Yale University
Department of Computer Science

Identifying k-CNF formulas from noisy examples

Dana Angluin and P. D. Laird, Yale University

YALEU/DCS/TR-478
June 1986

Research funded in part by the National Science Foundation, DCR-8404226.

Abstract

We consider the problem of identifying an unknown subset L of a universal set U, given an oracle that randomly samples and reports elements of U and classifies them as to membership in L. The oracle is assumed to make independent random errors in classifying the reported elements. Valiant has shown that in the absence of errors, there is an efficient procedure to identify sets described by k-CNF formulas that produces an "approximately correct" identification with "high probability". The main result of this paper is to extend Valiant's result to the case of noisy oracles, provided that the error rate is less than $1/2$. We also give some general results indicating how many examples suffice to compensate for this kind of error.

Contents

Identifying k-CNF formulas from noisy examples

Dana Angluin and P. D. Laird, Yale University

1 Introduction

The ability to form general concepts on the basis of particular examples is an essential ingredient of intelligent behavior. If the examples may contain errors, the task of useful generalization becomes harder. In this paper we address the question of how to compensate for randomly introduced errors, or "noise", in the example data. The examples are assumed to be generated by a sampling procedure that randomly mis-classifies examples of a concept as non-examples and vice versa. The criterion of correct identification we adopt is that of "probably approximately correct identification", introduced by Valiant [6]. The main result of the paper is that identification of k-CNF formulas is feasible under this error model, provided that the error probability is less than $1/2$.

In the remainder of this section we define the notion of "probably approximately correct identification", give an example of it, introduce our model of random errors in the data, compare it with Valiant's, and summarize our results.

1.1 Probably approximately correct identification

Valiant has proposed a general criterion of correct identification of a concept from examples in a stochastic setting [6]. The idea is that after randomly sampling examples and non-examples of a concept, an identification procedure should conjecture a concept that with "high probability" is "not too different" from the correct concept. This is quantified in the following definitions.

Let $L_1, L_2, \ldots$ be a countable family of subsets of a countable universe U, and let D be an unknown probability distribution on the elements of U. The task is to identify an unknown one of these sets, L_*, given access only to a sampling oracle $EX()$. Each call to $EX()$ randomly selects an element x from the universe

Research funded in part by the National Science Foundation, DCR-8404226.

U according to the distribution D and returns $\langle x, + \rangle$ if $x \in L_*$, and returns $\langle x, - \rangle$ otherwise

An identification procedure makes a number of calls to $EX()$ and then conjectures one of the sets, L_h. The success of the identification is measured by two parameters, ϵ and δ, which are given as inputs to the identification procedure.

The parameter ϵ is a bound on the "difference" between the conjectured set L_h and the unknown set L_*. Define

$$d(S,T) = \sum_{x \in S \oplus T} \Pr(x),$$

where S and T are any subsets of U, $S \oplus T$ is the symmetric difference of S and T, and Pr denotes probability with respect to the distribution D. Thus, $d(S,T)$ is precisely the probability that in one call to $EX()$ we will draw an element that is in one but not the other of the two sets.

The parameter δ is a confidence parameter; because the calls to $EX()$ are random experiments, there is always the possibility of getting a wildly unrepresentative sample and drawing a ridiculous conclusion. The parameter δ is a bound on how likely this is.

Putting this together, the identification procedure is said to do *probably approximately correct identification* of L_* if and only if

$$\Pr[d(L_*, L_h) \leq \epsilon] \geq 1 - \delta.$$

We abbreviate probably approximately correct identification as *pac-identification.*

Blumer et al. have investigated this notion of identification and given necessary and sufficient conditions for it in terms of the structure of the hypothesis space [3]. See their paper also for definitions encompassing the case of continuous distributions.

1.2 An example: k-CNF formulas

If n and k are positive integers, $CNF(n,k)$ denotes the class of all conjunctive normal form propositional formulas over the variables $x_1, x_2, ..., x_n$ with at most k literals per clause. For a fixed n, the universe U is the set of all truth assignments a mapping $x_1, x_2, ..., x_n$ to the set $\{0,1\}$. A formula ϕ in $CNF(n,k)$ is interpreted as representing the set of all assignments a from U that satisfy ϕ, i.e., such that $a(\phi) = 1$. The sampling oracle $EX()$ returns assignments (represented as n-vectors of 0's and 1's for concreteness) marked either $+$ or $-$ according to whether they satisfy the unknown formula ϕ_*.

Valiant [6] has shown that there is an identification procedure V that takes as input n, k, ϵ, and δ, has access to the sampling oracle $EX()$ for an unknown formula ϕ_*, runs in time polynomial in n^k, $1/\epsilon$, and $\log 1/\delta$, and does *pac*-identification of ϕ_*, for any ϕ_* from $CNF(n,k)$.

The procedure V calculates from n, k, ϵ, and δ a number, m, of samples to draw, makes m calls to $EX()$, and then outputs the conjunction of all clauses over $x_1, x_2, ..., x_n$ with at most k literals per clause that are satisfied by every positive example, i.e., by every assignment a such that some call to $EX()$ returned the value $\langle a, + \rangle$.

1.3 A simple model of errors

We introduce a model of random errors, or "noise", in the sampling oracle $EX()$. We assume that the sampling oracle is able to draw elements from the relevant distribution D without error, but that the process of determining and reporting whether the example is a positive or negative one is subject to independent random errors with some unknown probability $\eta < 1/2$. That is, the experiment performed by $EX()$ is now assumed to be: draw a random element x from U according to the distribution D, and then flip a coin that comes up heads with probability $1 - \eta$. If the coin comes up heads, report x with the correct sign, otherwise, report x with the reverse of the correct sign. To indicate that the oracle is subject to errors of this type, it will be denoted $EX_\eta()$. $EX_0()$ is the sampling oracle with no errors of reporting.

