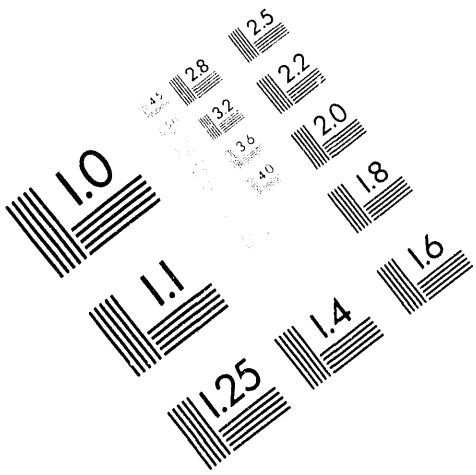
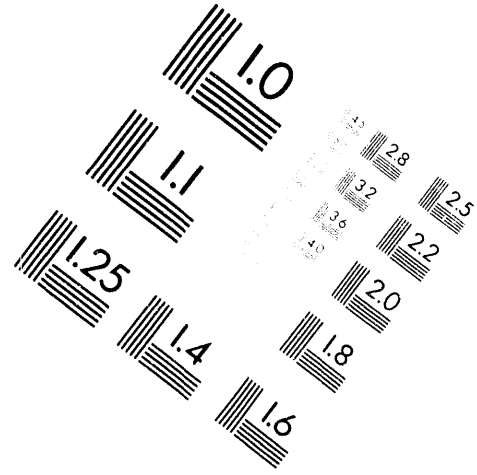


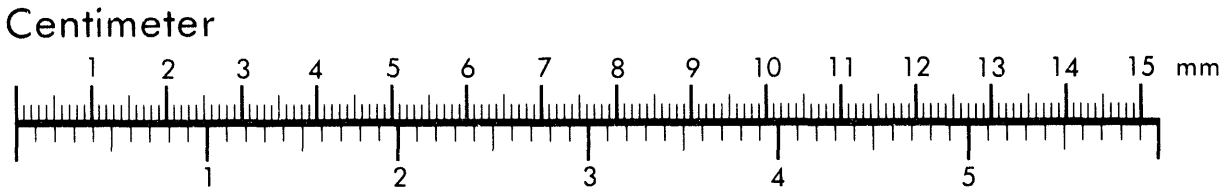
AIIM

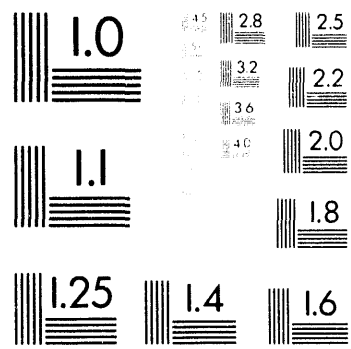
Association for Information and Image Management

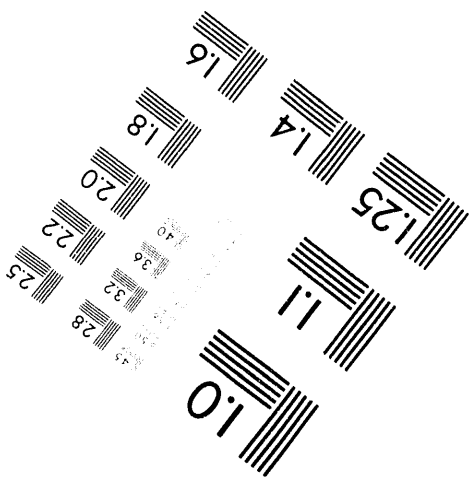
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202

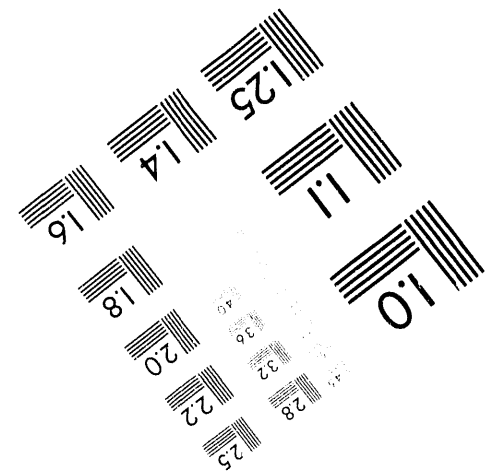








MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.



1 of 1

Learning One-Dimensional Geometric Patterns Under One-Sided Random Misclassification Noise

Paul W. Goldberg*
Department 1423
Sandia National Laboratories, MS 1110
P.O. Box 5800
Albuquerque, NM 87185-1110
pwgoldb@cs.sandia.gov

Sally A. Goldman†
Dept. of Computer Science
Washington University
St. Louis, MO 63130
sg@cs.wustl.edu

Abstract

Developing the ability to recognize a landmark from a visual image of a robot's current location is a fundamental problem in robotics. We consider the problem of PAC-learning the concept class of geometric patterns where the target geometric pattern is a configuration of k points in the real line. Each instance is a configuration of n points on the real line, where it is labeled according to whether or not it visually resembles the target pattern.

We relate the concept class of geometric patterns to the landmark recognition problem and then present a polynomial-time algorithm that PAC-learns the class of one-dimensional geometric patterns when the negative examples are corrupted by a large amount of random misclassification noise.

1 Introduction

Developing the ability to recognize a landmark from a visual image of a robot's current location is a fundamental problem in robotics. We consider the problem of PAC-learning the concept class of geometric patterns where the "target" geometric pattern is a configuration of k points in the real line. Each instance is a configuration of n points on the real line, where it is labeled according to whether or not it visually resembles the target pattern. To capture the notion of visual resemblance we use the *Hausdorff metric* (for example, see Gruber [Gru83]). Informally, two geometric patterns P and Q resemble each other under the Hausdorff metric, if every point on one pattern is "close" to some point on the other pattern.

As a motivation of this problem consider the problem of recognizing from a visual image from a robot's current location whether or not it is in the vicinity of a known landmark (where a landmark is a location that is visually different from other locations). Such an algorithm is needed for navigation where the navigation is performed by planning a path going between known landmarks, tracking the landmarks as it goes. Because of inaccuracies in effectors and possibly errors in the robot's internal map, when the robot believes it is at landmark L, before heading to the next landmark it can check that it is really in the vicinity of L. Then adjustments can be made if the robot is not at L by either re-homing to L and/or updating its map. We can apply our algorithm to learn geometric patterns to this problem by converting the visual image the robot has into a one-dimensional geometric pattern.

