

# A New PAC Bound for Intersection-Closed Concept Classes <sup>★</sup>

Peter Auer, Ronald Ortner

Department Mathematik und Informationstechnologie  
Montanuniversität Leoben, Franz-Josef-Straße 18, 8700-Leoben, Austria  
e-mail: {auer,rortner}@unileoben.ac.at

The date of receipt and acceptance will be inserted by the editor

**Abstract** For hyper-rectangles in  $\mathbb{R}^d$  Auer et al. [1] proved a PAC bound of  $O\left(\frac{1}{\varepsilon}(d + \log \frac{1}{\delta})\right)$ , where  $\varepsilon$  and  $\delta$  are the accuracy and confidence parameters. It is still an open question whether one can obtain the same bound for intersection-closed concept classes of VC-dimension  $d$  in general. We present a step towards a solution of this problem showing on one hand a new PAC bound of  $O\left(\frac{1}{\varepsilon}(d \log d + \log \frac{1}{\delta})\right)$  for arbitrary intersection-closed concept classes, complementing the well-known bounds  $O\left(\frac{1}{\varepsilon}(\log \frac{1}{\delta} + d \log \frac{1}{\varepsilon})\right)$  and  $O\left(\frac{d}{\varepsilon} \log \frac{1}{\delta}\right)$  of Blumer et al. [4] and Haussler et al. [7]. Our bound is established using the *closure algorithm*, that generates as its hypothesis the intersection of all concepts that are consistent with the positive training examples. On the other hand, we show that many intersection-closed concept classes including e.g. maximum intersection-closed classes satisfy an additional combinatorial property that allows a proof of the optimal bound of  $O\left(\frac{1}{\varepsilon}(d + \log \frac{1}{\delta})\right)$ . For such improved bounds the choice of the learning algorithm is crucial as there are consistent learning algorithms that need  $\Omega\left(\frac{1}{\varepsilon}(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta})\right)$  examples to learn some particular maximum intersection-closed concept classes.

## 1 Introduction

In the PAC model a learning algorithm generalizes from given examples to a hypothesis that approximates a target concept taken from a concept class known to the learner. The learning algorithm  $\mathcal{A}$  then *PAC learns* a concept class, if for  $\varepsilon, \delta > 0$  there is an  $m = m(\varepsilon, \delta)$ , such that with probability at

---

<sup>★</sup> A preliminary version of this paper appeared in: Learning Theory, Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004. Lecture Notes in Computer Science 3120, pp. 408–414.

least  $1 - \delta$  the algorithm outputs a hypothesis with error smaller than  $\varepsilon$ , when  $m$  random examples are given to  $\mathcal{A}$ . Bounds on  $m$  often depend on the *VC-dimension*, a combinatorial parameter of the concept class. For finite  $d$  the well-known bound of Blumer et al. [4] states that for *any* consistent learning algorithm  $O\left(\frac{1}{\varepsilon}(\log \frac{1}{\delta} + d \log \frac{1}{\varepsilon})\right)$  examples suffice for PAC learning concept classes of VC-dimension  $d$ . On the other hand, for the 1-inclusion graph algorithm a bound of  $O\left(\frac{d}{\varepsilon} \log \frac{1}{\delta}\right)$  was established in [7].

In this paper we give a complementing bound of  $O\left(\frac{1}{\varepsilon}(d \log d + \log \frac{1}{\delta})\right)$  when learning *intersection-closed* concept classes (cf. e.g. [1, 2, 8]) with the *closure algorithm*. Intersection-closed concept classes include quite natural classes such as hyper-rectangles in  $\mathbb{R}^d$  or the class of all subsets of size at most  $d$  of some finite set  $X$ . These specific intersection-closed concept classes as well as many others satisfy an additional combinatorial property that allows to prove an optimal bound of  $O\left(\frac{1}{\varepsilon}(d + \log \frac{1}{\delta})\right)$  in these particular cases (see [3] and Section 4 below, respectively). It is an open problem whether this optimal bound holds for intersection-closed concept classes in general. If so, it can be achieved only for special learning algorithms, since there are consistent learning algorithms that need  $\Omega\left(\frac{1}{\varepsilon}(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta})\right)$  examples to learn some intersection-closed concept classes (see Section 4 below).

## 2 Preliminaries

### 2.1 Intersection-Closed Concept Classes

A *concept class* over a (possibly infinite) set  $X$  is a subset  $\mathcal{C} \subseteq 2^X$ . For  $Y \subseteq X$  we set  $\mathcal{C} \cap Y := \{C \cap Y \mid C \in \mathcal{C}\}$ . The *VC-dimension* of a concept class  $\mathcal{C} \subseteq 2^X$  is the cardinality of a largest  $Y \subseteq X$  for which  $\mathcal{C} \cap Y = 2^Y$ .

**Definition 1** A concept class  $\mathcal{C} \subseteq 2^X$  is *intersection-closed* if for all  $C_1, C_2 \in \mathcal{C}$  the intersection  $C_1 \cap C_2$  is in  $\mathcal{C}$  as well.

For any set  $Y \subseteq X$  and any concept class  $\mathcal{C} \subseteq 2^X$  we define the *closure of  $Y$*  (with respect to  $\mathcal{C}$ ) as the intersection of all concepts in  $\mathcal{C}$  that contain  $Y$ , i.e.  $\text{clos}_{\mathcal{C}}(Y) := \bigcap_{Y \subseteq C \in \mathcal{C}} C$ . If it is clear to which concept class we refer, we often drop the index and write  $\text{clos}(Y)$ . Note that if there is no concept containing  $Y$ , then the closure is by definition of the nullary intersection the set  $X$  itself, so that  $Y \subseteq \text{clos}(Y)$  holds in general. The following proposition provides an alternative definition of intersection-closed concept classes for finite  $X$ .

**Proposition 1** A concept class  $\mathcal{C} \subseteq 2^X$  over finite  $X$  is *intersection-closed* if and only if for  $Y \subseteq C \in \mathcal{C}$  one always has  $\text{clos}(Y) \in \mathcal{C}$ .

