example with the hypothesis.

Learning Two-dimensional Concepts defined by the Directed Hausdorff Distance

The Occam learning algorithm given for learning concepts consisting of sets of points on the real line defined by the directed Hausdorff distance can be generalised to two dimensions. This involves discretising the learning problem in a relatively explicit way.

We will consider sets of points in the plane rather than polygons. We assume that we have two sets POS and NEG of positive and negative examples respectively, each of size m, with each example consisting of a set of n points in the plane. The arguments and algorithm apply for any of the standard distance functions for points in the plane. We will assume the Euclidean distance is being used, so that the set of points within r of some fixed point is contained in a circle in the plane. We will assume that concepts are closed sets, so that this set includes its boundary.

Theorem 4.12: *The concept class as defined above is Occam learnable.*

Proof: Suppose there are m positive and m negative examples, each example consisting of a set of n points. Hence there are $O(mn)$ example points in the sample.

All the circles around the example points decompose the plane into regions within which all points are contained in the same subset of this collection of circles, and there are $O((mn)^2)$ of these regions. This follows from the observation that circles in the plane have a V-C dimension of 3, which implies that the number of distinct regions which they can divide the plane into is quadratic in the number of circles (the growth function of [VC71].)

Our strategy is to find a representative point for each of these regions, and the learning task need only involve this finite and polynomial-sized collection of points, which will be denoted D (for discretisation).

D can be constructed in $O((mn)^2)$ time. As a result, the following is an (Occam) learning algorithm which finds a sufficiently small set of points in the plane which defines a "sphere" under the directed Hausdorff distance which

contains POS but no element of NEG. The algorithm uses the same greedy set cover method as *FIND-COMPONENT*.

Algorithm: 2D-FIND-COMPONENT

$POS := \{e_1, ..., e_m\}$;
$NEG := \{e'_1, ..., e'_m\}$;
Let D be the set of representative points defined above;
$I := \{x \in D : x$ is within r of some point in e_i, for $1 \leq i \leq m\}$;
{*I is found simply by checking every element of D, and corresponds to the I in algorithm FIND-COMPONENT*}
Let $N :=$ {elements of D not within r of any point in $e'_i : 1 \leq i \leq m$};
{ *Greedy set cover for negative examples:* }
repeat
find a point $p \in I$ lying in the maximum possible number of members of N;
add that point to centre of hypothesis;
$N := N -$ members of N containing p;
until $N = \varnothing$;

The algorithm works in much the same way as *FIND-COMPONENT* in one dimension, where discretisation of the domain was provided by pairs of points r distant from points forming the examples. In the above algorithm a set (D) of points characterising all regions in which points are indistinguishable from the point of view of the examples, is calculated in advance.

Chapter Five

Finding a Hypothesis Consistent with a Set of Positive Examples

In section 1.4 the property of minimal consistency was defined, and noted to be a necessary condition for learnability from positive examples in the PAC learning model. We observe here that none of the concept classes considered in this thesis are minimally consistent. Nevertheless, in this chapter we consider the computational aspects of finding a concept which contains some given sample of positive examples. Since the problem is trivial if hypotheses of arbitrarily large radius are allowed, it is assumed that concept classes are of fixed radius r.

In section 5.1 we consider the motivation for this approach, and in section 5.2 we present the results. As might be expected the problem is easy, indeed sometimes trivial, for many of our concept classes. The main result however, is that the problem remains NP-hard for sets of points under the Hausdorff metric, with minimisation over translations.

5.1. Motivation

The problem of finding a hypothesis consistent with positive examples only is a natural one, despite the fact that in the learning framework we have defined, no individual concept offers the required guarantees as a predictor of subsequent examples. An example of an application where only positive examples would be available is the problem in robotics of inferring the shape of a three-dimensional object from a collection of two-dimensional projections of it. It is also intuitively believable that learning of this sort takes place in practice, as is argued in, for example, [N87]. For concept classes that are not minimally consistent (such as the ones considered in this thesis) this assumption requires a less demanding learning model to be developed. For example [N87] gives a polynomial-time learning algorithm for bounded DNF formulae, subject to the restriction that the examples (assignments to the boolean variables) have a uniform distribution.

In the next chapter we consider an alternative relaxation of the requirements, namely that a hypothesis may consist of an intersection of concepts. This extension of the hypothesis class to its closure under intersection makes it minimally consistent, and turns out to yield feasible prediction algorithms from positive examples only, for some concept classes. The methods developed in this chapter serve as a basis for the prediction algorithms.

Another reason which can be advanced for this approach is suggested by the results, that reveal that tractability of the positive example consistent hypothesis problem can depend on whether a metric is minimised over a class of transformations of geometrical objects. The results point us to the source of the intractability. We have shown that it is hard to "sandwich" a hypersphere between two sets of points that its surface is supposed to separate, but it is now revealed that this is sometimes, but not always, due to the problem of just finding one of limited radius that holds all the points supposed to lie in its interior.

In the NP-completeness proofs in the previous chapter, negative examples played a key role in the encodings of CNF formulae. This is because a negative example naturally encodes a disjunction, since its negativeness must be the result of one or more of a collection of possible features of that example causing it to lie outside the concept. The elimination of negative examples offers renewed prospects for feasible algorithms in a suitably modified learning model.

If the radius of concepts was allowed to vary, one natural choice of hypothesis would be the sphere of smallest radius containing the sample. This is not a subset of all other consistent hypotheses (that would make the class minimally consistent) but may have other desirable properties, such as correctly classifying negative examples which are sufficiently distant from the figure defining the target concept. Approximations to this hypothesis could be found iteratively by searching for hypotheses of various radii containing the sample. The problem of finding such a concept can be alternatively viewed as that of finding Chebyshev centres (see [BK80].)

However, the value of this hypothesis remains uncertain, as the following example shows:

Example: A concept is a set of points in the plane all within r of some fixed point in the plane under the L_∞ metric, hence axis parallel squares of edge length $2r$ in the plane. The following diagram shows that a concept of minimum radius may exist which contains points up to $3r$ from the centre.

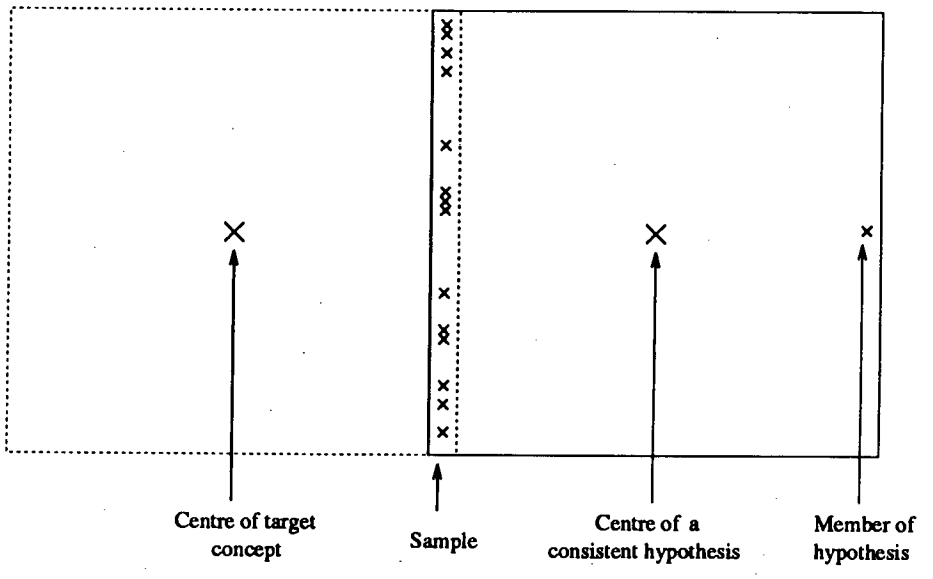

figure 5.1. Hypothesis of minimum radius with element $3r$ from centre of target concept

If we seek a hypothesis which is guaranteed to eliminate all negative examples whose distance from the figure defining the target concept is greater that some multiple of the radius, then a hypothesis of minimal radius may be no better that the rather trivial one obtained by choosing a sample point as the centre of the hypothesis and taking as the radius the maximum distance of it from any other sample point. Both hypotheses may contain points up to $3r$ from the centre.

Hence in this example the criterion of guaranteed elimination of all negatives of sufficiently large distance from the central object cannot always be used to distinguish the consistent hypothesis of minimal radius from a trivial consistent hypothesis. We note however that in this case there is a better choice of hypothesis of minimal radius consistent with the data which is guaranteed to contain points at most $2r$ from the centre.

5.2. Results

As noted earlier, it is in general easy to find a concept of fixed radius r which contains a set of examples known to belong to some target concept of radius r. We will consider concept classes consisting of sets of points on the real line, under various metrics.

Surprisingly however, if the objects are allowed to move under translations, so that the distance between two objects is the Hausdorff distance minimised over translations along $\mathbb{R}$, the CHP is once again NP-complete, and this is proved by a reduction from 3-CNF. It is a key feature of the proof that a positive example can encode a disjunction due to the fact that it must be within Hausdorff distance r of the hypothesis in one of a number of possible positions.

We have: *possible in polynomial time*

Theorem 5.1: *It is* ∧ *to find a concept of radius r which contains a set of positive examples of some target concept, for concept classes of fixed radius r whose concepts are defined by sets of points in the real line, using the following metrics:*

1.) The minimax metric (with or without minimisation over translations)

2.) The directed Hausdorff distance (with or without minimisation over translations)

3.) The Hausdorff metric (without minimisation over translations)

Proof:

Proof of (1): Without translations the problem is easy. Each point defining a consistent hypothesis must be placed within r of the set of m corresponding example points, where m is the number of examples.

With translations it can be solved efficiently as follows: Let $\langle x_1, ..., x_n \rangle$ denote a hypothesis. Each example imposes two linear constraints on each pair (x_i, x_j), $i \neq j$. We can identify the $n(n-1)$ strongest constraints, $x_i - x_j \geq c_{ij}$, and find a solution to this system of equations in polynomial time.