Why do we restrict η to be less than $1/2$? Clearly, when $\eta = 1/2$, the errors in the reporting process destroy all possible information about membership in the unknown set L_*, and no identification procedure could be expected to work. When $\eta > 1/2$, there is information about L_*, but it is equally information about the complement of L_* with the smaller error $1 - \eta$. While in principle we might be able to recognize this situation in domains which are not closed under complement with respect to U, we have chosen not to pursue this possibility.

What if η is very close to $1/2$? How could an identification procedure be expected to work? We assume that there is some information about η available as input to the identification procedure, namely an upper bound η_b such that $\eta \leq \eta_b < 1/2$. Just as an "efficient" identification procedure is permitted in the absence of errors to run in time polynomial in $1/\epsilon$ and $1/\delta$, in the presence of errors we will permit the polynomial to have $1/(1 - 2\eta_b)$ as one of its arguments. This quantity is inversely proportional to how close η_b is to $1/2$, so the closer the upper bound on the error rate is to $1/2$, the longer the identification procedure will be permitted to run.

How general is this model of errors? It seems appropriate to a setting in which there is an observable, reliable mechanism selecting examples, and a separate, noisy one classifying them. However, there are many situations for which this is not a reasonable assumption, for example the following one. Suppose that correct examples are being transmitted over a noisy line (say, with independent errors in each bit); then not only is the sign of the example subject to errors, but a given example x may be changed into another one x'. In this case, the examples x' reported by the sampling oracle may come from a different distribution D'. Even if our results were applicable in this situation, the "difference" of the hypothesis from the correct set will be measured with respect to the observed distribution D' instead of the true distribution D, which is not necessarily what is wanted.

1.4 Valiant's error model

Valiant [5] has considered a rather different error model for the problem of identifying disjunctive normal form formulas (or dually, conjunctive normal form formulas) from examples. He assumes a small rate of errors, but permits the errors to be maliciously rather than randomly chosen. In particular, for an error rate of η an error-prone sampling oracle first flips a coin with probability of success $1-\eta$. If the coin comes up heads, the sampling and reporting process proceed without error. If the coin comes up tails, the oracle may give any response. Valiant gives an algorithm that efficiently identifies k-DNF (dually, k-CNF) formulas in the presence of such errors, provided that the error rate is sufficiently small (proportional to the minimum of ϵ and δ and inversely proportional to the number of k-CNF clauses.)

Valiant suggests the possibility that "the learning phenomenon is only feasible with very low error rates". The results of this paper show that nontrivial learning is in principle still feasible at quite high error rates, provided the errors are randomly generated in a sufficiently straightforward way.

1.5 Summary of results

In the next section we consider any finite space of hypotheses, $L_1, L_2, ..., L_N$, and show that there is a polynomial p such that a sample of size

$$m \geq p(\log N, 1/\epsilon, \log 1/\delta, 1/(1-2\eta_b))$$

contains enough information to do *pac*-identification of each L_i from noisy samples. The following section shows that the approach used to prove this theorem is not in general computationally tractable, so that we need different methods to achieve *pac*-identification from noisy samples in a practical sense.

The main result of the paper is that there is an efficient procedure to do *pac*-identification of k-CNF formulas from noisy samples. More precisely, there is a procedure V' that takes as inputs n, k, ϵ, δ, and η_b, has access to a sampling oracle $EX_\eta()$ for some unknown $\eta \leq \eta_b$ and some unknown formula ϕ_*, runs in time polynomial in n^k, $1/\epsilon$, $\log 1/\delta$, and $1/(1-2\eta_b)$, and does *pac*-identification of ϕ_*, for every ϕ_* in $CNF(n,k)$.

Thus, even in the presence of fairly large random errors in the reporting of examples, k-CNF formulas can be efficiently *pac*-identified. Note in particular that the error rate η may be much larger than the accuracy and confidence parameters ϵ and δ. The procedure V' may be viewed as a modification of the procedure V of Valiant, in that it samples to estimate the error rate η and then discards clauses whose failure rate is "significantly larger" than this estimate. The procedure V' is presented and analyzed in Section 4.

2 How many examples suffice in the presence of noise?

Forgetting for a moment the question of computational feasibility, how can we be sure that there is enough information in a certain number of samples drawn from a noisy oracle to determine the unknown set L_* up to ϵ-equivalence with probability at least $1-\delta$? Suppose the space of possible hypotheses is finite, say $L_1, L_2, ..., L_N$. For the error-free case, Blumer et al. [3] have shown the following.

Theorem 1 *There is a polynomial p such that if L_i is any hypothesis that agrees with*

$$m \geq p(\log N, 1/\epsilon, \log 1/\delta)$$

samples drawn from the $EX()$ oracle, then

$$\Pr[d(L_i, L_*) \leq \epsilon] \geq 1-\delta.$$

Thus, in the absence of errors, it suffices to find any hypothesis that is consistent with all of a sufficiently large collection of examples.

How is this modified in the presence of errors? Because of the errors, there is no guarantee that any of the hypotheses will be consistent with all of a particular collection of samples drawn from some $EX_\eta()$. However, if we replace the notion of consistency with that of minimizing the number of disagreements with the examples, and permit the number of samples to depend on the upper bound η_b on the error rate, we get an analogous result, Theorem 2.

Let

$$\sigma = \langle x_1, s_1\rangle, \langle x_2, s_2\rangle, ..., \langle x_m, s_m\rangle$$

denote a sequence of samples drawn from an $EX_\eta()$ oracle, where each x_i is in the universe U and each s_i is either $+$ or $-$. If L_i is any possible hypothesis, let $F(L_i, \sigma)$ denote the number of indices j for which L_i disagrees with $\langle x_j, s_j \rangle$, that is, $s_j = +$ and x_j is not in L_i or $s_j = -$ and x_j is in L_i.