*Proof* First, it is clear by definition that  $\text{clos}(Y) \in \mathcal{C}$  for intersection-closed  $\mathcal{C}$ . Now suppose that for  $Y \subseteq C \in \mathcal{C}$  one always has  $\text{clos}(Y) \in \mathcal{C}$ , and let  $C_1, C_2 \in \mathcal{C}$ . Then because of  $C_1 \cap C_2 \subseteq C_1, C_2$  we have by definition of the

closure,  $\text{clos}(C_1 \cap C_2) \subseteq C_1, C_2$  and consequently  $\text{clos}(C_1 \cap C_2) \subseteq C_1 \cap C_2$ . On the other hand,  $C_1 \cap C_2 \subseteq \text{clos}(C_1 \cap C_2)$ , so that  $C_1 \cap C_2 = \text{clos}(C_1 \cap C_2) \in \mathcal{C}$ .  $\square$

Let  $\bar{\mathcal{C}} = \{\bigcap_{C \in \mathcal{C}'} C : \mathcal{C}' \subseteq \mathcal{C}\}$  be the concept class of *all* intersections of concepts in  $\mathcal{C}$ . Obviously,  $\text{clos}_{\mathcal{C}}(Y) \in \bar{\mathcal{C}}$  for any intersection-closed class  $\mathcal{C}$  (if  $Y$  is contained in some concept), and  $\text{clos}_{\mathcal{C}}(Y) = \text{clos}_{\bar{\mathcal{C}}}(Y)$ . The following proposition shows that we can move from  $\mathcal{C}$  to  $\bar{\mathcal{C}}$  without difficulty.

**Proposition 2** *If  $\mathcal{C}$  is intersection-closed and of VC-dimension  $VC(\mathcal{C}) = d$ , then  $VC(\bar{\mathcal{C}}) = d$ .*

*Proof* Obviously  $VC(\bar{\mathcal{C}}) \geq VC(\mathcal{C})$ . Assume that  $Y \subseteq X$  with  $|Y| = d + 1$  is shattered by  $\bar{\mathcal{C}}$ . Then  $Y$  is also shattered by  $\bar{\mathcal{C}} \cap Y = \mathcal{C} \cap Y$ , since  $\bar{\mathcal{C}} \cap Y$  is finite and  $\mathcal{C}$  is intersection-closed. This contradicts  $VC(\mathcal{C}) = d$ .  $\square$

Again, let  $Y \subseteq X$ . A *spanning set* of  $Y$  (with respect to an intersection-closed concept class  $\mathcal{C} \subseteq 2^X$ ) is any set  $S \subseteq Y$  such that  $\text{clos}_{\mathcal{C}}(S) = \text{clos}_{\mathcal{C}}(Y)$ . A spanning set  $S$  of  $Y$  is called *minimal* if no subset of  $S$  is a spanning set of  $Y$ . Finally, let  $\text{span}_{\mathcal{C}}(Y)$  denote the set of all minimal spanning sets of  $Y$ . Again we will often drop the index, if no ambiguity can arise. Note that if  $Y$  is finite, then  $\text{span}(Y) \neq \emptyset$ . The following theorem mentions a key property of intersection-closed concept classes.

**Theorem 1** *Let  $\mathcal{C} \subseteq 2^X$  be an intersection-closed class of VC-dimension  $d$ . Let  $Y \subseteq X$  be finite and contained in some concept of  $\mathcal{C}$ . Then all minimal spans of  $Y$  have size at most  $d$ .*

*Proof* [8] First, we consider the case where  $X$  is finite. Let  $Y \subseteq C_Y$  for some  $C_Y \in \mathcal{C}$ . We show that any minimal spanning set of  $Y$  is shattered by  $\mathcal{C}$ . Thus let  $S \in \text{span}(Y)$ . Since  $\mathcal{C}$  is intersection-closed, it is sufficient to show that (i)  $S \subseteq C$  for some  $C \in \mathcal{C}$  and (ii) for each  $x \in S$  there is a  $C \in \mathcal{C}$  such that  $S \setminus \{x\} \subseteq C$  and  $x \notin C$ .

Because of  $S \subseteq Y$  we have  $S \subseteq \text{clos}(S) \subseteq \text{clos}(Y) \subseteq C_Y \in \mathcal{C}$ , so that (i) holds. To see that (ii) holds as well, note that for  $x \in S$  one has  $\text{clos}(S) \neq \text{clos}(S \setminus \{x\})$  due to the minimality of  $S$ . It follows that  $S \setminus \{x\} \subseteq \text{clos}(S \setminus \{x\})$  and  $x \notin \text{clos}(S \setminus \{x\})$ . Since by Proposition 1,  $\text{clos}(S \setminus \{x\}) \in \mathcal{C}$ , this finishes the proof for finite  $X$ .

If  $X$  is infinite, one shows as above that  $S$  is shattered by  $\mathcal{C} \cap S$  and consequently by  $\mathcal{C}$ .  $\square$

*Remark 1* Note that Theorem 1 does not hold for arbitrary  $Y$ , as the following example shows. Let  $\mathcal{C}_{X,d} := \{C \subseteq X : |C| \leq d\}$  be the class of all subsets of  $X$  of size at most  $d$ . Then for each  $Y$  with  $|Y| > d$  one has  $\text{clos}(Y) = X$ , whereas the closure of each set of size smaller than  $d$  is the set itself. Hence, each spanning set of  $Y$  must consist of more than  $d$  elements.

However, it is easy to see that a minimal spanning set of an arbitrary finite  $Y \subseteq X$  cannot have more than  $d + 1$  elements.

**Corollary 1** *Let  $\mathcal{C} \subseteq 2^X$  be an intersection-closed class of VC-dimension  $d$ . Then all minimal spans of any finite  $Y \subseteq X$  have size at most  $d + 1$ .*

*Proof* Let  $\mathcal{C}' := \mathcal{C} \cup X$ . It is easy to see that the VC-dimension of  $\mathcal{C}'$  is at most  $d + 1$ , so that the corollary follows immediately from Theorem 1.  $\square$

Furthermore, we shall need the following well-known theorem.