The main result of this paper is a polynomial-time algorithm that PAC-learns the class of one-dimensional geometric patterns when the negative examples are corrupted by a large amount of random misclassification noise. Our algorithm can learn as long as the noise rate is strictly less than one and the expected number of truly positive examples is greater than the expected number of false positive examples. The time and sample complexity are polynomial in the inverse of the amount by which the noise rate is less than 1, and the inverse amount by which the ratio of true positive examples to total positive examples is greater than 1/2.

An interesting feature of this problem is that the target concept is specified by a k-tuple of points on the real line, while the instances are specified by n-tuples of points on the real line where n is potentially much larger than k. Although there are some important distinctions, in some sense our work illustrates a concept class in a continuous domain in which a large fraction of each instance can be viewed as "irrelevant". As in previous work on learning with a large number of irrelevant attributes in the Boolean domain (e.g. Littlestone's work [Lit88]), our algorithm's sample complexity

*This research was performed while visiting Washington University. Currently supported by the U.S. Department of Energy under contract DE-AC04-76AL85000.

†Supported in part by NSF Grant CCR-9110108 and an NSF NYI Grant CCR-9357707.

DISCLAIMER

(the best dual to a mistake-bound) depends polynomially on k and $\lg n$.

This paper is organized as follows. In the next section we formally define the concept class of one-dimensional geometric patterns. Next in Section 3, we describe in more detail how our algorithm could be applied to the landmark recognition problem described above. In Section 4 we describe the learning model and noise model used in this paper. Then, in Section 5, we describe our algorithm to PAC-learn the class of one-dimensional geometric patterns when the data is noise free. Then in Section 6 we present our main result, an algorithm to learn when the negative examples are corrupted by random misclassification noise. Finally we conclude in Section 7.

2 One-Dimensional Geometric Patterns

For the concept class considered here, the instance space $\mathcal{X}_n$ consists of all configurations of n points on the real line[1]. A concept is the set of all configurations from $\mathcal{X}_n$ within unit distance [2] under the Hausdorff metric of some "ideal" configuration of k points. The Hausdorff distance between configurations P and Q, denoted $H(P,Q)$, is:

$$\max\left\{\sup_{p\in P}\left\{\inf_{q\in Q}\{d(p,q)\}\right\},\sup_{q\in Q}\left\{\inf_{p\in P}\{d(p,q)\}\right\}\right\}$$

where d is the Euclidean distance between p and q.

Let P be any configuration of points on the real line. Then we define the concept c_P that corresponds to P by $c_P = \{X \in \mathcal{X}_n \mid H(P,X) \leq 1\}$. Figure 1 illustrates an example of such a concept. Thus one can view each concept as a sphere of unit radius in a metric space where P defines the center of the sphere. For any $X \in \mathcal{X}_n$ such that $X \in c_P$, we say that X is a *positive example* of c_P. Likewise, if $X \notin c_P$, we say that X is a *negative example* of c_P. Furthermore, all configurations of points that resemble the given configuration P are contained within this sphere. Finally, the concept class $\mathcal{C}_{k,n}$ that we study is defined as follows: $\mathcal{C}_{k,n} = \{c_P \mid P$ is a configuration of k points on the real line$\}$. As is standard in the neural network literature, we assume the unit cost model of real computation. (See Valiant [Val91] for a discussion of why this assumption is typically appropriate for geometric domains.)

As discussed in the introduction, n may be significantly greater than k. For example, the learner may be asked to predict if a configuration of 100 points is contained within a sphere defined by 3 points. This consideration is, in some sense, analogous to the notion of irrelevant attributes studied in the Boolean domain. Namely,

[1] Note that throughout this abstract, the word "point" will refer to a single point on the real line, and we shall use the term "a configuration of points" when speaking of an instance.

[2] All results presented here apply if unit distance is replaced by some fixed distance since we can just rescale.

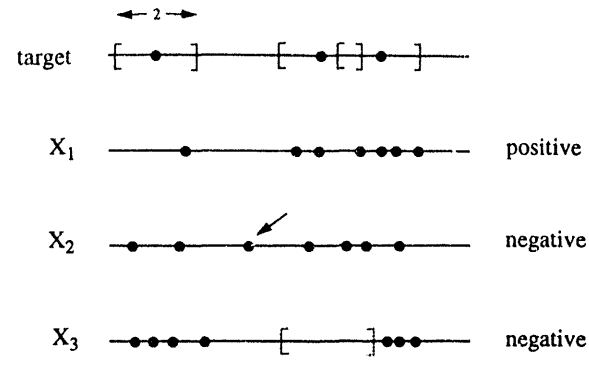


Figure 1: This figure illustrates an example concept from $\mathcal{C}_{3,7}$. The top line shows the target pattern. Around each target point we show an interval that covers all points within unit distance from that point. Every positive example must have every point within one of the above intervals *and* no interval can be empty (e.g. see X_1 above). For an example to be negative, there must be a point in it that is not within unit distance of any target point (e.g. see X_2 above) and/or there are no points in the example near some target point (e.g. see X_3 above).

given any positive (respectively, negative) example from $\mathcal{X}_n$, there exists a subset of k of the n points in that example such that the configuration of these k points is also a positive (respectively, negative) example. However, observe that unlike the Boolean domain, there is no fixed set of points of an instance that are "relevant". Thus if an arbitrary point is removed from an instance it can no longer be determined if that instance was positive or negative before the point was removed.

At first glance, there may appear to be some similarities between $\mathcal{C}_{k,n}$ and the class of the union of at most k intervals over the real line. However, the class of one-dimensional geometric patterns is really quite different (and significantly more complex) than the class of unions of intervals on the real line. One major difference is that for the union of intervals each instance is a single point on the real line, whereas for $\mathcal{C}_{k,n}$ each instance is a set of n points on the real line. Thus the notion of being able to independently vary the concept complexity and instance complexity does not exist for the class of union of intervals. Furthermore, observe that for $\mathcal{C}_{k,n}$ each instance (configuration of n points) is an element of a metric space, which has a measure of distance defined between any pair of instances. However, with the class of union of intervals there is no notion of a distance between instances. Finally, for the class of union of intervals, an instance is a positive example simply when the single point provided is contained within one of the k intervals. For $\mathcal{C}_{k,n}$ an instance is positive if and only if it satisfies the following two conditions:

1. Each of the n points in the instance are contained within one of the k, width 2 intervals defined by

the k target points.

2. There is at least one of the n points in the instance contained within the width 2 interval defined by each of the k target points.

Thus, these two classes are very different in character.