Theorem 2 *There is a polynomial p such that if we draw a sequence σ of*

$$m \geq p(\log N, 1/\epsilon, \log 1/\delta, 1/(1-2\eta_b))$$

samples from an $EX_\eta()$ oracle and find any hypothesis L_i that minimizes $F(L_i, \sigma)$, then

$$\Pr[d(L_i, L_*) \leq \epsilon] \geq 1 - \delta.$$

Note that the dependence on the size of the hypothesis space and the accuracy and confidence parameters remains essentially the same as in Theorem 1.
Proof: We analyze the expected rate of disagreement between any hypothesis L_i and sample sequences produced by the oracle $EX_\eta()$ with unknown set L_*. Let

$$d_i = d(L_i, L_*).$$

The probability that an example produced by $EX_\eta()$ disagrees with L_i is the probability that an example is drawn from $L_i \oplus L_*$ and reported correctly (which is just $d_i(1-\eta)$) plus the probability that an example is drawn from the complement of $L_i \oplus L_*$ and reported incorrectly (which is just $(1-d_i)\eta$.) Let p_i denote the probability that an example from $EX_\eta()$ disagrees with L_i, then we have

$$p_i = d_i(1-\eta) + (1-d_i)\eta.$$

In the case that the hypothesis L_i is equal to L_*, we have $p_i = \eta$, since disagreements will only arise as the result of reporting errors.

The expression for p_i may be rewritten as

$$p_i = \eta + d_i(1-2\eta).$$

Since $\eta < 1/2$, this shows that any hypothesis L_i has an expected rate of disagreement of at least η. In particular, if we define a hypothesis L_i to be *ϵ-bad* if and only if $d_i \geq \epsilon$, then for any ϵ-bad hypothesis L_i we have

$$p_i \geq \eta + \epsilon(1-2\eta).$$

Thus we have a separation of at least $\epsilon(1-2\eta)$ between the disagreement rates of correct and ϵ-bad hypotheses. By our assumptions, η is not known, but an upper

bound $\eta_b < 1/2$ is known, so we have a known lower bound on the separation of $\epsilon(1-2\eta_b)$.

The problem is reduced to guaranteeing that the number m of samples drawn from $EX_\eta()$ is sufficient to guarantee that no ϵ-bad hypothesis has a lower observed rate of disagreement with the samples than L_*, with probability greater than $(1-\delta)$.

At this point we need to introduce a little notation. If p and r are numbers between 0 and 1 and m is a non-negative integer, let $GE(p,m,r)$ denote the probability of at least rm successes in m independent Bernoulli trials with probability p, and let $LE(p,m,r)$ denote the probability of at most rm successes in m independent Bernoulli trials with probability p. Lemmas bounding these quantities in various ways are given in the Appendix.

Let $s = \epsilon(1-2\eta_b)$. Let σ denote a sequence of

$$m \geq \frac{48}{\epsilon^2(1-2\eta_b)^2} \ln \frac{2N}{\delta}$$

examples drawn from the noisy sampling oracle $EX_\eta()$. This bound is polynomial in $\log N$, $1/\epsilon$, $\log 1/\delta$, and $1/(1-2\eta_b)$.

In order for some ϵ-bad hypothesis to minimize $F(L_i,\sigma)$, either

$$F(L_*,\sigma)/m \geq \eta + s/2$$

or

$$F(L_i,\sigma)/m \leq \eta + s/2$$

for some ϵ-bad hypothesis L_i, or both. Applying Lemma 12 in the Appendix,

$$\begin{aligned} \Pr[F(L_*,\sigma)/m \geq \eta + s/2] &= GE(\eta, m, \eta + s/2) \\ &\leq \delta/2N \\ &\leq \delta/2, \end{aligned}$$

and if L_i is ϵ-bad then

$$\begin{aligned} \Pr[F(L_i,\sigma)/m \leq \eta + s/2] &\leq LE(\eta + s, m, \eta + s/2) \\ &\leq \delta/2N. \end{aligned}$$

Thus the probability that any ϵ-bad hypothesis L_i has $F(L_i,\sigma)/m \leq \eta + s/2$ is at most $\delta/2$, since there are at most $N-1$ ϵ-bad hypotheses. Putting these two inequalities together, the probability that some ϵ-bad hypothesis minimizes $F(L_i,\sigma)$ is at most δ. □

Applying Theorem 2 to the case of formulas from $CNF(n,k)$, the logarithm of the size of the hypothesis space is polynomial in n^k, so this result indicates that there is enough information in a sample of size polynomial in n^k, $1/\epsilon$, $\log 1/\delta$, and $1/(1-2\eta_b)$ to *pac*-identify formulas from $CNF(n,k)$ in the presence of errors. However, the particular method of minimizing conflicts with a sample is in general not a computationally feasible approach, as the next section indicates.

3 Minimizing disagreements may not be feasible

The computational approach suggested by Theorem 2 is to draw a sample sequence of the appropriate size from $EX_\eta()$ and then find a hypothesis that minimizes disagreements with this sequence of samples. In general this may not be a computationally feasible problem, as the following NP-hardness result demonstrates.

Let n be a positive integer. Let $PP(n)$ denote the set of all products of a subset of the literals $x_1, x_2, ..., x_n$. There are 2^n such products; the empty product is interpreted as equivalent to "true". Each product π in $PP(n)$ is interpreted as denoting the set of truth-value assignments that satisfy it. $PP(n)$ is a subset of the formulas in $CNF(n,1)$.

A sample sequence σ will consist of a finite sequence of ordered pairs of the form $\langle a_j, s_j \rangle$ where a_j is a truth-value assignment to the variables $x_1, x_2, ..., x_n$ and s_j is either $+$ or $-$. If $\pi \in PP(n)$ and σ is a sample sequence, then $F(\pi,\sigma)$ is the number of pairs $\langle a_j, s_j \rangle$ in σ such that $s_j = +$ and $a_j(\pi) = 0$ or $s_j = -$ and $a_j(\pi) = 1$. That is, $F(\pi,\sigma)$ is the number of disagreements between π and the sample sequence σ.