**Theorem 2 (Sauer's Lemma[9])** *Let  $\mathcal{C} \subseteq 2^X$  be a concept class of VC-dimension  $d$  over finite  $X$ . Then*

$$|\mathcal{C}| \leq \binom{|X|}{\leq d} = \sum_{i=0}^d \binom{|X|}{i}.$$

## 2.2 Learning

Learning a concept  $C \in \mathcal{C}$  means learning the characteristic function  $\mathbf{1}_C$  on  $X$ . Thus the learner outputs a hypothesis  $h : X \rightarrow \{0, 1\}$ . Given a probability distribution  $\mathcal{P}$  on  $X$ , the error of the hypothesis  $h$  with respect to  $C$  and  $\mathcal{P}$  is defined as  $\text{er}_{C, \mathcal{P}}(h) := \mathcal{P}(\{x \mid h(x) \neq \mathbf{1}_C(x)\})$ .

**Definition 2** *A concept class  $\mathcal{C} \subseteq 2^X$  is called PAC learnable, if for all  $\varepsilon, \delta \in (0, 1)$  there is an  $m = m(\varepsilon, \delta)$ , such that for all probability distributions  $\mathcal{P}$  on  $X$  and all  $C \in \mathcal{C}$ : when learning  $C$  from  $m$  examples, the output hypothesis  $h$  has  $\text{er}_{C, \mathcal{P}}(h) > \varepsilon$  with probability smaller than  $\delta$  in respect to the  $m$  examples drawn independently according to  $\mathcal{P}$  and labelled by  $C$ .*

## 3 A New PAC Bound

The property mentioned in Theorem 1 can be used together with Sauer's Lemma to modify the original proof of the bound of  $O\left(\frac{1}{\varepsilon}(\log \frac{1}{\delta} + d \log \frac{1}{\varepsilon})\right)$  for arbitrary concept classes by Blumer et al. [4] to obtain the following alternative bound.

**Theorem 3** *Let  $\mathcal{C} \subseteq 2^X$  be a well-behaved<sup>1</sup>, intersection-closed concept class of VC-dimension  $d \geq 10$ . Then  $\mathcal{C}$  is PAC learnable from*

$$m = \left\lceil \max \left\{ \frac{16}{\varepsilon} d \log d, \frac{6}{\varepsilon} \log \frac{7}{\delta} \right\} \right\rceil$$

*examples.*

---

<sup>1</sup> This is some modest measure-theoretic assumption on the concept class (cf. proof of Lemma 1 below).

### 3.1 The Closure Algorithm

Unlike the original theorem of [4], the result of Theorem 3 does not hold for *any* consistent learning algorithm, but for the *closure algorithm*. Given a set of labelled examples  $(x_1, y_1), \dots, (x_m, y_m)$  with labels  $y_i \in \{0, 1\}$ , the hypothesis generated by the closure algorithm is the closure of the positive examples, i.e.  $\text{clos}(\{x_i \mid y_i = 1, 1 \leq i \leq m\})$ . Thus, negative examples don't have any influence on the generated hypothesis.

*Example 1* (a) Consider the intersection-closed class of all convex sets in  $\mathbb{R}^n$ . In this case, the closure of a set is simply its convex hull. Thus, the hypothesis output by the closure algorithm is the convex hull of all positive examples, which is the smallest concept of the class that is consistent with the positive examples.

(b) Consider the class of open intervals  $(a, b)$  in  $\mathbb{R}$ . Adding the empty set one obtains an intersection-closed class. The hypothesis of the closure algorithm is the intersection of all open intervals that contain all the positive examples. Given at least two distinct positive examples this is the smallest *closed* interval that contains all positive examples.

As Example 1 (b) shows, the output hypothesis need not be a concept in  $\mathcal{C}$  but obviously in  $\bar{\mathcal{C}}$ . For a classification task with  $n$  labelled and  $m$  unlabelled examples given, one would restrict the concept class to the  $n + m$  given examples. With respect to the arising concept class  $\mathcal{C}'$ , the closure of the positively labelled examples is guaranteed to be a concept in  $\mathcal{C}'$ , which in turn can be used to classify the unlabelled examples.

**Proposition 3** *The hypothesis generated by the closure algorithm classifies all negative examples correctly.*

*Proof* The algorithm returns the intersection of all concepts that are consistent with the given positive examples. Consequently, if the output hypothesis classified any negative example incorrectly, there wouldn't be any concept in  $\mathcal{C}$  that is consistent with the given examples.  $\square$

### 3.2 Proof of Theorem 3

We start with some simple observations. Suppose we have some concept  $C \in \mathcal{C}$  and examples  $(x_1, y_1), \dots, (x_{2m}, y_{2m})$ ,  $k$  of which are misclassified by  $C$ . According to Proposition 3, these  $k$  examples must be positive. Let  $\ell$  be the number of positive examples among  $(x_1, y_1), \dots, (x_{2m}, y_{2m})$ . We define recursively sets  $X_i^+$  and  $S_i$  for  $i = 1, \dots, \ell$ , where  $X_1^+ := \{x_i \mid y_i = 1, 1 \leq i \leq 2m\}$  is the set of positive examples.  $S_i$  is an arbitrary element of  $\text{span}(X_i^+)$ , and for  $i > 1$  we set  $X_i^+ := X_{i-1}^+ \setminus S_{i-1}$ . The sets  $S_i$  can be considered as the outer “shells” of the set  $X_1^+$  that are removed step by step. The idea behind this construction is that for each  $X_i^+$  that contains

misclassified examples, there must be at least one misclassified example in the shell  $S_i$ , too: otherwise, if all examples in  $S_i$  were classified correctly, then since  $S_i \in \text{span}(X_i^+)$  one has  $X_i^+ \subseteq \text{clos}(X_i^+) = \text{clos}(S_i) \subseteq C$ , which means that the examples of  $X_i^+$  would be classified correctly as well. Thus removing shell  $S_i$  from  $X_i^+$  at least one misclassified example is removed, which leads to the following proposition.