3 Motivation: The Landmark Recognition Problem

In this section we explore one motivation for this work. Consider a robot designed to navigate through a large-scaled environment[3]. Suppose that we have selected a set of key "landmarks" of which the robot has prior knowledge. It is crucial that the robot be able to recognize whether or not it is in the vicinity of a given landmark from a visual image taken from the robot's current location. We shall refer to this problem as the *landmark matching problem*. In his doctoral thesis, Pinette [Pin93] says that "any general navigation algorithm must be able to match landmarks by their appearance." Namely, when performing navigation a robot plans a path by moving between known landmarks, tracking landmarks as it goes. Because of inaccuracies in effectors and errors in the robot's internal map, when the robot believes it has reached landmark L, before heading to the next landmark it can check that it is really in the vicinity of L. Then adjustments can be made if the robot is not at L by either re-homing to L and/or updating its map.

It is also crucial that the landmark matching algorithm can be performed in real-time. To reduce the processing time required by the landmark matching algorithm, some are proposing the use of imaging systems that generate a one-dimensional array of light intensities taken at eye-level [HTP+92, LL90, Pin93, SA88]. We now briefly describe one such imaging system (see Hong et al. [HTP+92] and Pinnette [Pin93]). In their robot a spherical mirror is mounted above an upward-pointing camera on a robot thus enabling it to instantaneously obtain a 360 degree view of the world. See Figure 2 for a picture of such a robot. The view of the world obtained by this imaging system and the processing performed are shown in Figure 3. All points along the eye-level-view of the robot (shown by the horizon line in Figure 2) project into a circle in the robot's 360 degree view. Figure 3 shows the panoramic view that results by scanning the 360 degree view (beginning at due north) in a circle around the robot's horizon line. The panoramic view is sampled along the horizontal line midway between the top and the bottom to produce a one-dimensional array of light intensities (or *signature*) as shown in Figure 3.

Most work on designing landmark matching algorithms uses a pattern matching approach by trying to match the current signature to the signature taken at landmark position L. If one's goal is to determine if the robot is standing exactly at position L, then the pattern matching approach can easily be implemented to work well. However, in reality, the matching algorithm must determine if the robot is in the vicinity of L (i.e. in a circle centered around L). Because the visual image may change significantly as small movements around L are made, the pattern matching approach encounters difficulties.

Rather than using a pattern matching approach to match the light intensity array from the current location with the light intensity array of the landmark, we instead propose using a learning algorithm to construct a good hypothesis for performing landmark recognition. We obtain the instances by converting the array of light intensities into one-dimensional geometric patterns by placing points where there are significant changes in light intensity. The target pattern could be constructed as follows: whenever there is an object at eye-level that would cause the light intensity received by the robot to change, a set of points are placed evenly spaced at distance two from each other along the image of the object. Thus if there is an object in view from the location of the landmark, then even though a relatively small number of points are placed in the "target pattern" the "example pattern" may have significantly many more points placed in this region. It is from this occurrence that we motivate looking at the situation in which the example complexity may be significantly larger than the target complexity (and thus leads to a notion that has similarities to the notion of irrelevant attributes in the Boolean domain). Then by applying our algorithm, giving it a set of positive examples (i.e. patterns obtained from locations in the vicinity of the landmark) and a set of negative examples (i.e. patterns obtained from locations not in the vicinity of the landmark), we can construct a hypothesis that can accurately predict whether or not the robot is near the given landmark.

4 The Learning Model and Model of Noise

We assume the reader is familiar with the PAC learning model as define by Valiant [Val84]. As commonly done, we allow the learner to output any polynomially evaluatable hypothesis. We now describe the hypothesis class used here. We define the hypothesis class $\mathcal{H}_\ell$ to be the intersection of at most $2(k+1)\lg m_1$ hypotheses from $\mathcal{C}_{\ell,n}$ where m_1 is the size of the sample required in the noise-free setting. Our algorithm for PAC-learning $\mathcal{C}_{k,n}$ from noise-free data uses $\mathcal{H}_{k+1}$ as the hypothesis class and the algorithm for PAC-learning $\mathcal{C}_{k,n}$ with noisy data uses $\mathcal{H}_{2k+1}$ as the hypothesis class.

In this work we consider a variant of the PAC model in which the negative examples from the oracle, EX, are corrupted by random misclassification noise. (This noise model is just a variation of the model of random misclassification noise introduced by Angluin and Laird [AL88] except here only negative examples are

[3] By a large-scaled environment we mean that not all landmarks are visible from all locations in the environment.

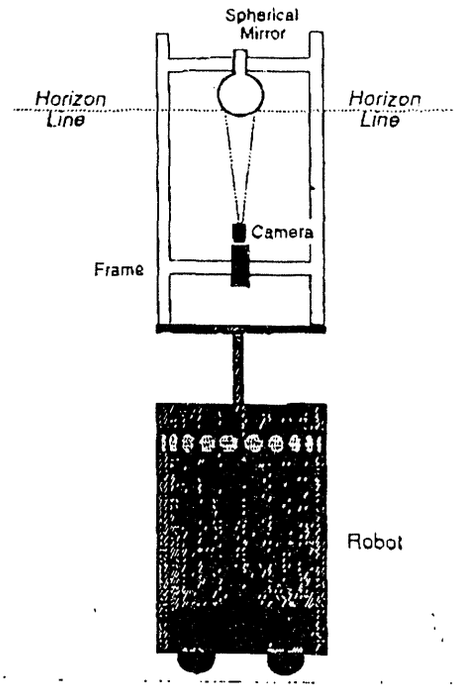


Figure 2: The imaging system on the robot. (This figure comes directly from Pinnette's thesis [Pin93].)