Theorem 3 *The problem of determining, given positive integers n and c and a sample sequence σ, whether there is an element $\pi \in PP(n)$ such that $F(\pi,\sigma) \leq c$ is NP-complete.*

Proof: The proof is a polynomial time reduction of the vertex cover problem to the specified problem. The vertex cover problem is specified by an undirected graph G of n vertices and a positive integer $c \leq n$, and the question is whether there exists a set C of at most c vertices of G such that every edge of G is incident to at least one vertex in C. (Such a set C is called a vertex cover.) The vertex cover problem is NP-complete.

Let a vertex cover problem, $\langle G, c \rangle$, be given. Suppose the vertices of G are $v_1, v_2, ..., v_n$. There will be n variables: $x_1, x_2, ..., x_n$. For each vertex v_i, define a truth assignment a_i that maps x_i to 0 and every other x_j to 1. For each edge $e = \{v_i, v_j\}$, define a truth assignment b_e that maps x_i and x_j to 0 and every other

x_k to 1. The sample sequence σ consists of one copy of $\langle a_i, + \rangle$ for each vertex v_i and $n+1$ copies of $\langle b_e, - \rangle$ for each edge e in G.

Then we claim that G has a vertex cover of at most c vertices if and only if there is an element π of $PP(n)$ such that $F(\pi, \sigma) \leq c$.

Suppose G has a vertex cover C of at most c vertices. Let π denote the product of those x_i such that v_i is in C. How many examples from σ disagree with π? For each vertex v_i, the assignment a_i assigns 0 to π if and only if $v_i \in C$. Thus, π disagrees with at most c positive examples from σ. For each edge $e = \{v_i, v_j\}$, the set C contains at least one of v_i or v_j, so the product π contains at least one of x_i or x_j. Since the assignment b_e is 0 on both x_i and x_j, it must be 0 on π. Thus, π agrees with all the negative examples in σ. Hence $F(\pi, \sigma) \leq c$, as claimed.

Suppose now that there exists some $\pi \in PP(n)$ such that $F(\pi, \sigma) \leq c$. Since $c \leq n$, this means that π must agree with all the negative examples in σ, since each one is repeated $n+1$ times. Hence π can only disagree with positive examples in σ, and at most c of them. Thus π must contain at most c literals x_i. Define the set C to be all those vertices v_i such that x_i appears in the product π. Then C contains at most c vertices; it remains to see that it is a vertex cover. If $e = \{v_i, v_j\}$ is any edge in G then the assignment b_e must assign 0 to π, since π agrees with all the negative examples. But b_e assigns 0 to π if and only if π contains at least one of x_i or x_j. Thus C contains at least one of v_i or v_j, so C is a vertex cover of G.

The computation of n, c, and σ from $\langle G, c \rangle$ can clearly be carried out in polynomial time. □

This indicates that even for a very simple domain the approach of directly trying to minimize the number of disagreements with the sample may not be computationally feasible. We show in the next section that a somewhat more sophisticated approach does permit efficient *pac*-identification of k-CNF formulas from noisy samples.

4 Efficient *pac*-identification of k-CNF formulas in the presence of noise

In this section we describe an efficient procedure V' that does *pac*-identification of k-CNF formulas. The inputs to the procedure are n, k, ϵ, δ, η_b, and a noisy oracle $EX_\eta()$ for an unknown formula ϕ_* from $CNF(n,k)$, using an unknown distribution D to sample truth-assignments. The accuracy and confidence parameters ϵ and δ must be between 0 and 1, and the error bound η_b is such that $0 \leq \eta \leq \eta_b < 1/2$.

Once n and k are fixed, there is a set $\mathcal{C}$ of all possible clauses over the variables $x_1, \ldots, x_n$ with at most k literals per clause. Let M denote the cardinality of $\mathcal{C}$.

Clearly M is at most $(2n+1)^k$.

4.1 Motivation for the procedure V'

Once D is fixed we define two probabilities for each clause C from $\mathcal{C}$:

$$\begin{aligned} p_0(C) &= \Pr[a(C) = 0] \\ p_1(C) &= \Pr[a(C) = 1]. \end{aligned}$$

If ϕ_* is also fixed, we may subdivide these probabilities into four cases, p_{rs}, for $r = 0, 1$ and $s = 0, 1$ as follows:

$$p_{rs}(C) = \Pr[a(C) = r \text{ and } a(\phi_*) = s].$$

Note that $p_0(C) = p_{00}(C) + p_{01}(C)$.

We use these probabilities to classify each clause as follows. A clause C is defined to be *important* if and only if

$$p_0(C) \geq Q_I,$$

where

$$Q_I = \epsilon(1 - 2\eta_b)/16M^2.$$

A clause C is defined to be *harmful* if and only if

$$p_{01}(C) \geq Q_H,$$

where

$$Q_H = \epsilon/2M.$$

Note that $Q_H \geq Q_I$, so every harmful clause is important. Note also that no clause contained in or implied by ϕ_* can be harmful.

The intuition is that a clause that is not important is almost always assigned the value 1 by assignments chosen according to D, so it may be included or not in the final hypothesis without significantly affecting the outcome. On the other hand, a harmful clause is one for which a significant fraction of the assignments chosen from D make the clause 0 but the correct hypothesis 1. If a harmful clause is included in the final hypothesis, it will cause a nontrivial probability of disagreement between the final hypothesis and the correct hypothesis. Thus, the strategy of the procedure V' is to attempt to include in the final hypothesis all the important clauses contained in ϕ_* and no harmful clauses. Our first lemma shows that if V' succeeds in this attempt, then the final hypothesis is indeed an ϵ-approximation of ϕ_*.