**Proposition 4** *Let  $C$  be some concept of an intersection-closed concept class. If  $C$  classifies  $k$  of given examples  $x_1, \dots, x_{2m}$  incorrectly, they are in  $\bigcup_{i=1}^k S_i$ .*

*Proof* By Proposition 3, misclassified examples must be in  $X_1^+$ . Now suppose there is a wrongly classified example that is not in  $\bigcup_{i=1}^k S_i$ . Since the  $S_i$  are disjoint, there must be an  $S_{i_0}$  that does not contain any misclassified example. Thus, all examples in  $S_{i_0}$  and consequently all examples in  $X_{i_0}^+$  are classified correctly. But this is only possible if all the  $k$  misclassified examples have been removed before, which means they are contained in  $\bigcup_{i=1}^{i_0} S_i \subseteq \bigcup_{i=1}^k S_i$ , which contradicts our assumption.  $\square$

**Lemma 1** *Let  $\mathcal{C} \subseteq 2^X$  be a well-behaved, intersection-closed concept class of VC-dimension  $d$ ,  $\mathcal{P}$  a probability distribution on  $X$ , and  $C \in \mathcal{C}$ . Then for all  $\varepsilon > 0$  and for all  $m > \frac{2}{\varepsilon}$ , given  $m$  independent random examples labelled by  $C$  and drawn according to  $\mathcal{P}$ , the probability that the hypothesis  $h$  generated by the closure algorithm has error  $\text{er}_{C, \mathcal{P}}(h) > \varepsilon$  is at most*

$$2 \sum_{k=p}^m 2^{-k} \binom{kd}{\leq d},$$

where  $p = \lceil \varepsilon m / 2 \rceil$ .

*Proof* As mentioned before, we modify the original proof of [4], pp.952ff. The basic first step is exactly as in the original proof. One shows that the probability in the lemma can be upper bounded using the so-called “doubling trick”:<sup>2</sup> We are interested in the probability that given  $m$  labelled examples  $(x_1, y_1), \dots, (x_m, y_m)$ , the generated hypothesis  $h$  has error greater than  $\varepsilon$ . This can be interpreted as follows: Given another  $m$  examples  $(x_{m+1}, y_{m+1}), \dots, (x_{2m}, y_{2m})$  drawn according to  $\mathcal{P}$ , the probability for each single example that it is misclassified by  $h$  is at least  $\varepsilon$ . On average,  $\varepsilon m$  of these  $m$  examples get misclassified. Now, basically applying Chebyshev’s

---

<sup>2</sup> In the following we only give an outline of Lemmata A1.1, A2.1 and the first part of Lemma A2.2 in [4], pp.954ff. For the proof of these lemmata one needs the already mentioned measure-theoretic assumptions on  $\mathcal{C}$  that guarantee that the following two sets are measurable: (i) the set of all  $m$ -tuples  $((x_1, y_1), \dots, (x_m, y_m))$  such that the hypothesis calculated from these examples has error larger than  $\varepsilon$ , and (ii) the set of all  $(2m)$ -tuples  $((x_1, y_1), \dots, (x_{2m}, y_{2m}))$  such that the hypothesis calculated from the first  $m$  elements misclassifies at least  $\lceil \varepsilon m / 2 \rceil$  of the second  $m$  elements (for a detailed discussion see [4], pp.952ff).

inequality, one can show that the probability that at least  $\varepsilon m/2$  examples are misclassified exceeds  $\frac{1}{2}$ , presupposed that  $m > \frac{2}{\varepsilon}$ . It follows that

$$\frac{1}{2}P(((x_1, y_1), \dots, (x_m, y_m)) : \mathbf{er}_{C, \mathcal{P}}(h) > \varepsilon) \leq P(((x_1, y_1), \dots, (x_{2m}, y_{2m})) : \mathbf{er}_{C, s}(h) > \varepsilon/2),$$

where the hypothesis  $h$  in both cases is calculated from the examples  $(x_1, y_1), \dots, (x_m, y_m)$ , and  $\mathbf{er}_{C, s}(h)$  denotes the error of the hypothesis  $h$  on the sample  $(x_{m+1}, y_{m+1}), \dots, (x_{2m}, y_{2m})$ . Now, the latter probability equals  $\frac{\pi(2m)}{(2m)!}$ , where  $\pi(2m)$  denotes the number of those permutations of a sample of  $2m$  examples where the hypothesis calculated from the first  $m$  examples misclassifies at least  $p = \lceil \varepsilon m/2 \rceil$  of the second  $m$  examples. It follows that

$$P(((x_1, y_1), \dots, (x_m, y_m)) : \mathbf{er}_{C, \mathcal{P}}(h) > \varepsilon) \leq 2 \frac{\pi(2m)}{(2m)!}.$$

Note that due to the doubling trick, we may restrict the concept class  $\mathcal{C}$  to the  $2m$  given examples, so that we may assume that the closure algorithm returns a concept in this restricted class. This is implicitly used in the application of Proposition 4 below.

Now we deviate from the original proof by giving a new an upper bound on  $\pi(2m)$ . One has to consider the number of *witnesses*, i.e. all subsets of  $\{x_1, \dots, x_{2m}\}$  that may occur as the  $k \geq p$  misclassified examples among the  $\pi(2m)$  permutations. Since we use the closure algorithm for hypothesis calculation, by Proposition 4, it is sufficient to consider the corresponding subsets of  $\bigcup_{i=1}^k S_i$  for  $k = p, \dots, m$ . By Theorem 1,  $|\bigcup_{i=1}^k S_i| \leq kd$  so that by Sauer's Lemma the number of witnesses for fixed  $k$  is at most  $\binom{kd}{\leq d}$ . Since for a given witness of  $k$  misclassified examples at most a fraction of  $2^{-k}$  permutations of all the  $(2m)!$  permutations is among the  $\pi(2m)$  (cf. [4], pp.955), the result follows after summing up over all  $k \in \{p, \dots, m\}$ .  $\square$

**Lemma 2** *If  $d \geq 10$  and*

$$m \geq \max \left\{ \frac{16}{\varepsilon} d \log d, \frac{6}{\varepsilon} \log \frac{7}{\delta} \right\} \text{ then } 2 \sum_{k=p}^m 2^{-k} \binom{kd}{\leq d} < \delta ,$$

where  $p = \lceil \varepsilon m/2 \rceil$ .