Proof of (2): Recall that under the directed Hausdorff distance, concepts are components of scenes and examples are scenes containing them. Hence the "empty component" is a trivial solution to any instance of this problem. If a non-empty component must be obtained, then for immobile scenes, it is easy to find a point within r of some point in each scene. For mobile scenes, any point will serve as a hypothesis, since it may be moved to become a component of any scene in a sample.

Proof of (3): We can find a hypothesis whose size (that is, number of points on $\mathbb{R}$ comprising the object at its centre) is minimal.

We start by presenting the hypothesis-finding algorithm for positive examples in a concept class of radius r. The sample is of size m and each object has n points.

Algorithm: 1D-POS-HYPOTHESIS

Let $\{e_1, ..., e_m\}$ be the (positive) examples;
$e_i = \langle x_{(i,1)}, x_{(i,2)}, ...x_{(i,n)} \rangle$ for $1 \leq i \leq m$;
$S_i := \{x \in \mathbb{R} :$ some point in e_i is within r of $x\}$;
{*S_i is a union of at most n intervals.*}
$I := \cap_{i=1}^{m} S_i$;
{*I is a union of $O(mn)$ intervals.*
Note that any consistent hypothesis must be a subset of I. The second phase of this algorithm generates a set of n points which forms a consistent hypothesis.}
$H := \varnothing$; $S := \cup_{1 \leq i \leq m, 1 \leq j \leq n} \{x_{i,j}\}$;
Repeat
  $p :=$ rightmost point in I within r of leftmost point in S;
  $H := H \cup \{p\}$;
  $S := S \setminus \{$all elements within r of $p\}$;
until $S = \varnothing$;
hypothesis := r-ball defined by H.

This greedy algorithm will eliminate all points in the examples on or before reaching n, the size of the concept. The number of points generated is minimal for a consistent hypothesis, since by an inductive argument, for all i the ith leftmost points in the hypothesis account for a maximal number of points occurring in examples.□

We now prove the main result of this section:

Theorem 5.2: *For a concept class of fixed radius r with spheres defined by the Hausdorff metric minimised over translations, it is NP-complete to find a consistent hypothesis.*

Proof: We reduce satisfiability of 3-CNF to the consistent hypothesis problem with $3n + k + 6$ examples each having up to $4n + 6$ points, where n is the number of variables and k the number of clauses in the formula.

We begin with a description of the general ideas of the proof before going through the technical details. Note first that this is a reduction from 3-CNF since unlike previous proofs there appears to be no natural reduction from general CNF. The proof uses the same technique of encoding the value of a boolean variable by a pair of intervals for which containment of a point in a consistent hypothesis is governed by an exclusive-or relationship, so that such a pair must contain a hypothesis point but one interval of the two must not. Then using the usual interpretation that a hypothesis point in the left-hand interval of the ith pair represents the value *true* for the ith variable v_i and containment by the right-hand interval represents *false*, we show how to define examples representing the constraints $v_i = \neg v_j$ and $v_i \vee v_j \vee v_k$, for $1 \leq i, j, k \leq$(number of pairs of intervals). Hence if we use $2n$ such interval pairs, we can use them to encode the n variables, together with their negations using examples of the first kind, and consequently disjunctions of three variables and their negations using examples of the second kind. So as before the formula can be encoded by a set of such examples.

It has been necessary to constrain the structure of the objects used in the reduction more than in previous proofs (where the constraints were minimised in order to indicate the salient features.) In fact every point will have its position fixed, so that examples can be described using a tuple of real numbers. We assume unit radius.

We start by including the following two examples (shown in fig. 5.2):

$$e_1 : \langle -10, -5, 0, 2{\cdot}5, 5, 7{\cdot}5, 10, ..., 10n \rangle$$

$$e_2 : \langle -8, -7, 0, 2{\cdot}5, 5, 7{\cdot}5, 10, ..., 10n \rangle$$

Since examples are movable along $\mathbb{R}$ the (Hausdorff) distance between them must be minimised over these translations. In the positions given that distance is 2, and in fact any shift from this relative position will increase this distance. The two leftmost points in each example ensure that this is the case. Hence they must take this relative position for a CH to exist (since a hypothesis is a sphere of radius 1).

A hypothesis consistent with these two examples must take the following form, where an interval in a tuple indicates that at least one point in that interval must be present. This is a description of the object lying at the centre

of the hypothesis, and is given in the position in which it is unit distance from e_1 and e_2 as expressed above:

$$\langle -9, -6, [-1, 1], [1{\cdot}5, 3{\cdot}5], [4, 6], [6{\cdot}5, 8{\cdot}5], [9, 11], ..., [10n - 1, 10n + 1] \rangle$$

The following two examples (shown in fig. 5.2) force the point(s) in alternate intervals $\{[5i - 1, 5i + 1]\}$ to lie at the centre of those intervals.

$$e_3 : \langle -8, -7, -1, 2{\cdot}5, 4, 7{\cdot}5, 9, 12{\cdot}5, 14, ..., 10n - 1 \rangle$$

$$e_4 : \langle -8, -7, 1, 2{\cdot}5, 6, 7{\cdot}5, 11, 12{\cdot}5, ..., 10n + 1 \rangle$$

A hypothesis now takes this form:

$$\langle -9, -6, 0, [1{\cdot}5, 3{\cdot}5], 5, [6{\cdot}5, 8{\cdot}5], 10, ..., 10n \rangle$$

Then in a similar way, the following two examples (also shown in fig. 5.2) reduce the size of the remaining intervals from 2 to 1.

$$e_5 : \langle -8, -7, 0, 2, 5, 7, 10, 12, 15, ..., 10n \rangle$$

$$e_6 : \langle -8, -7, 0, 3, 5, 8, 10, 13, 15, ..., 10n \rangle$$

A hypothesis now takes the following form (h_1 in fig. 5.2):

$$h_1 : \langle -9, -6, 0, [2, 3], 5, [7, 8], 10, [12, 13], ..., 10n \rangle$$

We have $2n$ intervals separated by fixed points, and the next set of examples ($2n$ of them) will split each of these intervals into exclusive-or subintervals, ready to encode boolean values.

The following example (e_7 in fig. 5.2) divides the leftmost of the $2n$ intervals in this way. The points at $-9{\cdot}5$, $-7{\cdot}5$, $-5{\cdot}5$ imply that it must move $1/2$ in either direction to be consistent with the hypothesis points at -9 and -6.

$$e_7 : \langle -9{\cdot}5, -7{\cdot}5, -5{\cdot}5, 0, 0{\cdot}6, 4{\cdot}4, 5, 7{\cdot}5, 10, 12{\cdot}5, 15, ..., 10n \rangle$$

So a hypothesis must now take the following form (h_2 in fig. 5.2):

$$h_2 : \langle -9, -6, 0, [2, 2{\cdot}1] xor [2{\cdot}9, 3], 5, [7, 8], 10, [12, 13], ..., 10n \rangle$$

The following set of examples applies the same treatment to the other intervals. For $1 \leq k \leq n-1$:

$$e_{7+k} : \langle -9{\cdot}5, -7{\cdot}5, -5{\cdot}5, 0, 2{\cdot}5, 5, ..., 5k, 5k + 0{\cdot}6, 5k + 4{\cdot}4, 5(k+1), ...$$

$$, 5(2n-1), 5(2n-1) + 2{\cdot}5, 10n\rangle$$

A hypothesis must be of the following form (h_3 in fig. 5.2):

$$h_3 : \langle -9, -6, 0, [2, 2{\cdot}1]xor[2{\cdot}9, 3], 5, [7, 7{\cdot}1]xor[7{\cdot}9, 8], 10,$$

$$[12, 12{\cdot}1]xor[12{\cdot}9, 13], ..., 10n\rangle$$

The following example (shown in fig. 5.2) encodes the formula $v_1 = \neg v_2$. Again, the three points at $-9{\cdot}5$, $-7{\cdot}5$, $-5{\cdot}5$ force it to move 1/2 in either direction to be consistent.

$$e_{7+n} : \langle -9{\cdot}5, -7{\cdot}5, -5{\cdot}5, 0, 1, 4, 5, 7, 8, 10, 12{\cdot}5, 15, 17{\cdot}5, ..., 5n\rangle$$

Finally the following example (shown in fig. 5.2) encodes the formula $v_1 \vee v_2 \vee v_3$.

$$e_{7+n+1} : \langle -9{\cdot}5, -5{\cdot}5, 0, 2, 5, 6{\cdot}5, 9{\cdot}5, 10, 11, 12{\cdot}5, 14{\cdot}5, 15, 17{\cdot}5, 20, ..., 5n\rangle$$

This completes the reduction, since we can use the two structures above to encode a disjunction of three variables or their negations. It is worth explaining how the last example works. In order for the two points on the left (ie. at $-9{\cdot}5$ and $-5{\cdot}5$) to fit in with the hypothesis, it may remain in the position given or move up to 1/2 in either direction.

The point at 2 is consistent with $v_1 = \mathit{true}$ in any of these positions but with $v_1 = \mathit{false}$ in the given or right positions only.

The points at 6·5 and 9·5 are consistent with $v_2 = \mathit{true}$ in any of the positions but with $v_2 = \mathit{false}$ in the right or left positions only.

The point at 13 is consistent with $v_3 = \mathit{true}$ in any of the positions but with $v_3 = \mathit{false}$ in the given or left positions only.