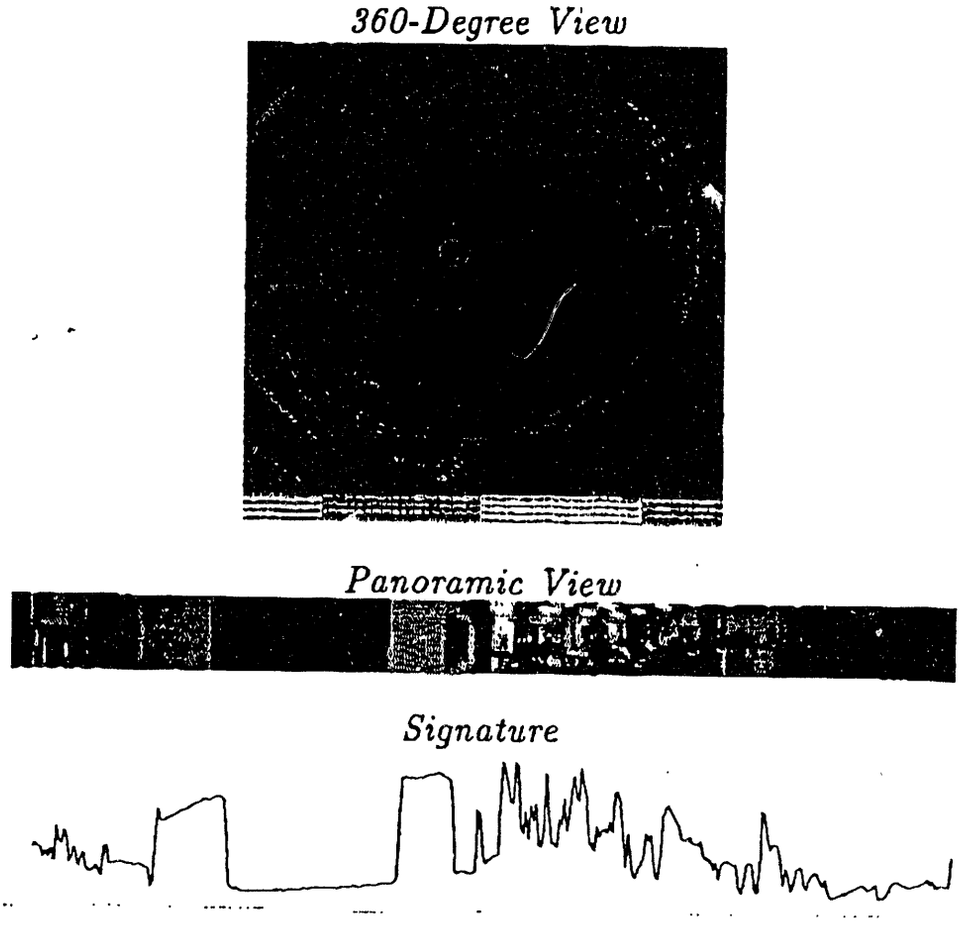


Figure 3: Stages of image processing. (This figure comes directly from Pinnette's thesis [Pin93].)

corrupted by the noise process.) Namely, we assume that for some noise rate ν every negative example drawn from EX is randomly and independently labeled as positive with probability ν and labeled as negative with probability $(1-\nu)$. We shall use EX_ν to denote the oracle after the noise process, as described above, has been applied. If ν is the noise rate and p_- (respectively, p_+) is the probability that a randomly drawn uncorrupted example is negative (respectively, positive), then our algorithm can learn as long as $\nu < 1$ (i.e. the noise rate is strictly less than one) and $\nu p_- < p_+$ (i.e. the expected number of truly positive examples is greater than the expected number of false positive examples). Our algorithm's time and sample complexity are polynomial in $\frac{1}{(1-\nu)}$ and $\left(\frac{p_+ + \nu p_-}{p_+ - \nu p_-}\right)$ where $\frac{1}{1-\nu}$ is the inverse of the amount by which the noise rate is less than 1, and $\left(\frac{p_+ + \nu p_-}{p_+ - \nu p_-}\right)$ is the inverse of the amount by which the ratio of true positive examples to total positive examples is greater than 1/2.

For ease of exposition, we assume that ν and νp_- are known, however, our results can be easily modified to work as long as an upperbound on both quantities is provided.

5 Learning $C_{k,n}$ in the Noise-Free Setting

The problem of learning one-dimensional geometric patterns has been previously studied by Goldberg [Gol92, Gol93]. He has developed an algorithm to PAC-learn $C_{n,n}$ in the noise-free setting [Gol93]. Our algorithm to learn $C_{k,n}$ is obtained by making straightforward modifications to Goldberg's algorithm. However, the modifications needed to handle the false positive errors are significantly more involved. We also note that Goldberg [Gol92] has shown that it is NP-complete to find a sphere in the given metric space (i.e. one-dimensional patterns of points on the line under the Hausdorff metric) consistent with a given set of positive and negative examples of an unknown sphere in the given metric space. In other words, given a set S of examples labeled according to some one-dimensional geometric pattern of k points it is NP-complete to find some one-dimensional geometric pattern (of *any* number of points) that correctly classifies all examples in S. Thus, assuming $NP \neq RP$, it is necessary to use a more expressive hypothesis space. To give even further evidence that the class of one-dimensional patterns is significantly more complex than the union of intervals on the real line, observe that the consistency problem for that class is trivial to solve.

Finally, the results of Goldberg and Jerrum [GJ93] can be used to show that the Vapnik-Chervonenkis dimension of $C_{k,n} \leq 2k\log(8enk) = O(k \lg n)$. We observe that as either k or n increases, and the other is held fixed, then the VC dimension can increase without limit. Hence both parameters are needed as upper bounds on concept and instance complexities.

We now present our algorithm for learning $C_{k,n}$ in the noise-free setting. Our algorithm is an Occam algorithm (see [BEHW89]). Namely, it draws a sufficiently large sample of size m_1 (polynomial in $k, \lg n, 1/\epsilon$, and $\lg 1/\delta$) and then outputs a consistent hypothesis from $\mathcal{H}_{k+1}$.

To build the hypothesis we use a greedy set cover algorithm that is based on the observation that it is possible, in polynomial time, to find a concept from $C_{k+1,n}$ consistent with all the positive examples and a fraction $\varphi = \frac{1}{2(k+1)}$ of the negative examples. Then the negative examples accounted for are removed and the procedure is repeatedly applied until all negative examples have been eliminated. Let r denote the number of rounds until all negative examples have be covered. Then it is easily shown that $r \leq 2(k+1)\lg m_1$. Finally, the hypothesis output is the intersection of the r concepts obtained in this manner.

By the results of Blumer, et al. [BEHW89] we get that the VC-dimension of $\mathcal{H}_{k+1}$ is at most $2dr\lg(3r)$ where $d = 2(k+1)\log 16en(k+1)$ and $r = 2(k+1)\lg m_1$, and thus any hypothesis that is consistent with a sample of size $m_1 = O\left(\frac{1}{\epsilon}\lg\frac{1}{\delta} + \frac{k^3 d}{\epsilon}\lg^3\left(\frac{kd}{\epsilon}\right)\right)$ will have error at most ϵ with probability at least $1-\delta$.