*Proof* First, we are going to use Proposition A2.1 (iii) of [4], which tells us that for  $k, d \geq 1$  one has

$$\binom{kd}{\leq d} \leq (ek)^d . \quad (1)$$

It is easy to check that for  $d \geq 10$  and  $x := 8 d \log d$  it holds that  $x \log 2 > 2d \log(ex)$ . It follows that for  $d \geq 10$  and  $k \geq 8 d \log d$

$$\frac{k}{2} > \frac{d \log(ek)}{\log 2}, \text{ or equivalently } (ek)^d < 2^{k/2} . \quad (2)$$

Hence for  $p \geq 8d \log d$  we have from (1) and (2)

$$\begin{aligned} 2 \sum_{k=p}^m 2^{-k} \binom{kd}{\leq d} &\leq 2 \sum_{k=p}^m 2^{-k} (ek)^d < 2 \sum_{k=p}^m 2^{-k/2} < 2 \cdot 2^{-p/2} \sum_{k=0}^{\infty} 2^{-k/2} \\ &= 2 \cdot 2^{-p/2} \sum_{k=0}^{\infty} \left(\frac{1}{\sqrt{2}}\right)^k = 2 \cdot 2^{-p/2} \frac{2}{2 - \sqrt{2}}. \end{aligned}$$

Setting  $K := \frac{4}{2 - \sqrt{2}}$  and substituting  $p = \lceil \varepsilon m / 2 \rceil$ , it is easy to see that for  $m \geq \frac{6}{\varepsilon} \log \frac{7}{\delta} > \frac{4}{\varepsilon \log 2} \log \frac{K}{\delta}$  one has  $p > \frac{2}{\log 2} \log \frac{K}{\delta}$ , whence  $\log 2^{p/2} > \log \frac{K}{\delta}$  or equivalently  $K \cdot 2^{-p/2} < \delta$ , which finishes the proof.  $\square$

*Proof of Theorem 3.* The theorem follows immediately from Lemmata 1 and 2.

*Remark 2* It is easy to see that the closure algorithm corresponds to one particular orientation of the 1-inclusion graph algorithm [7]. Thus, besides our bound of  $O\left(\frac{1}{\varepsilon}(d \log d + \log \frac{1}{\delta})\right)$  and the bound  $O\left(\frac{1}{\varepsilon}(\log \frac{1}{\delta} + d \log \frac{1}{\varepsilon})\right)$  of Blumer et al. [4], the bound of  $O\left(\frac{d}{\varepsilon} \log \frac{1}{\delta}\right)$  for the 1-inclusion graph algorithm of Haussler et al. [7] holds as well for learning intersection-closed classes with the closure algorithm.

#### 4 An Optimal PAC Bound for Intersection-Closed Classes with Homogeneous Spans

In this section we consider intersection-closed classes whose concepts have spanning sets with some additional combinatorial structure. For these classes one can obtain a bound of  $O\left(\frac{1}{\varepsilon}(d + \log \frac{1}{\delta})\right)$ , which meets the lower bound on the number of examples any algorithm needs to learn some particular concept classes of VC-dimension  $d$  (cf. [5]).

**Definition 3** *An intersection-closed concept class  $\mathcal{C} \subseteq 2^X$  of VC-dimension  $d$  is said to have homogeneous spans  $S$  if one can assign to each finite  $Y \subseteq C \in \mathcal{C}$  a (not necessarily minimal) spanning set  $S(Y)$  of size at most  $d$ , such that for all  $Y \subseteq X$  and all  $x \in S(Y)$ :*<sup>3</sup>

$$S(Y) \setminus x \subseteq S(Y \setminus x).$$

The following proposition is a straightforward consequence of Definition 3.

**Proposition 5** *Let  $\mathcal{C} \subseteq 2^X$  be an intersection-closed concept class  $\mathcal{C} \subseteq 2^X$  with homogeneous spans. Then for all finite  $Y \subseteq X$  and all  $Z \subseteq S(Y)$ :*

$$S(Y) \setminus Z \subseteq S(Y \setminus Z).$$

---

<sup>3</sup> For the sake of readability in the following we will often skip the brackets and write simply  $x$  for singletons  $\{x\}$ .

All interesting intersection-closed concept classes we are aware of have homogeneous spans. Some of them are mentioned in the following proposition.

**Proposition 6** *The following intersection-closed concept classes have homogeneous spans:*

- (a) *the class  $\mathcal{C}_{X,d} := \{C \subseteq X : |C| \leq d\}$  of all subsets of  $X$  of size at most  $d$ ,*
- (b) *any intersection-closed concept class with unique minimal spans,*
- (c) *all intersection-closed classes  $\mathcal{C}$  that are maximum (cf. [6]), that is, meet the bound of Sauer's Lemma (Theorem 2 above), i.e.  $|\mathcal{C}| = \binom{|X|}{\leq d}$ ,*
- (d) *hyper-rectangles  $\prod_{i=1}^d [0, a_i]$  in  $\mathbb{R}^d$ ,*
- (e) *hyper-rectangles  $\prod_{i=1}^n [0, a_i]$  in  $\mathbb{R}^n$  with at most  $d$  of the  $a_i \neq 0$ ,*
- (f) *intersections of half-spaces  $\bar{H}_{i,b}^- = \{x \in \mathbb{R}^n \mid v_i \cdot x \leq b\}$  from a fixed set of possible orientations  $\{v_1, \dots, v_d\}$ .*

*Proof* (a) Obviously, the spans obtained by setting  $S(Y) := Y$  are homogeneous.