Hence it requires one of these variables to have a *true* encoding in the hypothesis for this example to be consistent. □

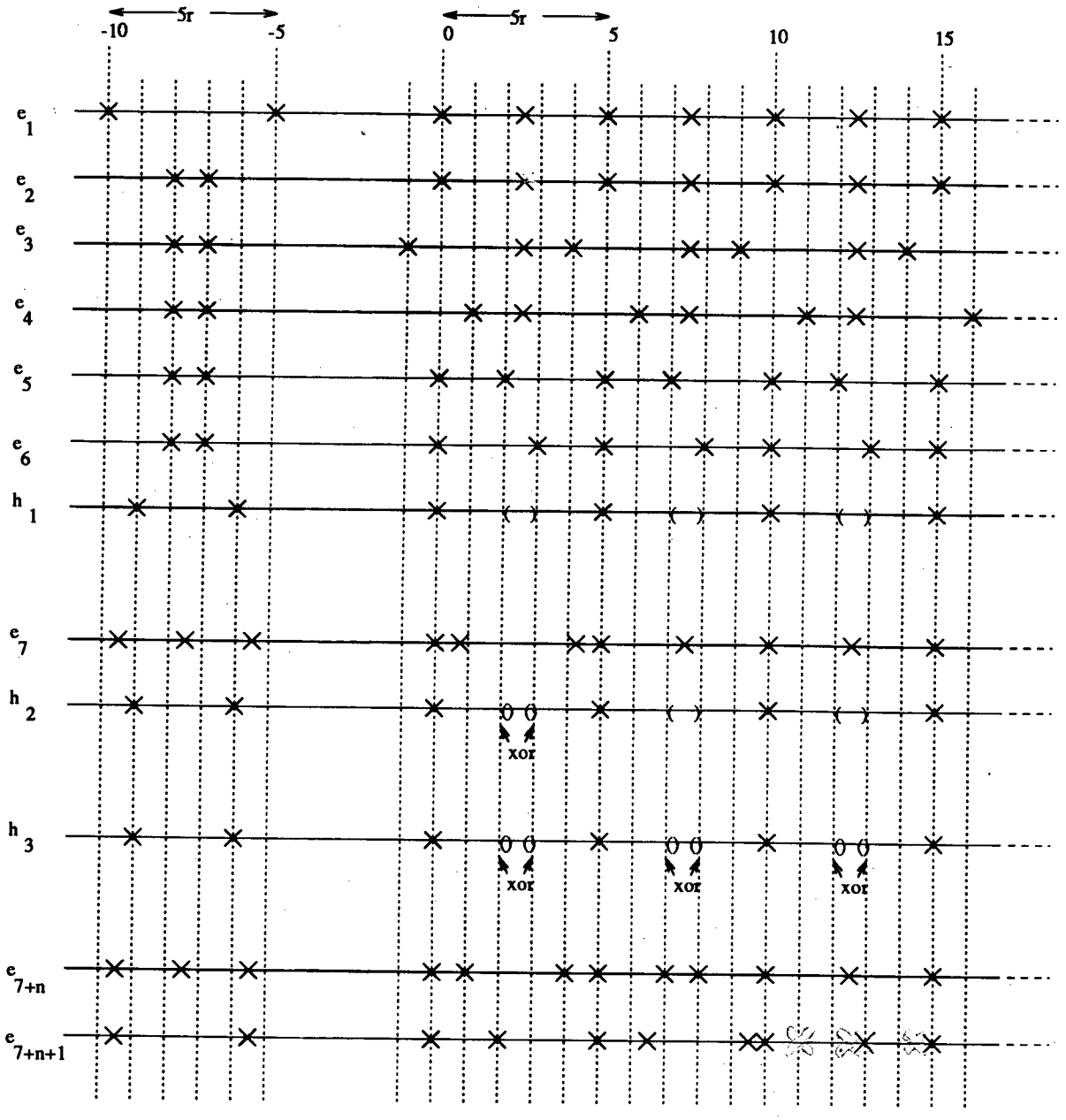
5r
-10
-5
5r
0
5
10
15
e 1
e 2
e 3
e 4
e 5
e 6
h 1
e 7
h 2
xor
h 3
xor
xor
xor
e 7+n
e 7+n+1

figure 5.2.

Chapter Six

Prediction Results

So far we have shown that for a wide range of problems concerned with the inductive learning of a geometrical shape, the learning problem is NP-hard, and that this remains the case if hypotheses are allowed to be defined by more complex figures than the target concept. In this chapter we consider the problem of *prediction*, in which a hypothesis is allowed to be any subset of the instance domain $\mathcal{X}$. The only restriction on it is that members or non-members of such a hypothesis must be recognisable by a polynomial-time algorithm.

We show that this relaxation of the learning task leads to polynomial-time algorithms in various formulations of the problem which were earlier shown to be NP-complete. Indeed some non-learnable concept classes turn out to be predictable from positive examples alone. In section 6.1 we describe a general method for prediction from positive examples, and show its applicability to some of the learning problems considered earlier.

For information-theoretic reasons, it remains impossible to predict concepts defined by sets of points on the real line using the Hausdorff metric, from positive examples only. However this class is predictable from positive and negative examples, and section 6.2 describes a prediction algorithm. Consequently these results shed some light on the importance of negative examples to this kind of learning.

Finally, in section 6.3 we present a non-predictability result for a geometrical concept class. The domain is polygons in the plane, where concepts are defined using either the Hausdorff or Fréchet metric minimised over sets of isometries which include rotations. This non-predictability is predicated on the non-predictability of disjunctive normal form formulae.

6.1. One-sided error prediction from Positive Examples

The method we present is an extension of the idea of the learning function for axis-parallel rectangles in the plane (example (1.1)). The hypothesis taken there was the smallest axis-parallel rectangle which contained all the positive

examples, and the negative examples were ignored. The reason why it is possible to ignore all the negative examples is that this concept class is minimally consistent, so the hypothesis can be a subset of all possible target concepts, giving learning with one-sided error.

Note that the hypothesis is the intersection of all concepts consistent with the positive examples. This suggests the following prediction method, in which the hypothesis class is extended to a minimally consistent class by letting hypotheses be certain intersections of concepts.

In general we have no guarantee that such a hypothesis will take the same form as a concept, which is why this constitutes prediction as opposed to learning, except in such concept classes as the example above. Given a sample POS of positive examples, our criterion for membership of the hypothesis derived is that any concept containing POS as a subset must contain the element being tested. This hypothesis can be characterised by any subset $POS^- \subseteq POS$ with the property that all concepts C which contain POS^- also contain POS. This is because a sample consisting of the set POS^- on its own should produce the same hypothesis as POS. In [HSW89] a minimal set with this property is called a *spanning set* for POS. We may express a hypothesis in terms of the set POS^-. This will not generally correspond to a concept if the concept class is not minimally consistent. This characterisation of a hypothesis for a prediction method is useful, since it will be shown that the existence of a set POS^- with this property having fixed finite size for all samples POS is a necessary and sufficient condition for uniform predictability from positive examples.

We denote by $\langle POS^- \rangle$ the hypothesis associated with POS^-. Then the membership testing algorithm for an unclassified example e for a hypothesis $\langle POS^- \rangle$ must determine the existence or non-existence of a concept containing POS^- as a subset but not $\{e\}$. If such a concept exists, classify e as negative, otherwise classify e as positive.

We use the following algorithm for generating POS^- from POS.

Algorithm: FIND-SUBSET

$POS^- := POS$;
for all elements $e \in POS^-$ do
begin
  if there exists a concept C s.t. $(POS^- - \{e\} \subseteq C)$ and $(e \notin C)$
    then do nothing
    else $POS^- := POS^- - \{e\}$;
end.

The algorithm works as follows. Define an element e of a set of positive examples to be *redundant* if all concepts containing the other members of the set contain e. Hence the other examples are enough to tell us that e is a positive example. *FIND-SUBSET* sets POS^- equal to POS initially and then takes each element of POS in turn, removing it if it is redundant in POS^-.

The algorithm can be seen to be polynomial-time provided that the "exist" conditions can be computed in polynomial time. This turns out to be the case for concept classes of immobile objects under the Hausdorff metric and minimax metric, but not, as we have seen, under the Hausdorff metric if the objects are movable and the radius of the concept class is given.

Properties of the method:

[KLPV87] show that learnability from positive examples implies learnability with one-sided error. The converse fails precisely when knowledge of just one negative example is necessary and sufficient for one-sided error learning. If no negative examples need to be seen then we have learning from positive examples alone. One-sided error learning is impossible if more than one negative example needs to be seen, since it will fail for distributions for which all negative examples are the same. An example of a class for which the converse fails is $\mathcal{X}$ = points on the unit circle, $\mathcal{C}$ = connected subsets (arcs) of the unit circle.

The same holds for prediction. Predictability from positive examples implies predictability with one-sided error (no false positives), since a hypothesis should not contain any member of $\mathcal{X}$ which could be negative for some possible target concept.

The set POS^- generated by the algorithm is not necessarily minimal. In fact, a minimal set characterising a given sample and having no redundant

elements may be hard to find. However there is an upper bound for the size of the set generated which we describe below, which characterises the information-theoretic predictability of a concept class from positive examples. (The hardness of finding a *minimal* set can be proved using a simple reduction from the NP-complete HITTING SET problem described in [GJ79]. That is, given a finite set S and a collection of subsets of S, find a subset $T \subset S$ of given size k such that every subset in the collection has an element in T.)

For the method to work on a particular concept class we require:

1.) We need to obtain a suitable set POS^- with size sublinear in the sample size m (the criterion for Occam learning). We will show that if POS^- is too large then the hypothesis is unlikely to be a good predictor.

2.) POS^- should be determined in polynomial time.

3.) The membership test for a hypothesis $\langle POS^- \rangle$ should be performable in polynomial time.

We will show that these conditions are satisfied for sets of points on the real line under the minimax metric, and for some restrictions of the concept class of points on the real line under the Hausdorff metric. For now we will observe that satisfaction of these properties is a sufficient condition for learning, since they imply an Occam algorithm whose hypothesis is consistent with the positive examples, and the negatives (assumed unseen).

Observation 6.1: The algorithm used to test for membership of this hypothesis is the same kind of existence test used in the generation of the hypothesis. Hence if condition (2) above is satisfied, then condition (3) is also satisfied. So we need only consider (1) and (2).

A hypothesis H produced by this method has the property that it is the largest hypothesis obtainable from positive examples only such that all negative examples are correctly classified. Since the probability distribution of negative examples is unknown this is a necessary feature of a hypothesis derived from positives. Any hypothesis H' containing an example $x \in \mathcal{X}$ outside H cannot have guaranteed PAC-ness unless negative examples are available to corroborate it. This is because for any point x outside H there exists a concept containing the given sample of positive examples which does not contain x. Hence we have no guarantee that random negative examples are unlikely to be the same as x.