We now summarize how the concept H from $C_{k+1,n}$ is selected in a given round. Recall that there are two ways for an example to be negative: either there is a point in the example that is not near[4] any target point (e.g. X_2 in Figure 1), or no points in the example are near some target point (e.g. X_3 in Figure 1). Let $\mathcal{N}$ be the set of negative examples that remain at the start of a round. Then either

> **Case 1:** At least $|\mathcal{N}|/2$ of the negative examples have no points near some target point. Thus, by an averaging argument, there is some width 2 interval I_1 containing at least one point from each of the positive examples that does not contain points in at least $\frac{|\mathcal{N}|}{2k}$ of the negative examples.
>
> **Case 2:** At least $|\mathcal{N}|/2$ of the negative examples have a point that is not near a target point. Since the portions of the real line that are not near any target point form at most $k+1$ contiguous intervals, by an averaging argument, there is some interval I_2 containing points from at least $\frac{|\mathcal{N}|}{2(k+1)}$ distinct negative examples and no points from the positive examples.

Using brute force (scanning the points in the sample from left to right) we can search for these two conditions, and are guaranteed to successfully find one. In the first case, we place a point in the hypothesis in the middle of interval I_1 and then cover all points from the positive

[4] For ease of exposition, we say that an example point within unit distance from a given target point is *near* that target point.

examples in a greedy fashion. In the second case, we build a hypothesis that covers all the points from the positive examples in a greedy manner, but that has no point in the hypothesis that is within unit distance of any point in interval I_2. It is easily seen that in both cases $k+1$ points placed in H suffices. Thus we obtain the following result.

Theorem 1 *The concept class $C_{k,n}$ is PAC-learnable from the hypothesis class $\mathcal{H}_{k+1}$ when the learner is given access to the noise-free oracle EX. The sample complexity of this algorithm is*

$$m_1 = O\left(\frac{1}{\epsilon}\lg\frac{1}{\delta} + \frac{k^3 d}{\epsilon}\lg^3\left(\frac{kd}{\epsilon}\right)\right)$$

where $d = 2(k+1)\log 16en(k+1)$ and $r = 2(k+1)\lg m_1$, and the time complexity is $O(r \cdot m_1)$.

6 Learning $C_{k,n}$ in the Presence of Noise

We now describe our algorithm to PAC-learn the class of one-dimensional patterns when the negative examples are corrupted by random misclassification noise of rate ν. (i.e. the oracle EX is replaced by EX_ν) As in the noise-free setting we use an Occam algorithm, in this case, outputting a hypothesis from $\mathcal{H}_{2k+1}$. Throughout this section we let m_1 denote the sample complexity for the noise-free case when the hypothesis is drawn from $\mathcal{H}_{2k+1}$, the accuracy parameter is $\epsilon/2$, and the confidence parameter is $\delta/3$. As in the previous section, it can be shown that VC-dimension of $\mathcal{H}_{2k+1}$ is at most $2dr\lg(3r)$ where here $d = 2(2k+1)\log 16en(2k+1)$ and $r = 2(k+1)\lg m_1$. From the derivations in Section 5 it is easily seen that $m_1 = O\left(\frac{1}{\epsilon}\lg\frac{1}{\delta} + \frac{k^3 d}{\epsilon}\lg^3\left(\frac{kd}{\epsilon}\right)\right)$.

We now describe our algorithm for learning $C_{k,n}$ in the presence of noise. The complete algorithm is shown in Figure 4. The learner begins by drawing a large enough sample $\mathcal{S}_{cover}$ so that with probability at least $1 - \frac{2\delta}{3}$ both of the following two conditions hold: (1) At least m_1 noise-free examples are obtained (to satisfy this condition, we need only require the minimal condition that $\nu < 1$), and (2) more than half of the positive examples in the sample are truly positive examples (to satisfy this condition we require that $\nu p_- < p_+$).

To compute the size of the sample (as a function of m_1, ν, p_-, p_+, and δ) we use Hoeffding's Inequality [Hoe63] (also referred as a form of Chernoff bounds) as stated below:

Lemma 2 (Hoeffding's Inequality) *Let $Y_1, \ldots, Y_m$ be a sequence of m independent Bernoulli trials, each succeeding with probability p. Let $S = Y_1 + \cdots + Y_m$ be the random variable describing the total number of successes. Then for $0 \le \gamma \le 1$, both of the following hold:*

$$\begin{aligned} \Pr[S > (p+\gamma)m] &\le e^{-2m\gamma^2} \\ \Pr[S < (p-\gamma)m] &\le e^{-2m\gamma^2} \end{aligned} \quad (1)$$

Also, for $0 \le \alpha \le p$,

$$\Pr[S \le \alpha m] \le e^{-2m(\alpha-p)^2} \quad (2)$$

Theorem 3 *Given a sample of size*

$$\max\left\{\frac{2m_1}{1-\nu}, \frac{2}{(1-\nu)^2}\ln\frac{6}{\delta}, 2\ln\frac{6}{\delta}\left(\frac{p_+ + \nu p_-}{p_+ - \nu p_-}\right)^2\right\}$$

$$= O\left(\frac{m_1}{1-\nu} + \frac{1}{(1-\nu)^2}\ln\frac{1}{\delta} + \left(\frac{p_+ + \nu p_-}{p_+ - \nu p_-}\right)^2\ln\frac{1}{\delta}\right)$$

obtained from the noisy oracle EX_ν, we are guaranteed with probability at least $1 - \frac{\delta}{3}$ that at least m_1 noise-free examples are in the sample and more than half of the positive examples in the sample are truly positive examples.

Proof: We individually compute the sample size needed for each condition so that the probability that the condition fails to hold is at most $\delta/6$. Thus the total probability of either condition failing is at most $\frac{\delta}{3}$ as desired.

To compute the sample size needed to ensure that Condition (1) holds we apply the bound given in Equation (2) with $m_2 = \frac{2m_1}{(1-\nu)}$, $p = (1-\nu)$, and $\gamma = \frac{(1-\nu)}{2}$ to obtain that

$$\Pr[\text{number of noise free exs} \le m_1] \le e^{-\frac{m_2}{2}(1-\nu)^2}.$$

Thus by selecting $e^{-\frac{m_2}{2}(1-\nu)^2} \le \delta/6$ we ensure that with probability at least $1 - \delta/6$ that at least m_1 noise-free examples are obtained. Thus by solving for m_2 we obtain that we must select $m_2 \ge \frac{2}{(1-\nu)^2}\ln\frac{6}{\delta}$. Thus by drawing a sample of size

$$m_2 = \max\left\{\frac{2m_1}{1-\nu}, \frac{2}{(1-\nu)^2}\ln\frac{6}{\delta}\right\}$$

we are guaranteed that Condition (1) holds with probability at least $1 - \delta/6$.