- (b) Setting  $S(Y)$  to the unique minimal span of  $Y$ , it is easy to see that  $\text{clos}(S(Y \setminus x) \cup x) = \text{clos}(Y)$ , so that by the uniqueness assumption  $S(Y) = S(S(Y \setminus x) \cup x)$ . Furthermore, for any  $Y$  one has  $S(Y) \subseteq Y$ , so that in particular  $S(S(Y \setminus x) \cup x) \subseteq S(Y \setminus x) \cup x$ . Combining these two gives  $S(Y) \subseteq S(Y \setminus x) \cup x$ .
- (c) Since by Theorem 1 the number of concepts  $|\mathcal{C}| = \binom{|X|}{\leq d}$  equals the number of possible minimal spans, the latter have to be unique and hence by (b) homogeneous.
- (d) One chooses  $S(Y) := \{\arg \max_{y=(y_1, \dots, y_d) \in Y} y_i \mid 1 \leq i \leq d\}$  with ties broken arbitrarily but consistently (e.g. take the minimal  $y$  with respect to the lexicographic order). Then for any  $x \in S(Y)$ ,  $S(Y \setminus x)$  contains the same elements as  $S(Y)$  only with  $x$  replaced with new maximal points in the respective directions.
- (e) If  $Y$  is contained in any of the hyper-rectangles,  $S(Y)$  can be defined analogously to (d) taking into account only the nonzero coordinates. Otherwise set  $S(Y) = \emptyset$ .
- (f) The homogeneous spans can be defined similarly as in the case of hyper-rectangles. First one chooses an arbitrary point  $y^* \in Y$ . Then one can define  $S(Y) := \{\arg \max_{y \in Y} (y - y^*) \cdot v_i \mid 1 \leq i \leq d\}$  breaking ties arbitrarily as before.  $\square$

However, one can give very simple intersection-closed classes that don't have homogeneous spans:

*Example 2* Consider the following intersection-closed concept class of VC-dimension 2:

$$\mathcal{C} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3, 4\}\}.$$

Possible spans for  $X = \{1, 2, 3, 4\}$  are  $\{2, 3\}$  or  $\{1, 4\}$ , which on the other hand are the only spans for  $\{1, 2, 3\} = X \setminus \{4\}$  and  $\{1, 2, 4\} = X \setminus \{3\}$ , respectively. However, after setting  $S(X) = \{1, 4\}$  the homogeneous span property is violated, since

$$S(X) \setminus \{4\} = \{1\} \not\subseteq \{2, 3\} = S(\{1, 2, 3\}) = S(X \setminus \{4\}).$$

On the other hand, setting  $S(X) = \{2, 3\}$  doesn't give homogeneous spans either:

$$S(X) \setminus \{3\} = \{2\} \not\subseteq \{1, 4\} = S(\{1, 2, 4\}) = S(X \setminus \{3\}).$$

**Theorem 4** *Let  $\mathcal{C}$  be a well-behaved, intersection-closed concept class of VC-dimension  $d$ . If  $\mathcal{C}$  has homogeneous spans, then it is PAC learnable from*

$$m = \left\lceil \frac{e}{\varepsilon} \left( d + \log \frac{1}{\delta} \right) \right\rceil$$

*examples.*

*Proof* Here we adapt the proof of the bound of  $O\left(\frac{1}{\varepsilon}(d + \log \frac{1}{\delta})\right)$  for hyperrectangles in [3], following the main lines of the proof of Theorem 7 on pp.381ff. Similar to the proof of Theorem 3, we start exactly as in the original proof by bounding the probability that  $\mathbf{er}_{\mathcal{C}, \mathcal{P}}(h) > \varepsilon$  with a “permutation trick”: Let  $h$  be the hypothesis generated by a sample of  $m$  labelled examples  $(x_1, y_1), \dots, (x_m, y_m)$ . Then one can show (cf. Lemma 8 of [3], pp.381f) that  $E(\mathbf{er}_{\mathcal{C}, \mathcal{P}}(h)^p)$  is equal to the probability, if one chooses  $m+p$  random examples, that the hypothesis calculated from the first  $m$  examples misclassifies all  $p$  remaining examples. Let  $\pi'(m+p)$  denote the number of possibilities to choose  $m$  from  $m+p$  given examples such that the hypothesis calculated from these  $m$  examples misclassifies the  $p$  remaining examples. Then we have by Markov's inequality that

$$\begin{aligned} P(((x_1, y_1), \dots, (x_m, y_m)) : \mathbf{er}_{\mathcal{C}, \mathcal{P}}(h) > \varepsilon) &= \\ &= P(((x_1, y_1), \dots, (x_m, y_m)) : \mathbf{er}_{\mathcal{C}, \mathcal{P}}(h)^p > \varepsilon^p) \\ &\leq \frac{E(\mathbf{er}_{\mathcal{C}, \mathcal{P}}(h)^p)}{\varepsilon^p} = \frac{\pi'(m+p)}{\varepsilon^p \binom{m+p}{p}}. \end{aligned}$$

Now we may consider the concept class  $\mathcal{C}' = \mathcal{C} \cap \{x_1, \dots, x_{m+p}\}$  instead of  $\mathcal{C}$  itself. Let  $\mathcal{C}''$  be the set of concepts in  $\mathcal{C}'$  that misclassify exactly  $p$

examples among  $(x_1, y_1), \dots, (x_{m+p}, y_{m+p})$ . We will show that  $|\mathcal{C}''| \leq \binom{d+p}{p}$ , so that the probability that  $\mathbf{er}_{\mathcal{C}, \mathcal{P}}(h) > \varepsilon$  is at most

$$\frac{\binom{d+p}{p}}{\varepsilon^p \binom{m+p}{p}} = \frac{(d+p) \cdot \dots \cdot (d+1)}{\varepsilon^p (m+p) \cdot \dots \cdot (m+1)} \leq \left( \frac{d+p}{\varepsilon m} \right)^p,$$

for which after choosing  $p = \lceil \log \frac{1}{\delta} \rceil$  and  $m \geq \frac{\varepsilon(d+p)}{\varepsilon}$  one obtains

$$\left( \frac{d+p}{\varepsilon m} \right)^p \leq \left( \frac{1}{e} \right)^{\lceil \log \frac{1}{\delta} \rceil} \leq \delta,$$

so that the theorem follows.