As we will see, even given that it is the largest hypothesis with this property, it can sometimes contain only the known positive examples, failing to generalise

to sufficiently many others – this usually happens when POS^- contains all or nearly all of the positive examples, failing condition (1) above. The intuition is that no inferences have been made from these examples, and so we are left with the sample that is already known.

We now describe the upper bound (associated with a concept class) on $|POS^-|$, and show that its size being polynomial in target concept complexity is a necessary and sufficient condition for the hypothesis to be a good predictor.

Recall the definition of d^+ introduced in section 1.4.

This is an upper bound on the size of the set POS^- returned by algorithm *FIND-SUBSET*, since POS^- is a set with the above properties.* It is also an upper bound for the V-C dimension since a larger set cannot be shattered. This function is analogous to the V-C dimension in terms of characterising predictability from positive examples only. Define "uniformly predictable from positive examples" analogously to the definition of "uniformly learnable". The following is a simple extension of theorem 1.5:

Theorem 6.1: *C is uniformly predictable from positive examples if and only if $d^+(C)$ is finite.*

Proof: If d^+ is finite then algorithm *FIND-SUBSET* is an Occam learning algorithm as in [BEHW89].

If d^+ is infinite then for an arbitrarily large set S with the above properties, we let positive examples be distributed uniformly over S. Given the parameters ϵ, δ the number of examples needed for PAC-learning is $O(|S|)$, unlimited. □

We have:

Corollary 6.2: *Let $C = \cup_{n\in\mathbb{N}} C_n$ be a stratified concept class having the property that given a set of examples belonging to some concept in C_n we can test in polynomial time whether they belong to a concept not containing some other given example. Then C is predictable in polynomial time if and only if $d^+(C_n)$ is polynomial in n.*

The prediction algorithm used is just *FIND-SUBSET* on the sample. Note that the size of POS^- never depends on the sample size, only on the complexity of the target concept. Hence if a concept class actually requires a hypothesis of complexity $n^c m^\alpha$, $\alpha \neq 0$, for prediction then it cannot be predicted in polynomial time from positive examples. Corollary 6.2 shows that if it is easy to

* ie. for all $x \in S$, there is a concept containing $S \setminus x$ but not x

find a concept consistent with a set of positive examples and one negative example, then the effectiveness of the above method characterises predictability from positive examples.

Applications

Theorem 6.3: *Sets of points on the real line under the translation-free minimax metric can be predicted from positive examples. (That is, the distance between two point sets is the greatest distance between any two corresponding points, without any minimisation over translations.)*

Proof: Use algorithm *FIND-SUBSET*.

Let C_n be concepts (with examples) having n points. Let POS^- be the subset of POS generated by *FIND-SUBSET*. Let e be an unclassified example to be tested for membership of (POS^-). A necessary and sufficient condition for the existence of $C \in C_n$ such that $e \notin C$ but $POS^- \subseteq C$ is that one or more points in e is to the right or to the left of all corresponding points in the other members of POS^-. Hence:

1.) The existence conditions can be calculated in polynomial time.

2.) $d^+(C_n) = 2n$ hence $|POS^-| \leq 2n$. POS^- consists of all examples in the sample POS containing a point which is rightmost or leftmost among corresponding points on all the other examples. There are at most $2n$ such examples. All others are redundant.

It follows from observation 6.1 that the hypothesis has a polynomial time membership test. □

Remarks: The form of the hypothesis obtained can be simplified. It is equal to the intersection of two new concepts, C_l and C_r, defined as follows. The ith point of the central object defining C_l is set to be the leftmost of the ith points in all the examples, and similarly the ith point for C_r is set to be the rightmost.

This only generalises to two dimensions using the L_1 or L_∞ norms, but not the L_2 norm. Under the L_2 norm, even concepts defined by a single point cannot be predicted. A concept defined by one point only is the points contained in a circle in the plane under the L_2 norm, and an axis-parallel square under the L_∞ norm. For sets of points inside unit circles in the plane, $d^+ = \infty$, but for sets of points inside axis-parallel unit squares in the plane, $d^+ = 4$.

Theorem 6.4: *Sets of points on the real line under the minimax metric can be predicted from positive examples.*

Proof: Let $\mathcal{C}_n$ be the above concept class restricted to concepts and examples defined by n points. The proof follows from two observations:

1.) It is easy to find a concept consistent with a given set of positive examples and one negative example.

2.) $d^+(\mathcal{C}_n) = n(n-1)$

For the second observation, we note that the constraints put on a consistent hypothesis defined by a set h of n points, by an example e are the set of $\frac{1}{2}n(n-1)$ pairs of linear constraints put on the relative positions of pairs of points in h by the pair of corresponding points in e. Hence a subset of a sample of size $n(n-1)$ is sufficient to make all other examples redundant, if it contains, for each pair of points, the example for which those two are the furthest apart, and the example for which they are closest. These will imply the linear constraints put on the distance between that pair of points by the other examples.

It is in fact possible to construct a sample for which $n(n-1)$ examples are necessary as well as sufficient to make the others redundant. □

Theorem 6.5: *Let $\mathcal{C}$ be the concept class of radius r whose objects are sets of points on the real line under the Hausdorff metric.*

Then $\mathcal{C}$ cannot be predicted from positive examples only.

Proof: We construct an infinite subset $\mathcal{C}^- \subseteq \mathcal{C}$ with the property that $d^+(\mathcal{C}^-) = \infty$, and all members of $\mathcal{C}^-$ are defined by two points.

The method fails in the following instance of the problem: Assuming unit radius, let the target concept be the open unit sphere around $\langle (x-1), (x+1) \rangle$, for some unknown fixed x, $0 \leq x \leq 1$. Let positive examples be sets of three points whose left point is at position 0 on the real line, the right point is at position 1, and the middle point occurs in the open intervals $(0, x) \cup (x, 1)$ under a uniform distribution. There is only one negative example, consisting of the tuple of points $\langle 0, x, 1 \rangle$.

The intersection of all concepts containing the positive examples is just the set of positive examples. This is because for any example e of the form above, $e = \langle 0, y, 1 \rangle$, e not included in the given sample of positive examples, there exists a concept (defined by $\langle (y-1), (y+1) \rangle$) for which that new example

could be negative, but is consistent with the known positive examples (that is, it contains them.) Hence that example could represent all the negative examples, which have been given a probability distribution which makes them take just one form.

So the hypothesis generated by this method is just the union of the positive examples. This means that almost all positive examples will be misclassified as negative. □

Remarks: When concepts are known to have points separated by at least $4r$, the problem is equivalent to learning under the minimax metric, since it can be seen which points correspond to which. Hence this restricted concept class is predictable, but not learnable.

A similar result holds for concepts defined using the directed Hausdorff distance (where concepts are components of scenes and examples are scenes containing them.) These are not predictable from positive examples – assuming unit radius for concepts, the set of components consisting of points in the interval $(1, 2)$ has $d^+ = \infty$. Examples of the form $\langle x, 2 + x \rangle$ for $0 \leq x \leq 1$ may contain all but any one of a given set of these components (assuming again that concepts are open sets.)

When the above concept class is restricted to those concepts defined by a component whose points are at integer positions, the method (using *FIND-SUBSET*) produces a hypothesis defined by a set POS^- of size $O(nr)$, where n is the number of points and r is the radius of the concept class. This is because the set of allowable positions for placing points at the centre of a consistent hypothesis is reduced by one for each example in the set POS^-. Any example in POS^- defines a set of $\leq nr$ positions where a point defining the target concept may occur, and considering the other members of POS^- consecutively, each must reduce the number of positions by at least one, or it would be redundant.

This discretisation of the concepts can be interpreted alternatively as the general learning situation with real numbers of limited accuracy or resolution. If d is the number of binary places of accuracy of real numbers, the hypothesis may be of size $\Theta(2^d n)$, showing how predictability from positive examples breaks down when concepts are defined by points at any position on the real line.

These observations indicate that predictability can depend crucially on our expectations of the form taken by a concept.

6.2. Prediction from Positive and Negative examples

We have seen that for essentially information-theoretic reasons, spheres around sets of points on the line under the Hausdorff metric cannot be predicted from positive examples only. With access to negative examples we can be sure of having enough information for prediction, since we have seen that the V-C dimension is polynomial. For this concept class the prediction problem is, moreover, tractable, and in this section we exhibit a prediction algorithm.

The algorithms that follow use a basic greedy set cover procedure in finding hypotheses that "explain" as many as possible of the examples, repeating until all examples have been explained. To show that not too many hypotheses are needed we will re-use the lemma on Occam hypotheses from Greedy Set Cover from the section on learning components from scenes.

Theorem 6.6: *The concept class of fixed radius r of sets of points on the real line under the Hausdorff metric is predictable from positive and negative examples.*

Proof: The following Occam algorithm gives a consistent hypothesis, which means that it is a good prediction algorithm.

Algorithm: ELIMINATE-NEGATIVES

POS := positive examples; NEG := negative examples;
repeat
  find hypothesis H_1 using method 1;
  find hypothesis H_2 using method 2;
    { *These methods are explained below. H_1 and H_2 will be consistent with all positive examples but not in general with all negatives.* }
  If $|NEG - H_1| > |NEG - H_2|$ then $H := H_1$ else $H := H_2$;
    { *Choose the hypothesis that omits the larger number of negatives* }
  remove from NEG all elements not in H;
until $NEG = \varnothing$.

The idea of the algorithm is based on the observation that it is possible in polynomial time to find a concept consistent with all the positive examples and a significant fraction of the negatives. Methods 1 and 2 are two procedures for finding such a hypothesis, at least one of which is guaranteed to eliminate sufficiently many negatives. Hence if we remove the negative examples accounted for by the concept obtained, we can repeatedly apply this technique until

all negative examples have been eliminated. The intersection of all concepts obtained by doing this is consistent with the sample, and assuming that enough negatives are removed at each stage, it is an Occam algorithm. It will be shown that the hypothesis obtained eliminates at least $\frac{m}{2(n+1)}$ negative examples.