Lemma 4 *Let D and ϕ_* be fixed. Let ϕ be any product of clauses from $\mathcal{C}$ that contains every important clause in ϕ_* and contains no harmful clauses. Then $d(\phi, \phi_*) < \epsilon$.*

Proof: We analyze the probability of an assignment a such that $a(\phi_*) = 1$ and $a(\phi) = 0$ or vice versa. Let $\phi - \phi_*$ denote the set of clauses in ϕ but not in ϕ_*.

$$\begin{aligned}
\Pr[a(\phi_*) = 1 \text{ and } a(\phi) = 0] &\leq \sum_{C \in \phi - \phi_*} p_{01}(C), \\
&< MQ_H, \text{ since no element of } \phi - \phi_* \text{ is harmful,} \\
&< \epsilon/2.
\end{aligned}$$

For the other side,

$$\begin{aligned}
\Pr[a(\phi) = 1 \text{ and } a(\phi_*) = 0] &\leq \sum_{C \in \phi_* - \phi} p_0(C), \\
&< MQ_I, \text{ since no element of } \phi_* - \phi \text{ is important,} \\
&< \epsilon/2.
\end{aligned}$$

Thus,

$$\Pr[a(\phi) \neq a(\phi_*)] < \epsilon/2 + \epsilon/2.$$

□

The procedure V' has no direct information about whether a clause is important or harmful – it must rely on the noisy oracle $EX_\eta()$ for its information about D and ϕ_*. Since the oracle $EX_\eta()$ reports assignments according to the distribution D, $p_0(C)$ can be directly estimated by sampling the oracle and calculating the fraction of assignments that assign 0 to C. The procedure V' uses this to construct a set I that with high probability contains all the important clauses C from $\mathcal{C}$. If this is accomplished, the remaining problem is to identify all the harmful clauses in I. (Note that V' depends in an essential way upon the fact that, in this model, the distribution D is not perturbed by the presence of noise.)

However, the definition of a harmful clause refers to the values of assignments on ϕ_*, which are subject to reporting errors and cannot be estimated directly. For each clause C we define two more probabilities:

$$\begin{aligned}
p_{0-}(C) &= \Pr[\text{a sample } \langle a, s \rangle \text{ drawn from } EX_\eta() \text{ has } a(C) = 0 \text{ and } s = -], \\
p_{0+}(C) &= \Pr[\text{a sample } \langle a, s \rangle \text{ drawn from } EX_\eta() \text{ has } a(C) = 0 \text{ and } s = +].
\end{aligned}$$

These may be directly estimated using calls to $EX_\eta()$. A sample $\langle a, s \rangle$ will have $a(C) = 0$ and $s = +$ if and only if either $a(C) = 0$ and $a(\phi_*) = 1$ and there was no

reporting error, or $a(C) = 0$ and $a(\phi_*) = 0$ and there was a reporting error. Thus

$$\begin{aligned} p_{0+}(C) &= (1-\eta)p_{01}(C) + \eta p_{00}(C) \\ &= \eta(p_{00}(C) + p_{01}(C)) + (1-2\eta)p_{01}(C) \\ &= \eta p_0(C) + (1-2\eta)p_{01}(C). \end{aligned}$$

If $p_0(C) \neq 0$ then

$$\frac{p_{0+}(C)}{p_0(C)} = \eta + \frac{p_{01}(C)}{p_0(C)}(1-2\eta).$$

Since $\eta < 1/2$, this quantity is always greater than or equal to η and is equal to η if C is contained in or implied by ϕ_*. Since $p_0(C) \leq 1$, for all clauses C such that $p_0(C) \neq 0$,

$$\frac{p_{0+}(C)}{p_0(C)} \geq \eta + p_{01}(C)(1-2\eta).$$

If C is a harmful clause, then $p_{01}(C) \geq Q_H$, so

$$\frac{p_{0+}(C)}{p_0(C)} \geq \eta + Q_H(1-2\eta).$$

The quantity $p_{0+}(C)/p_0(C)$ is the proportion of those assignments falsifying C that are reported with $s = +$. The preceding calculation shows that there is a separation of at least $Q_H(1-2\eta)$ in the expected value of this quantity between clauses that are to be retained (important clauses in ϕ_*) and clauses that are to be discarded (harmful clauses). Since η is unknown, we replace this separation by a (known) lower bound

$$s = Q_H(1-2\eta_b).$$

Moreover, $p_{0+}(C)/p_0(C)$ can be estimated by sampling the oracle $EX_\eta()$. (Recall that I contains clauses to which a nontrivial number of samples assign $s = 0$, so for elements of I this estimate will be sufficiently accurate.)

The procedure V' calculates an estimate η' of η and identifies as harmful all those clauses $C \in I$ whose estimated value of $p_{0+}(C)/p_0(C)$ is greater than $\eta' + s/2$. The final output is the product of all the other clauses in I. In order for this to work, V' needs a sufficiently accurate estimate η' for η. Where does this come from? If I contains any clause C in or implied by ϕ_*, then the estimate of $p_{0+}(C)/p_0(C)$ will be close to η. In this case, the minimum estimate of $p_{0+}(C)/p_0(C)$ for all clauses C in I will be close to η.

However, it may happen that no clause in I is contained in or implied by ϕ_*, and this minimum value may not be a good estimate of η. In this case, provided all the important clauses are in I, we know that ϕ_* does not contain any important

clauses. This means that almost all assignments drawn from D assign the value 1 to ϕ_*. In this case, the observed overall rate of negative examples will be sufficiently close to η. Thus, the estimate of η is taken to be the minimum of two estimates: the estimated fraction of negative examples and the minimum estimated value of $p_{0+}(C)/p_0(C)$ over all clauses C in I. We now summarize the description of V'.

4.2 Concise description of V'

From n, k, ϵ, δ, and η_b, the procedure V' calculates the following:

$$\begin{aligned}
\mathcal{C} &= \{C : C \text{ is a clause over } n \text{ variables with at most } k \text{ literals}\},\\
M &= |\mathcal{C}|,\\
K &= 3 \cdot 2^{17},\\
m &= \left\lceil \frac{KM^6}{\epsilon^3(1-2\eta_b)^3} \ln \frac{6M}{\delta} \right\rceil,\\
Q_H &= \epsilon/2M,\\
s &= Q_H(1-2\eta_b) = \epsilon(1-2\eta_b)/2M,\\
Q_I &= s/8M = \epsilon(1-2\eta_b)/16M^2.
\end{aligned}$$

V' draws m samples from the oracle $EX_\eta()$, say

$$\sigma = \langle a_1, s_1\rangle, \ldots, \langle a_m, s_m\rangle,$$

where each a_i is a truth-value assignment to the variables $x_1, \ldots, x_n$ and each s_i is either $+$ or $-$.