Again using the closure algorithm, only the positive examples  $X_1^+ = \{x_i \mid y_i = 1, 1 \leq i \leq m+p\}$  are relevant for hypothesis calculation and evaluation. Since examples may occur more than once, we also consider the corresponding multiset  $X_1^{+, \text{mult}}$  with possible multiple occurrences of elements.<sup>4</sup>

Now we want to encode each concept  $C$  in  $\mathcal{C}''$  according to its classification of some particular examples  $x_1, \dots, x_{d+p} \in X_1^+$ . That is,  $C \in \mathcal{C}''$  is encoded as a word in  $\{0, 1\}^{d+p}$  as follows: a 1 on the  $j$ -th position means that  $C$  classifies  $x_j$  correctly, while a 0 indicates that  $x_j$  is misclassified by  $C$ . In order to be able to guarantee the uniqueness of each code word, we impose an arbitrary yet fixed order on the elements in  $X_1^+$ . Now choose the smallest element  $x_1$  from  $S(X_1^+)$ . If  $C$  classifies  $x_1$  correctly, that is,  $x_1 \in C$ , we claim that  $x_1 \in S(C)$ . Indeed, setting  $Z := X_1^+ \setminus C$ , it follows by Proposition 5 that

$$x_1 \in S(X_1^+) \setminus Z \subseteq S(X_1^+ \setminus Z) = S(C).$$

We set the first letter of the code word for  $C$  to 1 and continue with the smallest element  $x_2 \in S(X_1^+) \setminus \{x_1\}$ . Otherwise, if  $C$  misclassifies  $x_1$ , the first letter of  $C$ 's code word is set to 0, and we continue with the smallest element from  $S(X_1^{+, \text{mult}} \setminus \{x_1\})$ . Repeating this procedure we obtain for each  $C$  a sequence of elements  $x_i$  in  $X_1^{+, \text{mult}}$  which corresponds to a word  $\in \{0, 1\}^*$ . Note that each sequence is uniquely determined by the corresponding word, so that it is not possible that two different concepts obtain the same code word.

Now, each positive classification of an  $x_j$  determines an element of  $S(C)$  and there are exactly  $p$  misclassifications. Since by Definition 3 the spanning set  $S(C)$  has at most  $d$  elements, it follows that each concept  $\in \mathcal{C}''$  can be encoded as word in  $\{0, 1\}^{d+p}$  consisting of  $p$  letters 0 and  $d$  letters 1. The number of these words is equal to  $\binom{d+p}{p}$ , which finishes our proof.  $\square$

---

<sup>4</sup> Accordingly, the operation  $\setminus$  is adapted such that e.g.  $\{1, 1, 2, 3\} \setminus \{1\} = \{1, 2, 3\}$ .

The following theorem shows that for our new bounds the choice of the learning algorithm is essential, as there are concept classes that need  $\Omega\left(\frac{1}{\varepsilon}(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta})\right)$  examples to be learned by particular consistent learning algorithms.

**Theorem 5** *Let  $X$  be an arbitrary set and  $\mathcal{C}_{X,d}$  the class of all subsets of  $X$  of size at most  $d$ . Furthermore, let  $\mathcal{A}$  be an algorithm that chooses as its hypothesis not the smallest concept consistent with the given examples (as the closure algorithm does), but an arbitrarily selected largest consistent concept. Then  $\mathcal{A}$  needs  $\Omega\left(\frac{1}{\varepsilon}(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta})\right)$  examples to learn  $\mathcal{C}_{X,d}$ .*

*Proof* The main step is to show a lower bound of  $\Omega\left(\frac{d}{\varepsilon} \log \frac{1}{\varepsilon}\right)$ . Let  $X$  consist of  $n := \lceil \frac{d}{\varepsilon} \rceil$  elements, and let  $\mathcal{P}$  be the uniform distribution on  $X$ . When learning the target concept  $\emptyset \in \mathcal{C}_{X,d}$  the error of the algorithm's hypothesis is smaller than  $\varepsilon$  only if at least  $(1 - \varepsilon)n$  distinct examples appear among the training examples. Since  $n - d \geq (1 - \varepsilon)n > n - (d + 1)$ , this means that one needs  $\geq n - d$  distinct examples in the training set. The probability that a certain example is not among  $m$  randomly chosen training examples is  $(1 - \frac{1}{n})^m$ . Let  $Z$  be a random variable denoting the number of examples in  $X$  that are not in the training set. Thus,  $\mu := E(Z) = n(1 - \frac{1}{n})^m$ . Furthermore, the probability that two distinct examples are not among the training examples is  $(1 - \frac{2}{n})^m$ , so that  $E(Z^2) = (n^2 - n)(1 - \frac{2}{n})^m + n(1 - \frac{1}{n})^m$ .

Now for  $a < \mu$  we have

$$P(Z \leq a) = P(\mu - Z \geq \mu - a) \leq P((Z - \mu)^2 \geq (\mu - a)^2).$$

Choosing  $a = \frac{\mu}{2}$ , it follows by Chebyshev's inequality that

$$P(Z \leq \frac{\mu}{2}) \leq \frac{\text{Var}(Z)}{(\mu - \frac{\mu}{2})^2} = 4 \cdot \frac{E(Z^2) - \mu^2}{\mu^2}. \quad (3)$$

Some straightforward calculation gives

$$\begin{aligned} \frac{E(Z^2)}{\mu^2} &= \frac{n(n-1) \left(\frac{n-2}{n}\right)^m}{n^2 \left(\frac{n-1}{n}\right)^{2m}} + \frac{n \left(\frac{n-1}{n}\right)^m}{n^2 \left(\frac{n-1}{n}\right)^{2m}} \\ &\leq \left(\frac{n-2}{n}\right)^m \left(\frac{n}{n-1}\right)^{2m} + \frac{1}{n} \left(\frac{n}{n-1}\right)^m \\ &\leq \left(\frac{n(n-2)}{(n-1)^2}\right)^m + \frac{1}{n} \left(\frac{n}{n-1}\right)^m \leq 1 + \frac{1}{n} \left(\frac{n}{n-1}\right)^m. \end{aligned} \quad (4)$$