We have seen how to generate a hypothesis in this concept class which is consistent with a set of positive examples. Let I be the allowable region for concept points, as in algorithm *1D-POS-HYPOTHESIS*. (That is, I is the set of all points on the real line which are within r of some point in every example.) We now show how to choose this concept in order to avoid at least $\frac{m}{2(n+1)}$ of a set of m given negative examples.

We note that under the Hausdorff metric, an example e belongs to a concept $B_r(s)$ (the r-ball around object s) if and only if

1.) Each point in s is within r of some point in e, and

2.) Each point in e is within r of some point in s.

Every member of a set *NEG* of negative examples must be negative as a result of failing condition (1) or condition (2). Hence either at least $m/2$ elements of *NEG* fail condition (1) or at least $m/2$ of them fail condition (2) (or both). Method 1 accounts for at least $m/2n$ negatives if condition (1) is failed by most of them, and method 2 accounts for at least $\frac{m}{2(n+1)}$ if condition (2) is failed by most of them. So both can be tried to obtain a hypothesis H accounting for at least $\frac{m}{2(n+1)}$ negatives.

Method 1.

We assume that for at least $m/2$ elements of *NEG* there is a point in s more than r from the nearest point on any of these objects. Since s has n points, at least one point of s is more than r from the nearest point in $m/2n$ of these objects. By sorting all the points in the examples and doing exhaustive search in I we can find a suitable point in time $O\big((mn)^2 \log(mn)\big)$. The other points forming s can be constructed as in algorithm *1D-POS-HYPOTHESIS* in chapter 5.

Method 2.

We assume in this case that for at least $m/2$ members of *NEG* there are points more than r from the nearest point in s. These points may occur to the right or to the left of all points in s, or in between two consecutive points in s. Out of these $n+1$ possibilities at least one must account for at least $\frac{m}{2(n+1)}$ members of *NEG*.

We can check for a suitable s with points in at least $\frac{m}{2(n+1)}$ members of *NEG* to its left by constructing s as in algorithm *1D-POS-HYPOTHESIS* and similarly for the right-hand side, but in the reverse direction. Checking for points occurring between two consecutive points of e is done using the algorithm *FIND-INTERVAL* below. This is basically an exhaustive search on the set of possible pairs of consecutive points which are as far apart as possible while still being suitable for inclusion in a hypothesis consistent with the positive examples:

Algorithm: FIND-INTERVAL

$\mathcal{P}$:= all points contained in members of POS, the positive examples
I := all points within r of some point in each positive example
repeat
  p_1 := leftmost point in $\mathcal{P}$
  p_2 := leftmost point in $\mathcal{P} \setminus \{p_1\}$
  q_1 := leftmost point in I within r of p_1
  q_2 := rightmost point in I within r of p_2
  if $q_2 - r > q_1 + r$ then $J := (q_1 + r, q_2 - r)$ else $J := \varnothing$
  N^- := members of NEG having a point in J
  $\mathcal{P} := \mathcal{P} \setminus \{p_1\}$
until $|N^-| \geq m/2(n+1)$ or q_2 is rightmost point in I
construct e containing q_1 and q_2 consecutively
end.

The above algorithm works as follows. p_1 and p_2 are the coordinates of two consecutive points occurring in positive examples. The loop tests consecutive pairs p_1 and p_2 working from left to right. q_1 and q_2 are two points which could occur in a consistent hypothesis, chosen to be within r of p_1 and p_2 respectively, and as far apart as possible. A consistent hypothesis need contain no points between q_1 and q_2, hence a consistent hypothesis may contain q_1 and q_2 consecutively. A hypothesis containing q_1 and q_2 as consecutive points will be consistent with all negative examples containing points between q_1 and q_2, and more than r distant from both q_1 and q_2. J is the interval in which such points may occur. The algorithm tests whether sufficiently many negative examples contain points in J.

To prove correctness of *FIND-INTERVAL*, suppose that there is a hypothesis s which is consistent with all positive examples, and contains two consecutive points at q_1' and q_2', $q_2' > q_1'$, such that there are at least $\frac{m}{2(n+1)}$ negative examples containing points between q_1' and q_2', more than r distant from q_1' and q_2'. Then $q_2' - q_1' > 2r$. Let p_1 be the rightmost point in $\mathcal{P}$ within r of q_1' and let p_2 be the leftmost point in $\mathcal{P}$ within r of q_2'. Then $p_2 > p_1$, and p_1 and p_2 are consecutive points in $\mathcal{P}$. When *FIND-INTERVAL* tests p_1 and p_2, the associated points q_1 and q_2 will satisfy $q_1 \leq q_1'$ and $q_2 \geq q_2'$, hence a hypothesis will be found consistent with the $\frac{m}{2(n+1)}$ negative examples.

To show that *ELIMINATE-NEGATIVES* is an Occam algorithm, we have $|NEG|, |POS| = O(m)$. Each hypothesis H reduces the size of NEG by a factor of at least $\frac{1}{2(n+1)}$. Hence an increase in the sample size by a factor of $\frac{2n+3}{2n+2}$ will increase the complexity of the hypothesis by a constant. So given a sample of size $O(m)$, the number of concepts whose intersection forms the hypothesis is $O(\log_{\frac{2n+3}{2n+2}} m) = O(\log_{(1+\frac{1}{2n+2})} m)$, so the complexity of the hypothesis is $O(n)O(\log_{(1+\frac{1}{2n+2})} m)$.

By lemma 4.6, this is $O(n^c m^\alpha)$ for suitable $c \in \mathbb{N}$ and any $0 \leq \alpha < 1$.

From theorem 3.3, we have the upper bound $v \leq O(n)(\log v + O(n^\omega) \log sd)$, where $v(n)$ is V-C dimension, n is number of free variables, $s(n)$ is length of expression, $d(n)$ is degree of the polynomials, ω is the (constant) number of quantifiers.

Hence here for arbitrarily small $\alpha > 0$,

$$v \leq O(n^c m^\alpha)\big(\log v + O((n^c m^\alpha)^\omega) \log(n^c m^\alpha sd)\big)$$

Consequently this is an Occam algorithm. □

Theorem 6.7: *The concept class of fixed radius r of sets of points in the plane under the Hausdorff metric is predictable from positive and negative examples.*

Proof: We may use a strategy similar to the previous one, looking for hypotheses consistent with all positive and a sufficiently large fraction of the negative examples. Using the algorithm *ELIMINATE-NEGATIVES*, we need to define method 1 and method 2 analogously for sets of points on the plane. It will be shown that at least one of these methods accounts for $\frac{1}{n^2}$ of the negative examples.

In searching for a suitable hypothesis we discretise the problem by dividing the plane into the regions formed by the r-circles around every point in each example. There are $O((mn)^2)$ points of intersection of these circles. Each point characterises a region of unique proximity status (ie. within distance r or otherwise) with respect to all points other that the two whose circles cause the intersection point. (Degeneracies such as three circles meeting at a point may cause duplication of representative points, but this is not a problem.) From these points we may construct a set D (for "discretisation") of size $O((mn)^2)$ of representative points for each region.

We consider first how to find a hypothesis consistent with the positive examples only.

Let I be the set of points in D which are within one unit of a point on every positive example. I corresponds to the intersection I in algorithm *1D-POS-HYPOTHESIS* and can be found in polynomial time by testing all members of D. A consistent hypothesis is a subset I^- of I such that every positive example point is within r of some member of I^-. We know that there exists such a subset of size n, namely the centre of the target concept (modulo discretisation: every point in that object must be shifted to a point in D representing it). Then the greedy set cover algorithm, choosing each time the point in I that is close to the greatest number of positive example points unaccounted for, yields a set I^- of size $O(\log_{1+n^{-1}}(mn)) = O(n \log nm)$. We now give the two-dimensional equivalents of methods 1 and 2.

Method 1.

To find a hypothesis which is consistent with the positive examples, but whose centre contains a point greater than r distant from all points on at least a fraction $\frac{1}{2n}$ of the negative examples (assuming one exists).

Construct a hypothesis consistent with the positive examples, as explained above. Then check every member of I for a point which is at least r distant from all points on at least $\frac{1}{2n}$ of the negative examples. I has polynomial size and testing each point takes polynomial time. When a suitable point has been located, add it to the hypothesis. The hypothesis now satisfies all required conditions.

Method 2.

To find a hypothesis which is consistent with the positive examples, but whose points are all greater than r distant from some point on at least a fraction

$\frac{1}{2n^2}$ of the negative examples. (This assumes that most negative examples are negative due to having points greater than r distant from any point defining the target concept.)

Claim: Assuming method 2 is applicable, there exists a region bounded by at most two arcs of r-circles and two vertical lines, which contains points lying in at least a fraction $\frac{1}{2n^2}$ of the negative examples, and no points within r of some consistent hypothesis of size n.

Before this claim is proved, let us describe the algorithm that exploits this fact.

We can discretise the set of vertical lines that need to be considered according to how they divide D into those lying to the left and those lying to the right. There are at most mn ways to divide D into two with a vertical line. Let L be a set of vertical lines, $|L| \leq mn$, each of which uniquely represents a division of the set D (all possible divisions represented.)

Algorithm FIND-REGION

$size := mn$;
$\mathcal{E}$:= all points contained in examples;
$\mathcal{P}$:= points contained in the positive examples;
For all pairs $\{p_1, p_2\} \in \mathcal{E}$ and all pairs $\{l_1, l_2\} \subseteq L$ do
  begin
  R := points in I outside the r-circles around p_1, p_2 and between l_1 and l_2;
  If R contains points in $\geq \frac{1}{2n^2}$ of the negative examples then
    begin
    Using greedy set cover, look for a set $I^- \subseteq I \setminus R$ such that for all $p \in \mathcal{P}$ there is a point $q \in I^-$ with $d(p,q) \leq r$;
    If $|I^-| \leq size$ then ($S := I^-$; $size := |I^-|$);
    end;
  end;
hypothesis H := r-ball around object defined by S;
end.