The following quantities are defined using σ:

$$\begin{aligned}
Z_- &= |\{j : s_j = -\}|,\\
Z_0(C) &= |\{j : a_j(C) = 0\}|,\\
Z_{0+}(C) &= |\{j : a_j(C) = 0 \text{ and } s_j = +\}|.
\end{aligned}$$

Z_- is the overall number of negative samples, $Z_0(C)$ is the number of samples that assign 0 to the clause C, and $Z_{0+}(C)$ is the number of samples that assign 0 to C and are reported with the sign $+$. For each clause C in $\mathcal{C}$ such that $Z_0(C) \neq 0$, define

$$h(C) = Z_{0+}(C)/Z_0(C).$$

$h(C)$ is the estimated value of the quantity $p_{0+}(C)/p_0(C)$.

The procedure V' calculates one estimate of η:

$$\eta_1 = Z_-/m,$$

which is just the observed fraction of negative examples.

The procedure V' then forms the set I by including all those clauses C in $\mathcal{C}$ such that

$$Z_0(C)/m \geq Q_I/2.$$

If I is non-empty then V' calculates a second estimate of η as follows:

$$\eta_2 = \min\{h(C) : C \in I\}.$$

If I is empty then $\eta_2 = +\infty$.

V' then calculates

$$\eta' = \min\{\eta_1, \eta_2\}.$$

The final output ϕ of V' is the product of all those clauses $C \in I$ such that

$$h(C) \leq \eta' + s/2.$$

It is clear from this description that V' runs in time polynomial in n^k, $1/\epsilon$, $\log 1/\delta$, and $1/(1 - 2\eta_b)$. In the next section we show that it achieves *pac*-identification of the formulas in $CNF(n,k)$.

4.3 Proof of correctness of V'

Theorem 5 *For every $\phi_* \in CNF(n,k)$, V' pac-identifies ϕ_*, that is,*

$$\Pr[d(\phi, \phi_*) \leq \epsilon] \geq 1 - \delta.$$

The proof proceeds by showing that, with high probability, the set I contains all the important clauses, and given that I contains all the important clauses, η' is a good estimate of η with high probability, and finally, given that I contains all the important clauses and η' is a good estimate of η, all the harmful clauses are excluded from ϕ and all of the important clauses in ϕ_* are retained in ϕ with high probability. The net effect is to show that with high probability, the output ϕ contains all the important clauses in ϕ_* and no harmful clauses. Applying Lemma 4, we conclude that with high probability, $d(\phi, \phi_*) < \epsilon$.

Lemma 6 *With high probability the set I contains all the important clauses, i.e.,*

$$\Pr[I \textit{ excludes some important clause}] \leq \delta/6.$$

Proof: Consider any important clause C. By definition, $p_0(C) \geq Q_I$. Each time an assignment is drawn from D, there is a probability of $p_0(C)$ that it assigns 0 to C. The probability that in m assignments drawn from D the fraction that assign 0 to C is less than or equal to $Q_I/2$ is at most $LE(Q_I, m, Q_I/2)$. That is,

$$\Pr[C \text{ is not included in } I] \leq LE(Q_I, m, Q_I/2).$$

By Lemma 13 in the Appendix,

$$LE(Q_I, m, Q_I/2) \leq \delta/6M.$$

Summing over the (at most M) important clauses, we find that

$$\Pr[I \text{ excludes some important clause }] \leq \delta/6.$$

□

Lemma 7 *With high probability η_1 is not "too small", that is,*

$$\Pr[\eta_1 \leq \eta - s/4] \leq \delta/6.$$

If ϕ_* does not contain any important clause then with high probability, η_1 is not "too big", that is,

$$\Pr[\eta_1 \geq \eta + s/4] \leq \delta/6.$$

Proof: We analyze the expected rate of negative samples from $EX_\eta()$ for any formula $\phi_* \in CNF(n,k)$. Let

$$\begin{aligned} p_0(\phi_*) &= \Pr[a(\phi_*) = 0], \\ p_-(\phi_*) &= \Pr[\text{a sample } \langle a, s\rangle \text{ drawn from } EX_\eta() \text{ has } s = -]. \end{aligned}$$

An assignment a chosen by $EX_\eta()$ is reported as a negative example if and only if either $a(\phi_*) = 0$ and there is no error of reporting or $a(\phi_*) = 1$ and there is an error of reporting, so

$$p_-(\phi_*) = (1-\eta)p_0(\phi_*) + \eta(1 - p_0(\phi_*)).$$

We rewrite this as

$$p_-(\phi_*) = \eta + p_0(\phi_*)(1 - 2\eta).$$

Since $\eta < 1/2$, $p_-(\phi_*) \geq \eta$ for all formulas $\phi_* \in CNF(n,k)$. Moreover, since $(1-2\eta)$ is at most 1,

$$p_-(\phi_*) \leq \eta + p_0(\phi_*).$$

To establish the first part of the lemma, we note

$$\Pr[\eta_1 \leq \eta - s/4] \leq LE(\eta, m, \eta - s/4).$$

By Lemma 13 in the Appendix,

$$LE(\eta, m, \eta - s/4) \leq \delta/6.$$

To prove the second part of the lemma, assume that ϕ_* is any element of $CNF(n,k)$ that contains no important clauses. Then

$$\begin{aligned} p_0(\phi_*) &\leq \sum_{C \in \phi_*} p_0(C), \\ &< \sum_{C \in \phi_*} Q_I, \text{ since no clause in } \phi_* \text{ is important,} \\ &< MQ_I, \\ &< s/8. \end{aligned}$$