To compute the size of the sample needed to guarantee that Condition (2) holds with sufficiently high probability, we use Equation (2) with $\alpha = 1/2$ and $p = p_+/(p_+ + \nu p_-)$ (i.e. the probability that an example labeled as positive is truly positive). From this bound we get that by drawing a sample of size at least $m_2 = 2\ln\frac{6}{\delta}\left(\frac{p_+ + \nu p_-}{p_+ - \nu p_-}\right)^2$, then condition (2) holds with probability at least $1 - \delta/6$.

This completes the proof of the theorem. □

Observe that $\frac{1}{1-\nu}$ is the inverse of the amount by which the noise rate is less than 1, and $\left(\frac{p_+ - \nu p_-}{p_+ + \nu p_-}\right)$ is the inverse of the amount by which the ratio of true positive examples to total positive examples is greater than 1/2.

Next we perform a preprocessing phase that detects some of the false positive examples. For each positive

LEARN-$\mathcal{C}_{k,n}$-FROM-EX_ν

1. Let m_1 denote the sample complexity for the noise-free case when the hypothesis is drawn from $\mathcal{H}_{2k+1}$, the accuracy parameter is $\epsilon/2$ and the confidence parameter is $\delta/3$.
2. Draw a sample $\mathcal{S}_{cover}$ of size $m_2 = \max\left\{\frac{2m_1}{1-\nu}, \frac{2}{(1-\nu)^2}\ln\frac{6}{\delta}, 2\ln\frac{6}{\delta}\left(\frac{p_+ + \nu p_-}{p_+ - \nu p_-}\right)^2\right\}$.
3. For each positive example $X \in \mathcal{S}_{cover}$, let I_X be the portion of the real line within unit distance of any one of the n points in X. Let J be all points x on the real line such that x is contained within at least half of the I_X where X ranges over all positive examples in $\mathcal{S}_{cover}$.
4. Remove from $\mathcal{S}_{cover}$ any positive example in which a point of the example is not within unit distance of a point in the set J.
5. Let $\mathcal{P}$ be the set of positive examples that remain in $\mathcal{S}_{cover}$, and let $\mathcal{N}$ be the set of negative examples in $\mathcal{S}_{cover}$.
6. Initialize H to be the always true hypothesis.
7. Repeat the following until $\mathcal{N} = \emptyset$:
   (a) Search for an interval of width 2 containing at least one point from half of the examples in $\mathcal{P}$ that does not contain points in at least $\frac{|\mathcal{N}|}{2k}$ of the examples in $\mathcal{N}$.
   (b) If such an interval I_1 is found in Step 7a Then
      i. Compute $H_1 \in \mathcal{H}_{2k+1}$ by placing a point in the center of I_1 and then greedily covering the points in $\mathcal{P}$.
      ii. $H \leftarrow H \bigcap H_1$.
      iii. $\mathcal{N} \leftarrow \mathcal{N} - \{N \in \mathcal{N} \mid H_1(N) = 0\}$.
   (c) Else (Case 2 applies)
      i. Draw a sample $\mathcal{S}_{estimate}$ of size $m_3 = \frac{8r^2(1-\nu)^2}{\epsilon^2}\left(\ln\frac{6}{\delta} + \ln r\right)$.
      ii. Let $\mathcal{I}$ be the set of minimum-sized intervals that contain at least one point from $\frac{|\mathcal{N}|}{2(k+1)}$ of the examples in $\mathcal{N}$.
      iii. For each $I \in \mathcal{I}$, compute $\hat{\lambda}(I)$ from $\mathcal{S}_{estimate}$.
      iv. Let I_{min} be the interval $I \in \mathcal{I}$ which minimizes $\hat{\lambda}(I)$
      v. Compute $H_2 \in \mathcal{H}_{2k+1}$ by greedily covering all points in $\mathcal{P}$ under the restriction that no point be placed within unit distance of I_{min}.
      vi. $H \leftarrow H \bigcap H_2$.
      vii. $\mathcal{N} \leftarrow \mathcal{N} - \{N \in \mathcal{N} \mid H_2(N) = 0\}$.
8. Output H.

Figure 4: Algorithm to PAC-learn $\mathcal{C}_{k,n}$ from the hypothesis class $\mathcal{H}_{2k+1}$ when the learner is given access to the noisy oracle EX_ν.

example $X \in \mathcal{S}_{cover}$, let I_X be the portion of the real line within unit distance of any one of the n points in X. Recall that in any true positive example, each point in the example is near some target point and there is a point in the example near each target point. Therefore, if we could intersect the I_X for all *true* positive examples X in $\mathcal{S}_{cover}$, then the k target points would be contained in the intersection. Thus, by Condition (2) on the sample, the candidate intervals for the target points can be reduced to those portions of the real line that are contained within at least half of I_X where X ranges over all examples reported as positive. Finally, any positive example with a point not within unit distance from a candidate interval for the target points is a false positive example and is discarded. This preprocessing is important to ensure that in each round we will be able to cover all points of the positive examples with $2k+1$ points in candidate intervals for the target.

Lemma 4 *With probability at least $1-\frac{\delta}{3}$, after the preprocessing phase is completed, any false positive examples that remain in $\mathcal{S}_{cover}$ are within 2 units from a target point.*

Proof: By Theorem 3, with probability at least $1-\frac{\delta}{3}$, greater than half of the positive example are truly positive. For a truly positive example X, observe that all points in I_X are within 2 units from a target point. Thus it follows that all false positive examples with a point greater than 2 units from a target point will be removed by the preprocessing. □

We now describe the modifications that we make in the portion of the noise-free algorithm in which a fraction $\frac{1}{2(k+1)}$ of the remaining negative examples are eliminated.

Case 1. Recall that in the noise-free setting if at least half of the negative examples in $\mathcal{N}$ have no points near some target point, then there is an interval of width 2 containing at least one point from each positive example that does not contain points in at least $\frac{|\mathcal{N}|}{2k}$ of the negative examples where $\mathcal{N}$ is the set of negative examples that remain.

We now show that this case is easily modified to handle the false positives errors that occur in the sample. Since more than half of the positive examples are real, it follows that in an interval near a target point, more than half of the positive examples are represented. Furthermore, since only false positive examples are located in an interval not near a target point, less than half of the positive examples are represented there. Thus, it suffices to find an interval I_1 of width two containing at least one point from half of the positive examples that does not contain points in at least $\frac{|\mathcal{N}|}{2k}$ of the negative examples. Finally, we compute $H_1 \in \mathcal{H}_{2k+1}$ that is to be added to the hypothesis by placing a point in the middle of interval I_1 and then covering the rest of the points in the positive examples in a greedy fashion.