The above algorithm does an exhaustive search over all possible subsets of D bounded by two vertical lines $l_1, l_2 \in L$ and the exteriors of two r-circles centred at p_1, p_2. There are $O((mn)^4)$ of these. So for the points of D this

constitutes a test of all regions R bounded by two vertical lines and two r-circles. The test for R is whether it contains points from $\frac{1}{2n^2}$ of the negative examples and if so whether there is a consistent hypothesis isolating it (ie. all its points are $> r$ distant from it). There should be one which admits a sufficiently small consistent hypothesis found by greedy set cover, namely one which the target hypothesis isolates. The algorithm picks the best consistent hypothesis found in this way.

Assuming such a region exists, this algorithm will find a hypothesis consistent with the positive examples and at least $1/n^2$ negatives in time $O((mn)^6)$, allowing time $O((mn)^2)$ for greedy set cover. The smallest (lowest number of points) hypothesis will be returned.

Proof of claim: A region R exists which satisfies the properties given:

Let T be the set of n points defining the target concept, and suppose that most ($\geq \frac{m}{2}$) of the negative examples have points lying outside the union of the r-circles around each point of T. There are $O(n^2)$ vertical lines which are either tangential to these circles or which pass through a point of intersection of two of them. The n circles and $O(n^2)$ lines divide the region outside the circles into $O(n^2)$ subregions. A little thought will convince the reader that each of these regions is bounded by at most two circles and two lines (see fig. 6.1.) Since points in $\geq m/2$ negative examples are contained in these regions, the resulting hypothesis will account for $O(m/2n^2)$ negative examples.

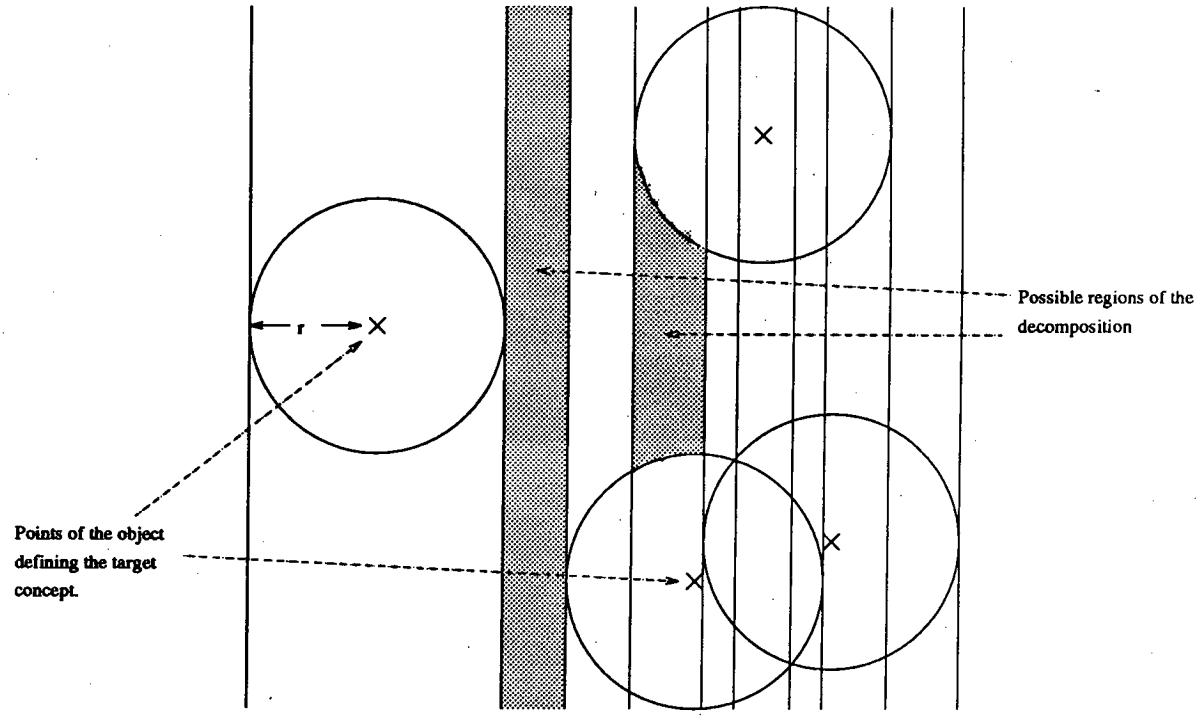

figure 6.1. Decomposition of region containing points in $\geq \frac{m}{2}$ negatives into simple sub-regions

To show that *ELIMINATE-NEGATIVES* in two dimensions is an Occam algorithm:

The size of the hypothesis is the product of the number of concepts generated for the intersection and the size of these concepts.

$O(\log_{1+n^{-2}} m)$ concepts are needed, of size $O(n^2 \log_{1+n^{-1}} m)$.

Hence the size of the hypothesis is $O\big(n^2(\log_{1+n^{-1}} m)(\log_{1+n^{-2}} m)\big)$.

Using the observation accompanying lemma 4.6, this product is $O(n^c m^\alpha)$ for suitable values of c and α. By a similar argument to that in theorem 6.6, the V-C dimension is consequently small enough for Occam learning. □

6.3. A Non-predictability Result

In this final section, we show that some geometrical concept classes are likely to be hard to predict. These are classes of planar figures for which the distance between two of them is minimised over a set of isometries which includes rotations. For example, the result applies to planar polygons under the Hausdorff or Fréchet metrics, and also to finite sets of points in the plane under the Hausdorff metric, provided that these figures are assumed to be rotatable. Thus this can be viewed as a result of learning the "shape" of an object rather than a particular depiction of it.

It is convenient to prove the result for planar polygons under the Hausdorff metric, after which it may be readily observed to apply to the other classes mentioned above.

The proof uses the "prediction-preserving reduction" method of Pitt and Warmuth [PW90]. The prediction problem used is that of predicting a class of boolean formulae shown to be hard to predict in [J91], and this non-predictability is based on the assumed non-predictability of monotone disjunctive normal form formulae, defined below.

Definition: A *monotone disjunctive normal form (MDNF)* formula is a DNF formula where each clause is a monomial, that is a conjunction of un-negated boolean variables.

Definition: *Translation-closed Monomials (TCMs)* are MDNF formulae with the following additional restriction. Denote the variables $\{v_0, v_1, \ldots, v_{n-1}\}$. Then the formula has n monomial clauses generated from one monomial by adding (modulo n) i, for $i = 0, 1, ..., n-1$ to the index of each variable in that monomial.

For example, over the variables $\{v_0, v_1, ..., v_5\}$, the TCM associated with the monomial $v_0 \wedge v_1 \wedge v_3$ is

$$(v_0 \wedge v_1 \wedge v_3) \vee (v_1 \wedge v_2 \wedge v_4) \vee (v_2 \wedge v_3 \wedge v_5) \vee (v_3 \wedge v_4 \wedge v_0) \vee (v_4 \wedge v_5 \wedge v_1)$$
$$\vee (v_5 \wedge v_0 \wedge v_2)$$

Theorem 6.8: [J91] *The concept class TCM is polynomially predictable if and only if the concept class DNF is.*

Proof: The prediction problem for TCM is shown to be equivalent to that for MDNF, which in turn is equivalent to that for DNF, using a theorem of [KLPV87].

TCM is a restriction of MDNF, which gives the reverse implication. A prediction-preserving reduction is given from MDNF to TCM, to show that a polynomial prediction algorithm for TCM would yield a such an algorithm for MDNF. □

Theorem 6.9: *If MDNF is hard to predict, then the concept class of spheres of fixed radius defined by planar polygons under the Hausdorff metric minimised over rotations, is hard to predict.*

Proof: We exhibit a prediction-preserving reduction from the prediction problem for TCMs to that for the geometrical concept class.

The reduction is from n-variable TCMs to $3n$-sided polygons. The idea will be that an n-variable TCM is descriptive of a set of features of a polygon with $3n$ sides, belonging to a particular subclass of $3n$-gons.

Basically, a TCM over n variables is mapped to a polygon which conforms to a pattern having rotational symmetry of order n, and similarly for a set of value assignments to n variables. Then there are essentially n positions for a polygon corresponding to value assignments which need to be tested to see if the polygons are close under the Hausdorff metric. These correspond to testing satisfaction of each of the n clauses by a set of value assignments.

The reduction is best described using an example. Suppose that we are trying to find a consistent hypothesis for TCMs over six variables $x_0, ..., x_5$. This problem is to be reduced to that of finding a consistent hypothesis for a set of 18-gons.

A concept is defined by a polygon of the form shown in fig. 6.2, where between two adjacent spikes there are two possible positions for the intermediate edge. If x_i is included in the monomial generating the corresponding formula, then its corresponding edge is the outer of the two, otherwise it is the inner one. It can be seen that all TCMs generated by monomials occurring in a particular TCM will be mapped to polygons that are identical modulo rotation about their centre.