Hence

$$p_-(\phi_*) < \eta + s/8.$$

Then

$$\Pr[\eta_1 \geq \eta + s/4] \leq GE(\eta + s/8, m, \eta + s/4),$$

and by Lemma 13 in the Appendix,

$$GE(\eta + s/8, m, \eta + s/4) \leq \delta/6.$$

□

Lemma 8 *With high probability, η_2 is not "too small", that is,*

$$\Pr[\eta_2 \leq \eta - s/4] \leq \delta/6.$$

If ϕ_* contains some important clause, then with high probability, η_2 is not "too large", that is,

$$\Pr[\eta_2 \geq \eta + s/4 \mid I \textit{ contains every important clause}] \leq \delta/6.$$

Proof: If I is empty then $\eta_2 = +\infty$ and the first part of the lemma is true.

So, assume I is nonempty and let C be any clause in I. The probability that a sample $\langle a, s \rangle$ reported by $EX_\eta()$ will have $s = +$ given that $a(C) = 0$ is just

$p_{0+}(C)/p_0(C)$. This is the expected value of the observed quantity $h(C)$. Since C is in I, we have at least $\lceil mQ_I/2 \rceil$ independent trials of this kind.

We have shown that $p_{0+}(C)/p_0(C)$ is greater than or equal to η, so

$$\Pr[h(C) \leq \eta - s/4] \leq LE(\eta, \lceil mQ_I/2 \rceil, \eta - s/4).$$

By Lemma 13 in the Appendix,

$$LE(\eta, \lceil mQ_I/2 \rceil, \eta - s/4) \leq \delta/6M.$$

Summing over the (at most M) elements of I,

$$\Pr[\text{for some } C \in I, h(C) \leq \eta - s/4] \leq \delta/6.$$

Thus in either case,

$$\Pr[\eta_2 \leq \eta - s/4] \leq \delta/6,$$

proving the first part of the lemma.

For the second part of the lemma, assume that D and ϕ_* are such that ϕ_* contains at least one important clause. Assume also that the sampling of $EX_\eta()$ produces the outcome that I contains all the important clauses, so in particular, it will contain at least one important clause from ϕ_*, say C_*. We have shown that in this case, $p_{0+}(C_*)/p_0(C_*) = \eta$, so

$$\Pr[h(C_*) \geq \eta + s/4] \leq GE(\eta, \lceil mQ_I/2 \rceil, \eta + s/4),$$

and by Lemma 13 in the Appendix,

$$GE(\eta, \lceil mQ_I/2 \rceil, \eta + s/4) \leq \delta/6.$$

Since in this case η_2 is at most $h(C_*)$, we conclude that when ϕ_* contains some important clause,

$$\Pr[\eta_2 \geq \eta + s/4 \mid I \text{ contains every important clause}] \leq \delta/6.$$

□

Lemma 9 *With high probability, every harmful clause C in I will have a "large" value of $h(C)$, that is,*

$$\Pr[\textit{for some harmful } C \in I, h(C) \leq \eta + 3s/4] \leq \delta/6.$$

With high probability, every important clause in both ϕ_* and I will have a "small" value of $h(C)$, that is,

$$\Pr[\textit{for some important } C \in I \cap \phi_*, h(C) \geq \eta + s/4] \leq \delta/6.$$

Proof: Suppose C is a harmful clause in I. We have shown that since C is harmful,

$$p_{0+}(C)/p_0(C) \geq \eta + s.$$

Thus,

$$\Pr[h(C) \leq \eta + 3s/4] \leq LE(\eta + s, \lceil mQ_I/2 \rceil, \eta + 3s/4),$$

and by Lemma 13 in the Appendix,

$$LE(\eta + s, \lceil mQ_I/2 \rceil, \eta + 3s/4) \leq \delta/6M.$$

Summing over the (at most M) harmful clauses in I,

$$\Pr[\text{for some harmful } C \in I, h(C) \leq \eta + 3s/4] \leq \delta/6.$$

Suppose C is an important clause in I and in ϕ_*. Then since

$$p_{0+}(C)/p_0(C) = \eta,$$

we have

$$\Pr[h(C) \geq \eta + s/4] \leq GE(\eta, \lceil mQ_I/2 \rceil, \eta + s/4),$$

and by Lemma 13 in the Appendix,

$$GE(\eta, \lceil mQ_I/2 \rceil, \eta + s/4) \leq \delta/6M.$$

Summing over the (at most M) important clauses in $I \cap \phi_*$, we obtain

$$\Pr[\text{for some important } C \in I \cap \phi_*, h(C) \geq \eta + s/4] \leq \delta/6.$$

□

Now we have all the pieces, and may conclude the proof of Theorem 5. We analyze the probability that V' fails, that is, produces an output ϕ such that $d(\phi, \phi_*) > \epsilon$. We analyze two cases separately: whether or not ϕ_* contains some important clause.

Assume ϕ_* contains no important clause. V' can fail only if at least one of the following events occurs.

1. $\eta_1 \geq \eta + s/4$.
2. For some harmful $C \in I$, $h(C) \leq \eta + 3s/4$.

To see this suppose that neither of these events occurs. Then the value of η' is less than $\eta + s/4$, so the value of $\eta' + s/2$ is less than $\eta + 3s/4$. Also, for every harmful clause $C \in I$, $h(C)$ is greater than $\eta + 3s/4$, so no harmful clause will be included in ϕ. Since there are no important clauses in ϕ_* by the assumption of this case, ϕ will (vacuously) contain every important clause in ϕ_*. Applying Lemma 4, we conclude that $d(\phi, \phi_*) < \epsilon$.

By Lemma 7, the probability of event (1) above is at most $\delta/6$, and by Lemma 9, the probability of event (2) above is at most $\delta/6$, so in this case, the probability of failure of V' is at most $\delta/3$.