Now if  $m = n \log \frac{1}{16\varepsilon}$ , it follows from (3) and (4) that

$$\begin{aligned} P(Z \leq \frac{\mu}{2}) &\leq \frac{4}{n} \left( \left(\frac{n}{n-1}\right)^n \right)^{\log \frac{1}{16\varepsilon}} = \frac{4}{n} \left( \frac{n}{n-1} \left(\frac{n}{n-1}\right)^{n-1} \right)^{\log \frac{1}{16\varepsilon}} \\ &\leq \frac{4}{n} \left(\frac{n}{n-1}\right)^{\log \frac{1}{16\varepsilon}} e^{\log \frac{1}{16\varepsilon}} = \frac{4}{n} \cdot \frac{1}{16\varepsilon} \left(\frac{n}{n-1}\right)^{\log \frac{1}{16\varepsilon}}. \end{aligned} \quad (5)$$

Substituting  $n = \lceil \frac{d}{\varepsilon} \rceil$ , we have on one hand  $\frac{4}{n} \cdot \frac{1}{16\varepsilon} \leq \frac{1}{4d}$ . On the other hand, since  $d \geq 1$  we have  $(\frac{n}{n-1})^{\log \frac{1}{16\varepsilon}} \leq (\frac{1}{1-\varepsilon})^{\log \frac{1}{16\varepsilon}} < 2$ . It follows from (5) that  $P(Z \leq \frac{\mu}{2}) < \frac{1}{2d} \leq \frac{1}{2}$  and consequently  $P(Z > \frac{\mu}{2}) > \frac{1}{2}$ .

Now assume that  $\varepsilon \leq \frac{1}{16}$  such that  $\frac{d}{\varepsilon} = \lceil \frac{d}{\varepsilon} \rceil$ . Then since  $\log(1 - \frac{\varepsilon}{d}) \geq -\frac{\varepsilon}{d} - \frac{\varepsilon^2}{d^2}$  for  $\frac{\varepsilon}{d} < \frac{1}{2}$ , we have

$$\begin{aligned} \mu &= \frac{d}{\varepsilon} \left(1 - \frac{\varepsilon}{d}\right)^{\frac{d}{\varepsilon} \log \frac{1}{16\varepsilon}} \geq \frac{d}{\varepsilon} \exp \left\{ \frac{d}{\varepsilon} \log \frac{1}{16\varepsilon} \left( -\frac{\varepsilon}{d} - \frac{\varepsilon^2}{d^2} \right) \right\} \\ &= \frac{d}{\varepsilon} \exp \left\{ -\log \frac{1}{16\varepsilon} \left(1 + \frac{\varepsilon}{d}\right) \right\} = \frac{d}{\varepsilon} \left( e^{-\log \frac{1}{16\varepsilon}} \right)^{1 + \frac{\varepsilon}{d}} \\ &= \frac{d}{\varepsilon} (16\varepsilon)^{1 + \frac{\varepsilon}{d}} = 16d (16\varepsilon)^{\frac{\varepsilon}{d}} \geq 16d (16\varepsilon)^\varepsilon > 8d, \end{aligned}$$

since for  $\varepsilon \leq \frac{1}{16}$  one has  $(16\varepsilon)^\varepsilon > \frac{1}{2}$ . Summarizing, this yields  $\frac{\mu}{2} > \frac{\mu}{8} \geq d$  for sufficiently small  $\varepsilon > 0$  with  $\frac{d}{\varepsilon} = \lceil \frac{d}{\varepsilon} \rceil = n$  and for  $m = n \log \frac{1}{16\varepsilon}$ .

Now because  $P(Z > \frac{\mu}{2}) > \frac{1}{2}$ , it follows that  $P(Z \geq d+1) > \frac{1}{2}$ . Since the probability that the error of the algorithm's hypothesis exceeds  $\varepsilon$  is at least as large as  $P(Z \geq d+1)$ , it follows that one needs at least  $\Omega\left(\frac{d}{\varepsilon} \log \frac{1}{\varepsilon}\right)$  examples to learn  $\mathcal{C}_{X,d}$ . Combining this with the well-known lower bound of  $\Omega\left(\frac{1}{\varepsilon} \log \frac{1}{\delta}\right)$  (cf. [4] for details), one obtains the claimed lower bound of  $\Omega\left(\frac{1}{\varepsilon} \left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}\right)\right)$ .  $\square$

## 5 Final Remarks

The extension of our result for intersection-closed concept classes with homogeneous spans to intersection-closed concept classes in general seems to be far from trivial. Proposition 4 is obviously not strong enough to impose some kind of structure that is sufficient to obtain the desired bound. Thus, we think that some new insights into the combinatorial properties of intersection-closed classes will be needed to make further progress.

*Acknowledgements* We would like to thank Manfred Warmuth and Thomas Kormort for helpful discussion. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

## References

1. Auer, P. (1997). Learning Nested Differences in the Presence of Malicious Noise. *Theor. Comput. Sci.*, 185:1, 159–175.
2. Auer, P., & Cesa-Bianchi, N. (1998). On-Line Learning with Malicious Noise and the Closure Algorithm. *Ann. Math. Artif. Intell.*, 23:1-2, 83–99.

3. Auer, P., Long, P.M., & Srinivasan, A. (1998). Approximating Hyper-Rectangles: Learning and Pseudorandom Sets. *J. Comput. Syst. Sci.*, 57:3, 376–388.
4. Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1989). Learnability and the Vapnik-Chervonenkis Dimension. *J. ACM*, 36:4, 929–965.
5. Ehrenfeucht, A., Haussler, D., Kearns, M.J., & Valiant, L.G. (1989). A General Lower Bound on the Number of Examples Needed for Learning. *Inf. Comput.*, 82:3, 247–261.
6. Floyd, A., & Warmuth, M. (1995). Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension. *Machine Learning*, 21:3, 269–304.
7. Haussler, D., Littlestone, N., & Warmuth, M. (1994). Predicting  $\{0,1\}$ -Functions on Randomly Drawn Points. *Inf. Comput.* 115:2, 248–292.
8. Helmbold, D., Sloan, R., & Warmuth, M. (1990). Learning Nested Differences of Intersection-Closed Concept Classes. *Machine Learning* 5, 165–196.
9. Sauer, N. (1972). On the Density of Families of Sets. *J. Combin. Theory Ser. A* 13, 145–147.