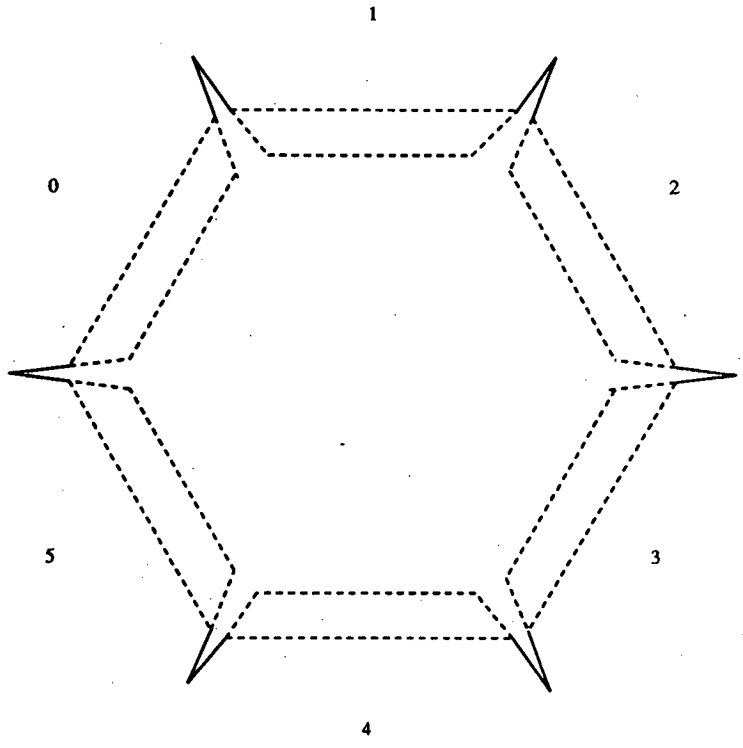

figure 6.2. General form of Polygon defining a concept for 6 variables

For example the TCM generated by $(x_0 \wedge x_1 \wedge x_3)$ is mapped to a concept defined by the following polygon (fig. 6.3):

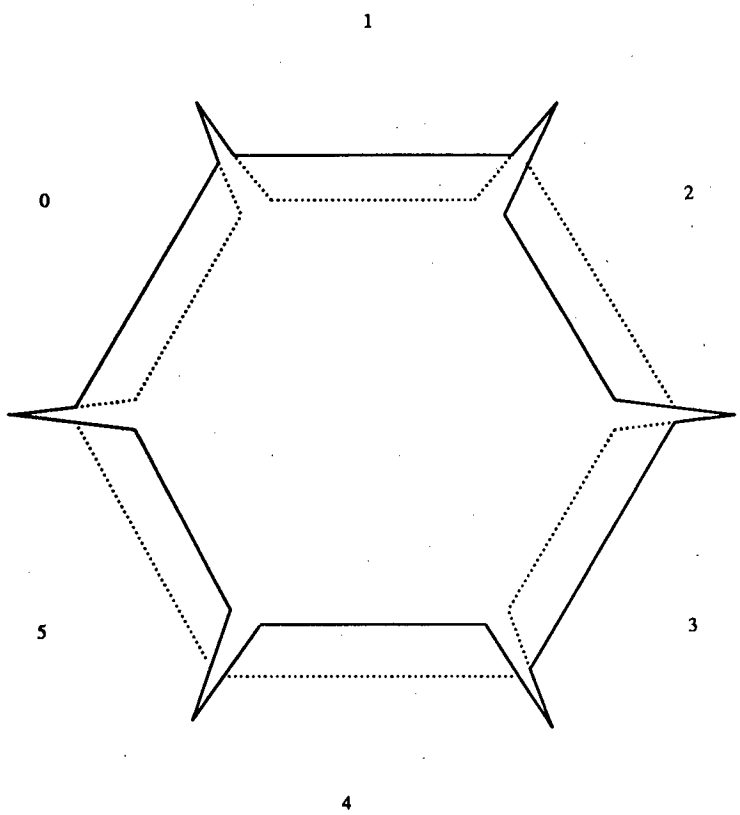

figure 6.3.

An assignment of values to the boolean variables $x_0, ..., x_5$ is encoded by a polygon fitting a similar template, in which true variables are assigned sides consistent with any position of a corresponding side in a hypothesis, while assignments to false have corresponding sides on the inner dotted line, which is more than one unit away from the outer dotted line, and consequently inconsistent. For example, $x_1, x_2, x_3, x_4 = T$, $x_5, x_0 = F$ is represented by the following polygon (fig. 6.4):

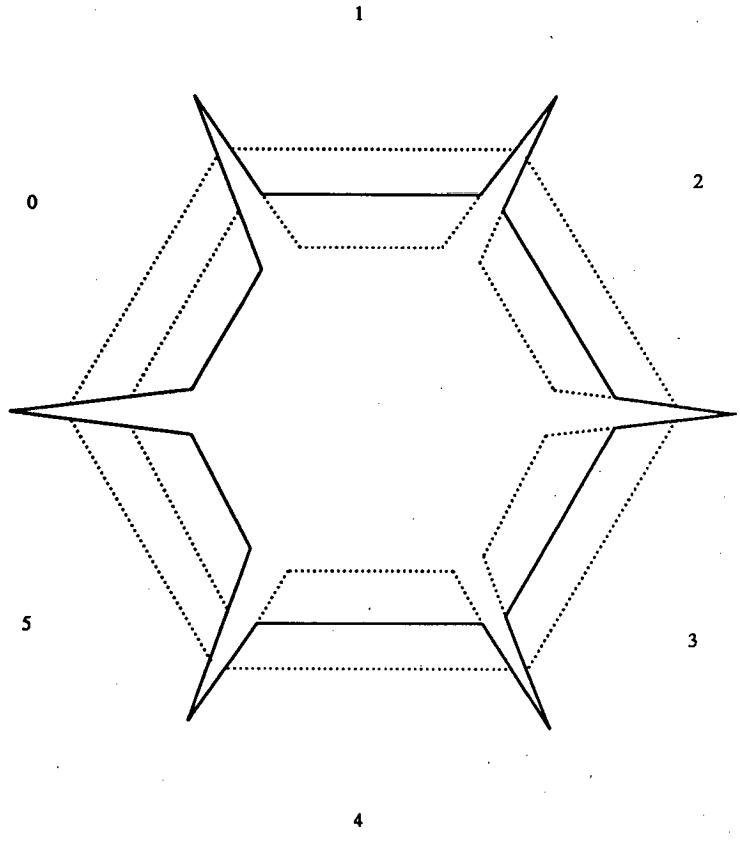

figure 6.4.

Again, note that by rotational symmetry, this is equivalent to any assignment of four consecutive variables to true and the other two to false, since polygons are identical modulo rotation. Also note that this reduction does not make any use of translations, whose effect on the situation is eliminated by making the length of the protruding spikes on the example polygons one unit less than the ones on the template, so that there are only six positions worth considering for each example, generated by rotation. □

It can be seen that this reduction is valid if polygons are compared under the Fréchet metric, instead of the Hausdorff metric. It is also fairly easy to see that a similar construction can be used to give a similar result for finite sets of points in the plane, under the Hausdorff metric. Hence we have a non-predictability result which is in direct contrast to the earlier predictability result where sets of points are immobile.

Chapter Seven

Conclusions and Open Problems

In this thesis we have found out a great deal about what aspects of a geometrical pattern learning problem may make it intractable. Some of the results were counterintuitive to the author, such as the non-learnability of fixed geometrical objects under the Hausdorff metric, or the fact that it is hard to find a concept of given radius consistent with a sample of positive examples under the Hausdorff metric modulo translations.

We have seen the importance of negative examples of a concept under various assumptions on the form of a learning problem. An important conclusion emerging from the prediction results is that learning is often easier when viewed as a means of acquiring the ability to classify correctly, rather than to obtain some underlying truth.

The results presented here have several gaps. We will list the open problems that have arisen:

1.) Theorem 3.3 gives an upper bound on the V-C dimension of concept classes parametrised by $\mathbb{N}$ whose membership test is given by a class of formulae $\Phi_n, n \in \mathbb{N}$, in the first-order theory of the real numbers. This upper bound is:

$$v \leq O(n)(\log v + O(n^{\omega}) \log(sd))$$

where $s(n)$ is the length of Φ_n, $d(n)$ is the degree of the polynomials it contains and ω is the quantification depth.

The length and degree bound may not be superexponential in n — a concept class defined by a class of sufficiently large formulae can have arbitrarily large V-C dimension. However it would be of interest to know whether the exponential dependence on the number of quantifiers in Φ_n can be improved. Results giving tighter upper or lower bounds would be interest, since they would show whether quantifiers can extract much information from real numbers. Note that if Φ_n is any polynomial-time algorithm, its associated concept class *may* have exponential V-C dimension.*

2.) In our reductions, the maximum distance between two points in a geometrical figure P is usually proportional to the number of points defining P.

* This seems to require however that the $\lfloor\ \rfloor$ operator (rounding real numbers down to the nearest integer) be computable in constant time.

The exception is sets of point under the minimax metric, for which this distance may be $O(1)$, and points were only separated in the interest of clarity. It may interesting and relevant to take this notion of size of a geometrical figure into consideration.

For example, let $\mathcal{X}$ be sets of points on the line, $\mathcal{C}$ be spheres of points in $\mathcal{X}$ under the reversed directed Hausdorff distance minimised over translations. (Hence a concept is basically defined by a "scene", and examples are sets of points which can be translated so as to be close to components of that scene.) Assuming unit radius, if there is some limit L placed on the distance between points in examples, the concept defined by L points at unit intervals will contain any sample of positive examples. It is not so clear however how to find a concept consistent with "sparse" positive examples.

3.) Our results for learning two-dimensional objects are all valid only for the L_1 or L_∞ norms. We do not know whether it is possible to learn polygons under the directed Hausdorff distance with the Euclidean norm, or whether an NP-hardness result holds for learning polygons under the Hausdorff distance with the Euclidean norm.

Some results are affected by the choice between the L_2 and L_1/L_∞ norms. In particular, while it is possible to predict planar point sets under the minimax metric with the L_∞ norm, this cannot be done under the L_2 norm. It may be appropriate to develop a weaker form of learning that is not sensitive to these changes.

4.) Figure 5.1 illustrates a situation where a consistent hypothesis may contain a point more than $3r$ distant from the centre of a target concept. However a hypothesis may be chosen with better worst-case performance guarantees. A concept defined by a point close to the sample need not contain any point more than $2r$ distant from the centre of any target concept. It is unknown whether there is a concept class which lacks this property.

5.) It is not clear how to generalise the algorithms that have been presented for sets of points in the line, to polygons in the plane. Most of the algorithms only appear to be naturally extendible to sets of points in the plane. Examples are the prediction algorithm for point sets under the Hausdorff metric, and learning under the directed Hausdorff distance.

6.) Is it possible to predict sets of points modulo translations, rather than rotations? Only rotations are used to show apparent non-predictability of

point sets or polygons in the plane under the Hausdorff metric modulo a class of isometries. But the prediction algorithm given for fixed point sets under the Hausdorff metric does not work for point sets under the Hausdorff metric modulo translations, since this requires a hypothesis consistent with a sets of positive examples to be found, which the main result of chapter five suggests is hard.