Case 2. Recall that in the noise-free setting if at least half of the negative examples in the sample have a point that is not within unit distance of a target point then there is some interval containing points from at least $\frac{|\mathcal{N}|}{2(k+1)}$ distinct negative examples and no points from the positive examples where $\mathcal{N}$ is the set of negative examples that remain.

While in the noise-free setting in each positive example there are *no* points in intervals not within unit distance of any target point, with false positive examples we must find intervals having the desired number of negative examples represented and "few enough" points from positive examples. We then ignore these points, potentially introducing error if they were near a target point. The main complication comes from the observation that in an interval within unit distance from a target point, there could be a very high concentration of points from negative examples. Thus simply finding a group of $\frac{|\mathcal{N}|}{2(k+1)}$ negative examples with the fewest number of positive examples interleaved (thus treating the positive examples as false positives) could cause significant error in the final hypothesis.

For any interval I of the real line, let $p_+(I)$ (respectively, $p_-(I)$) denote the probability that a point from a randomly drawn positive (respectively, negative) example from the noise-free oracle EX is in I. We use $\lambda(I)$ to denote the expected ratio of observed positive to observed negative examples in interval I, and $\hat{\lambda}(I)$ to denote the estimated value for $\lambda(I)$. Let $\mathcal{I}_{away}$ denote the set of intervals for which all points in any $I \in \mathcal{I}_{away}$ are not within unit distance of a target point. Let $\mathcal{I}_{light}$ denote the set of intervals such that a portion of each $I \in \mathcal{I}_{light}$ is within unit distance of a target point, yet for all $I \in \mathcal{I}_{light}$, $p_+(I) \leq \frac{\epsilon}{2r}$, Finally, let $\mathcal{I}_{heavy}$ denote the set of intervals such that for $I \in \mathcal{I}_{heavy}$, $p_+(I) > \frac{\epsilon}{2r}$ (and thus a portion of each such interval is within unit distance of a target point).

Observe that for any interval I, $\lambda(I) = \frac{p_+(I)+\nu p_-(I)}{(1-\nu)p_-(I)} = \frac{\nu}{1-\nu} + \left(\frac{1}{1-\nu}\right)\frac{p_+(I)}{p_-(I)}$. Thus we get the following key observations.

- For any interval $I \in \mathcal{I}_{away}$, $p_+(I) = 0$ and thus $\lambda(I) = \frac{\nu}{1-\nu}$.
- For any interval $I \in \mathcal{I}_{heavy}$ since $p_+(I) > \epsilon/(2r)$ it follows that $\lambda(I) > \frac{\nu}{1-\nu} + \frac{\epsilon}{2r(1-\nu)}$.

We now show that this separation is sufficient so that if we draw a large enough, but polynomial size, sample then with high probability, we know that in each round if Case 2 applies (which we know if Case 1 fails) then of the intervals containing a point from at least $\frac{|\mathcal{N}|}{2(k+1)}$ distinct examples from $\mathcal{N}$, the interval I with the lowest value of $\hat{\lambda}(I)$ provides a good set of negative examples from $\mathcal{N}$ to eliminate.

Lemma 5 *Assume that Case 1 does not apply and let*

$\mathcal{I}$ be the set of minimum-sized intervals that contain at least one point from $\frac{|\mathcal{N}|}{2(k+1)}$ distinct examples from $\mathcal{N}$. For each $I \in \mathcal{I}$ we compute $\hat{\lambda}(I)$ from $\mathcal{S}_{estimate}$. Let I_{min} be the interval $I \in \mathcal{I}$ which minimizes $\hat{\lambda}(I)$. Then, with probability at least $1 - \frac{\delta}{3r}$, the error introduced by not allowing a positive point within unit distance of I_{min} will be at most $\epsilon/(2r)$.

Proof Sketch: Since Case 1 does not apply, we know that Case 2 applies and thus there must be some interval I that is greater than unit distance from any target point that contains points from at least $\frac{|\mathcal{N}|}{2(k+1)}$ distinct examples from $\mathcal{N}$. Since the minimum separation between the value of λ for intervals in $\mathcal{I}_{away}$ and intervals in $\mathcal{I}_{heavy}$ is at least $\frac{\epsilon}{2r(1-\nu)}$, using Hoeffding's inequality it can be shown that when the estimates for the λs are computed from the sample $\mathcal{S}_{estimate}$ of size $\frac{8r^2(1-\nu)^2}{\epsilon^2}\left(\ln\frac{6}{\delta} + \ln r\right)$ where $r = 2(k+1)\lg m_1$, then each of the following conditions hold with probability at least $1 - \frac{\delta}{6r}$.

$$\text{For any } I \in \mathcal{I}_{away}, \quad \hat{\lambda}(I) < \frac{\nu}{1-\nu} + \frac{\epsilon}{4r(1-\nu)}. \quad (3)$$

$$\text{For any } I \in \mathcal{I}_{heavy}, \quad \hat{\lambda}(I) > \frac{\nu}{1-\nu} + \frac{\epsilon}{4r(1-\nu)}. \quad (4)$$

Let $s = \frac{\epsilon}{2r(1-\nu)}$ denote the minimum separation between the value of λ for intervals in $\mathcal{I}_{away}$ and intervals in $\mathcal{I}_{heavy}$. We use sample $\mathcal{S}_{estimate}$ to compute $\hat{\lambda}(I)$ for interval I. We now apply Hoeffding's inequality as given in Equation (1) with $\gamma = s/2$ to get that for any $I \in \mathcal{I}_{away}$,

$$\Pr[\hat{\lambda}(I) \geq \frac{\nu}{(1-\nu)} + \frac{s}{2}] \leq e^{-2m_3(s/2)^2}$$

and for any $I \in \mathcal{I}_{heavy}$,

$$\Pr[\lambda(I) \leq \frac{\nu}{(1-\nu)} + \frac{s}{2}] \leq e^{-2m_3(s/2)^2}.$$

Then to ensure the probability (for each condition above) that the condition does not hold is at most $\frac{\delta}{6r}$ we require that $e^{-2m_3(s/2)^2} \leq \frac{\delta}{6r}$ in both cases.

Solving for m_3 yields that a sample of size

$$\frac{2}{s^2}\left(\ln\frac{6}{\delta} + \ln r\right) = \frac{8r^2(1-\nu)^2}{\epsilon^2}\left(\ln\frac{6}{\delta} + \ln r\right)$$

suffices. Solving for m_3 yields that the given sample suffices.