7.) It may be worthwhile to study learning under the other metrics mentioned in chapter 2.

References

[A87] D. Angluin (1987). Queries and Concept Learning. *Machine Learning,* **2** pp. 343-370.

[ABB91] H. Alt, B. Behrends, J. Blömer (1991). Approximate Matching of Polygonal Shapes. *Procs. of the 7th annual ACM Symposium on Computational Geometry* pp. 186-193.

[ACHKM89] E.M. Arkin, L.P. Chew, D.P. Huttenlocher, K. Kedem, J.S.B. Mitchell (1989). An efficiently computable metric for Comparing Polygonal Shapes. *Tech. Rept. TR 89-1007 May 1989 Dept. of Computer Science, Cornell University.* also *IEEE Trans. Pattern Analysis and Machine Intelligence* **13** No. 3, 1991.

[AS83] D. Angluin, C.H. Smith (1983). Inductive Inference: Theory and Methods. *Computing Surveys,* **15** No. 3, pp. 237-269.

[BBM90] Ben-David, Benedek, Mansour. A Parametrization Scheme for Classifying Models of Learnability. *Conference on Learning Theory,* 1989.

[BEHW87] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth (1987). Occam's Razor. *Information Processing Letters* **24** pp. 377-380.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth (1989). Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the Association for Computing Machinery,* **36** No. 4, pp. 929-965.

[BG90] V. Bruce, P.R. Green (1990). *Visual Perception: Physiology, Psychology and Ecology.* 2nd edition 1990, Lawrence Erlbaum Assocs, publishers

[BK80] J. Borwein, L. Keener (1980). The Hausdorff Metric and Chebyshev Centres. *Journal of Approximation Theory* **28** pp. 366-376.

[BSS89] L. Blum, M. Shub, S. Smale (1989). On a Theory of Computation and Complexity over the Real Numbers: NP-Completeness, Recursive Functions and Universal Machines. *Bull. of the American Math. Soc.* **21** No. 1, pp. 1-46.

[F06] M. Fréchet (1906). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Mathematico di Palermo,* **22**, pp. 1-74.

[F74] H. Freeman (1974). Computer Processing of Line-Drawing Images. *Computing Surveys,* **6** No. 1, pp. 57-98.

[G67] M. Gold (1967). Landuage Identification in the Limit. *Information and Control.* **10** pp. 447-474.

[G83] P.M. Gruber (1983). Approximation of convex bodies. *Convexity and its Applications*, eds. P.M. Gruber and P.M. Willis, pp. 131-162 Birkhäuser Verlag 1983

[G91] M. Godau (1991). A natural metric for curves — Computing the distance for polygonal chains and approximation algorithms. *Symposium on Theoretical Aspects of Computer Science 1991,* pp. 127-136

[GJ79] Garey, Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York: W.H. Freeman and Company

[H88] D. Haussler (1988). Quantifying Inductive Bias: AI learning algorithms and Valiant's framework. *Artificial Intelligence*, **36** (2), pp. 177-222.

[HKLW91] D. Haussler, M. Kearns, N. Littlestone, M.K. Warmuth (1991). Equivalence of models for polynomial learnability. *Information and Computation* **95** No. 2, pp. 129-161. *(see also Conference on Learning Theory, 1988 pp. 42-55)*

[HKR91] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge (1991). Comparing Images Under the Hausdorff Distance Under Translation. *Tech. Rept.* TR91-1211, June 1991. *Dept. of Computer Science, Cornell University.*

[HKS91] D.P. Huttenlocher, K. Kedem, M. Sharir (1991). The Upper Envelope of Voronoi Surfaces and its Applications. *Tech. Rept., Dept. of Computer Science, Cornell University.* Feb. 1991.

[HLW88] D. Haussler, N. Littlestone, M.K. Warmuth (1988). Predicting 0,1-functions on randomly drawn points. *Conference on Learning Theory, 1988,* pp. 280-296. *also in Procs. of the 29th ACM symposium on Foundations of Computer Science* pp. 100-109

[HSW89] D. Haussler, R. Sloan, M.K. Warmuth (1989). Learning Nested Differences of Intersection-closed Concept Classes. *Conference on Learning Theory 1989,* pp. 41-56.

[HW87] D. Haussler, Welzl (1987). Epsilon-nets and simplex range queries. *Discrete Computational Geometry* **2**, pp. 127-151.

[ISI89] K. Imai, S. Sumino, H. Imai (1989). Minimax Geometric Fitting of Two Corresponding Sets of Points. *Procs. of the 5th annual ACM symposium on Computational Geometry* pp. 266-275.

[J91] M. Jerrum (1991). Simple Translation-invariant Concepts are Hard to Learn. *Internal Report* CSR-12-91, Jan. 1991. *Dept. of Computer Science, Edinburgh University. to appear in Information and Computation*

[KLPV87] M. Kearns, M. Li, L. Pitt, L. Valiant (1987). On the learnability of Boolean formulae. *Procs. of the 19th ACM Symposium on Theory of Computing,* pp. 285-295.

[KV89] M. Kearns, L. Valiant (1989). Cryptographic Limitations on Learning Boolean Formulae and Finite Automata. *Procs. of the 21st ACM Symposium on Theory of Computing,* pp. 289-295.

[M64] J. Milnor (1964). On the Betti Numbers of Real Varieties. *Procs. of the American Mathematical Society,* **15**, pp. 275-280.

[M87] D. Mumford (1987). The problem of robust shape descriptors. *First Intl. Conf. on Computer Vision* IEEE Computer Soc. Press, 1987, pp. 602-606.

[N87] B.K. Natarajan (1987). On Learning Boolean Functions. *Procs. of the 19th Annual Symposium on Theory of Computation,* pp. 296-304.

[N91a] B.K. Natarajan (1991). *Machine Learning: A Theoretical Approach.* Morgan Kaufman Publishers, Inc. ISBN 1-55860-148-1

[N91b] B.K. Natarajan (1991). Probably approximate learning of sets and functions. *SIAM Journal of Computing* **20** No. 2, pp. 328-351.

[PV88] L. Pitt, L. Valiant (1988). Computational Limitations on Learning from Examples. *Journal of the ACM,* **35** No. 4, pp. 965-984.

[PW90] L. Pitt, M.K. Warmuth (1990). Prediction-preserving Reducibility. *Journal of Computer and System Sciences* **41** pp. 430-467.

[R91] G. Rote (1991). Computing the minimum Hausdorff distance between two point sets on a line under translation. *Information Processing Letters* **38** No. 3, pp. 123-127.

[KP89] M. Kearns, L. Pitt (1989). A polynomial-time algorithm for learning k-variable pattern languages from examples. *Conference on Learning Theory,* 1989, pp. 57-71.

[R92] J. Renegar (1992). On the Computational Complexity and Geometry of the First-Order Theory of the Reals. Part 1 (of 3). *Journal of Symbolic Computation* **13**, pp. 255-299.

[VC71] V.N. Vapnik, A.Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* **16**, No. 2 pp. 264-280.

[V84] L. Valiant (1984). A Theory of the Learnable. *Communications of the ACM*, **27** No. 11, pp. 1134-1142.

[V85] L. Valiant (1985). Learning Disjunctions of Conjunctions. *Procs of the 9th International Joint Conference on AI*, pp. 560-566.

[V91] L. Valiant (1991). A View of Computational Learning Theory. *NEC Research Symposium: Computation and Cognition (ed. C.W. Gear), SIAM, Philadelphia, 1991.*

[W74] R.M. Wharton (1974). Approximate Language Identification. *Information and Control* **26**, pp. 236-255.

[WD80] R.S. Wenocur, R.M. Dudley (1981). Some Special Vapnik-Chervonenkis Classes. *Discrete Mathematics* **33** pp. 313-318.

Notation and Conventions

Conventional Symbols in the text:

What follows is a list of symbols and what they usually denote in this thesis. Some, such as ϵ and δ, are general conventions, others are chosen by the author.

ϵ: error bound on hypothesis

δ: uncertainty bound

C: A concept

T: Target concept

$\mathcal{X}$: Instance domain

$\mathcal{C}$: A concept class

$\mathcal{H}$: A hypothesis class (also called hypothesis space)

V-C dimension: Vapnik-Chervonenkis dimension

$dim(\mathcal{C})$: Vapnik-Chervonenkis dimension of $\mathcal{C}$

m: sample size

n: complexity of an example or concept

CHP: consistent hypothesis problem

POS: positive examples in a sample

NEG: negative examples in a sample

r-interval: interval of length r

r-circle: circle of radius r

An *object* is a geometrical figure, an element of some instance domain under consideration. A concept *defined by* an object x is a concept consisting of a sphere in a metric space, whose centre is x. Hence we may refer to the "radius of a concept", usually denoted r, and if all concepts in a concept class have radius r we say that r is the radius of the concept class. In learning "under a metric", we refer to learning a concept class whose concepts are spheres based on that metric.

Diagrammatic conventions:

Most of the diagrams in this thesis in chapters four and five depict collections of configurations of finite sets of points on the real line. While they are not essential to the proofs, it is anticipated that they will aid understanding.

The following conventions are used. Each set of points on the line is shown as a horizontal line with crosses marking the positions of the points. The scale is uniform within any diagram, and each vertical dotted line represents a single position on the real line, and cuts each depiction of the real line in a diagram at that position. A dashed section of any horizontal line indicates a section of the real line that is too long to include conveniently in a diagram.

Each line is labelled on the left with a symbol (occurring in the relevant proof) indicating an example or general form that a consistent hypothesis may take. In the case of an example, there is also a $+$ or $-$ sign, indicating whether the examples is positive or negative. In the case of a hypothesis, there may appear pairs of brackets as well as crosses along the line itself. A cross indicates a position where a point must occur in a consistent hypothesis, and a pair of brackets enclose an interval where at least one point must occur. If two intervals are labelled "xor", then points must be contained in one but not both of those intervals.