We now complete the proof of the lemma. Since $I \in \mathcal{I}_{away}$ by Equation (3) we have that with probability at least $1 - \frac{\delta}{6r}$,

$$\hat{\lambda}(I) < \frac{\nu}{1-\nu} + \frac{\epsilon}{4r(1-\nu)}.$$

Furthermore, $I \in \mathcal{I}$, and thus it will be included in the minimum. By Equation (4) the probability that the estimate for any interval in $\mathcal{I}_{heavy}$ is selected as the minimum is at most $\delta/(6r)$ and thus, with probability at least $1 - \frac{\delta}{3r}$, $I_{min} \in \mathcal{I}_{away} \bigcup \mathcal{I}_{light}$. The concept from $\mathcal{H}_{2k+1}$ placed in the hypothesis prevents an example with a point in I_{min} to be classified as positive. Finally by the definitions of $\mathcal{I}_{away}$ and $\mathcal{I}_{light}$ this introduces error at most $\frac{\epsilon}{2r}$ giving the desired result. □

Putting this all together we get our main result.

Theorem 6 *There is an algorithm to PAC-learn the concept class $C_{k,n}$ from the hypothesis class $\mathcal{H}_{2k+1}$ when the learner is given access to the noisy oracle EX_ν that has time complexity and sample complexity polynomial in k, $\lg n$, $1/\epsilon$, $\lg 1/\delta$, $\frac{1}{(1-\nu)}$, and $\left(\frac{p_+ + \nu p_-}{p_+ - \nu p_-}\right)$.*

Proof Sketch: Our algorithm is shown in Figure 4. By the choice of m_1, if the algorithm's final hypothesis were consistent with m_1 properly labeled examples then it would have error at most $\epsilon/2$ with probability at least $1 - \frac{\delta}{3}$. Furthermore, from Theorem 3 the probability that either Condition (1) or Condition (2) on $\mathcal{S}_{cover}$ does *not* hold is at most $\delta/3$. Thus if the final hypothesis were consistent with all examples from $\mathcal{S}_{cover}$ then it would have error at most $\epsilon/2$ with probability at least $1 - \frac{2\delta}{3}$.

Given that Condition (2) holds then if Case 1 applies (i.e. at least half of the negative examples in $\mathcal{N}$ have no points within distance two of any target point), then there is an interval of width two containing at least one point from each from each truly positive example in $\mathcal{P}$ that does not contain points in at least $\frac{|\mathcal{N}|}{2k}$ of the negative examples where $\mathcal{N}$ is the set of negative examples that have not yet been eliminated. Since more than half of the points in $\mathcal{P}$ are truly positive if Case 1 applies then there exists an interval I_1 of width two containing at least one point from half of the positive examples that does not contain points in at least $\frac{|\mathcal{N}|}{2k}$ of the negative examples. Thus this interval will be found. Finally, by placing a positive point in the middle of I_1 all examples in $\mathcal{N}$ with a point in I_1 will be classified as negative by H_1 and the algorithm will return to the top of the loop.

Now consider the case in which Case 1 does not apply, and thus Case 2 applies (i.e. at least half of the examples in $\mathcal{N}$ have a point that is not near the target point). From Lemma 5 it follows that at each round the probability that the error introduced in H is greater than $\frac{\epsilon}{2r}$ is at most $\frac{\delta}{3r}$. Thus, given that Conditions (1) and (2) hold for $\mathcal{S}_{cover}$, the probability that total additional error incurred in Case 2 is greater than $\epsilon/2$ is at most $\frac{\delta}{3}$. Finally, we show that all positive points can be properly classified by a hypothesis in $\mathcal{H}_{2k+1}$ for which there is no point within unit distance of I_{min}. From Lemma 4 we know that all positive examples in $\mathcal{P}$ (truly positive and the false positives not eliminated by the preprocessing) are within distance 2 from a target point. Thus by greedily covering the positive examples we know that at most $2k$ additional points will be needed.

Combining the above with the guarantees given that Conditions (1) and (2) on $\mathcal{S}_{cover}$ hold, we get that the error of the final hypothesis output by our algorithm is at most ϵ with probability at least $1-\delta$. The sample complexity of our algorithm is $O(|\mathcal{S}_{cover}|+r\cdot|\mathcal{S}_{estimate}|)$ and the time complexity is $O(r\cdot|\mathcal{S}_{cover}|+r\cdot|\mathcal{S}_{estimate}|)$.

□

7 Concluding Remarks

We are currently beginning to implement and test our algorithm on data from a robot with an imaging system as shown in Figure 2. Such experimental work will enable us to see how this approach to solve the landmark matching problem compares to a pattern matching approach. Also this experimental work may suggest modifications in the theoretical model of noise that we have studied so that it better models the type of noise found in real data.

We are also looking at techniques to reduce the time and sample complexity of the algorithm presented here, and studying the situation in which both the positive and negative examples are corrupted by random misclassification noise. Finally, it would be interesting to consider extensions of this work when the points in the target and example configurations are drawn from the plane.

Acknowledgements

We thank Brian Pinette for allowing us to include his figures in our paper. We also thank Stephen Judd and Tom Hancock for several very useful discussion about the material in Section 3. Finally, we thank the COLT committee members for their comments.

References

[AL88] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, October 1989.

[Gol92] Paul W. Goldberg. *PAC Learning Geometrical Figures*. PhD thesis, University of Edinburgh, 1992.

[Gol93] Paul W. Goldberg. Geometrical pattern learning. Unpublished Manuscript, April 1993.

[GJ93] Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of Concept Classes Parameterized by Real Numbers. *Conference on Computational Learning Theory*, July 1993.

[Gru83] P.M. Gruber. Approximation of convex bodies. In P.M. Gruber and P.M. Willis, editors, *Convexity and its applications*. Brikhauser Verlag, 1983.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[HTP+92] Jiawei Hong, Xiaonan Tan, Brian Pinette, Richard Weiss, and Edward M. Riseman. Image-based homing. *IEEE Control Systems Magazine*, 12(1):38–45, 1992.

[LL90] Todd S. Levitt and Daryl T. Lawton. Qualitative navigation for mobile robots. *Artificial Intelligence*, 44(3): 305–360, 1990.

[Lit88] Nick Littlestone. Learning when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[Pin93] Brian Pinette. *Image-Based Navigation Through Large-Scaled Environments*. PhD thesis, University of Massachusetts, Amherst, November 1993.

[SA88] Hisashi Suzuki and Suguru Arimoto. Visual control of autonomous mobile robot based on self-organizing model for pattern learning. *Journal of Robotic Systems*, 5(5):453–470, 1988.

[Val84] Leslie Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

[Val91] Leslie Valiant. A view of computational learning theory. In C.W. Gear, editor, *NEC Research Symposium: Computation and Cognition*. SIAM, Philadelphia, 1991.

Doc I

DATE FILMED

8/18/94

END