For the second case, assume that ϕ_* does contain at least one important clause. Then V' can fail only if at least one of the following events occurs.

1. Not every important clause is included in I.
2. $\eta_1 \leq \eta - s/4$.
3. $\eta_2 \leq \eta - s/4$.
4. $\eta_2 \geq \eta + s/4$, given that I includes every important clause.
5. For some harmful $C \in I$, $h(C) \leq \eta + 3s/4$.
6. For some important clause $C \in \phi_* \cap I$, $h(C) \geq \eta + s/4$.

To see this, suppose none of the above events occurs. Then every important clause is included in I, η' is strictly between $\eta - s/4$ and $\eta + s/4$, so $\eta' + s/2$ is strictly between $\eta + s/4$ and $\eta + 3s/4$. Moreover, for every harmful clause C in I, $h(C) > \eta + 3s/4$, so no harmful clause is included in ϕ. Also, for every important clause in $\phi_* \cap I$ (which includes every important clause in ϕ_*), $h(C) < \eta + s/4$, so every important clause in ϕ_* is included in ϕ. Applying Lemma 4, we conclude that $d(\phi, \phi_*) < \epsilon$.

By Lemma 6, event (1) has probability at most $\delta/6$, by Lemma 7, event (2) has probability at most $\delta/6$, by Lemma 8, events (3) and (4) each have probability at most $\delta/6$, and by Lemma 9, events (5) and (6) each have probability at most $\delta/6$, so the probability that V' fails at most δ.

Thus in either case, the probability that V' produces an output ϕ such that $d(\phi, \phi_*) \leq \epsilon$ is at least $1 - \delta$, which proves Theorem 5.

5 Remarks

Because of the duality of conjunctive and disjunctive normal forms, k-DNF formulas can be *pac*-identified from noisy examples by the dual of the procedure V'.

The procedure V' does not depend very strongly on the properties of conjunctive normal form, and should generalize to somewhat weaker notions of normal form. Proving similar results for more general error models would be quite interesting.

We have made no particular attempt to minimize the number of samples used by the procedure V', so there is undoubtedly room for improvement in that aspect of the algorithm. A lot of work would likely be necessary to "tune" the procedure V' for practical use; such an effort might well pay off in improved approaches to the problem.

It would be interesting to explore the effect of errors in a situation which calls for queries as well as random sampling. For example, could the polynomial time procedure to identify regular sets given a sampling oracle and membership queries in [1] be modified to compensate for random errors in the sampling and query responses? Another interesting direction is to attempt to incorporate errors into the general "refinement" approach to inference [4].

6 Appendix: Bounding lemmas

We establish some simple tools for bounding the accuracy of estimates of Bernoulli variables. For p and r between 0 and 1 and any positive integer m, let $LE(p,m,r)$ denote the probability of at most rm successes in m independent trials of a Bernoulli variable with probability of success p, and $GE(p,m,r)$ the probability of at least rm successes. Thus,

$$GE(p,m,r) = \sum_{k=\lceil rm \rceil}^{m} \binom{m}{k} p^k (1-p)^{m-k},$$

and

$$LE(p,m,r) = \sum_{k=0}^{\lfloor rm \rfloor} \binom{m}{k} p^k (1-p)^{m-k}.$$

It is not difficult to show that for p increasing, $GE(p,m,r)$ is nondecreasing and $LE(p,m,r)$ is nonincreasing. We extend LE to have the value 0 if its third argument is less than 0, and similarly GE has the value 0 if its third argument is greater than 1.

The basic lemma we use is from [2]:

Lemma 10 *If* $0 \leq p \leq 1, 0 \leq \beta \leq 1$, *and* m *is any positive integer then*

$$LE(p,m,(1-\beta)p) \leq e^{-\beta^2 mp/2},$$

and

$$GE(p, m, (1+\beta)p) \leq e^{-\beta^2 mp/3}.$$

We apply this to obtain a simple bound on the number of samples required to assure that an estimate of p is within a distance s of the correct value with probability at least $1-\delta$.

Lemma 11 *Let* $0 \leq p \leq 1$, $0 < s < 1$, *and* $0 < \delta < 1$. *If*

$$m \geq \frac{12}{s^2} \ln \frac{1}{\delta}$$

then

$$LE(p, m, p-s) \leq \delta,$$

and

$$GE(p, m, p+s) \leq \delta.$$

Proof: To prove the first inequality, we consider two cases. Assume that $p < s$. Then $p - s < 0$, so $LE(p, n, p-s) = 0$, which is certainly less than or equal to δ. Assume that $p \geq s$, and let

$$\beta = s/p.$$

Then $0 \leq \beta \leq 1$ and

$$p - s = (1-\beta)p.$$

Applying Lemma 10,

$$\begin{aligned} LE(p, m, p-s) &\leq e^{-(s/p)^2 mp/2}, \\ &\leq e^{-s^2 m/2p}, \\ &\leq e^{-s^2 m/2}, \text{ since } p \leq 1, \\ &\leq \delta, \text{ since } s^2 m/2 \geq 6\ln(1/\delta). \end{aligned}$$

For the second inequality, we also argue in two cases. Assume that $p \leq s/2$. Then

$$\begin{aligned} GE(p, m, p+s) &\leq GE(s/2, m, p+s), \text{ by the monotonicity of } GE \text{ in } p, \\ &\leq GE(s/2, m, s), \\ &\leq e^{-sm/6}, \text{ applying Lemma 10 with } \beta = 1, \\ &\leq \delta, \text{ since } sm/6 \geq (2/s)\ln(1/\delta). \end{aligned}$$

Assume that $p > s/2$. Let

$$\beta = s/2p,$$

note that $\beta < 1$, and

$$p + s/2 = (1 + \beta)p.$$

Applying Lemma 10,

$$\begin{aligned} GE(p, m, p + s) &\leq GE(p, m, p + s/2), \\ &\leq e^{-(s/2p)^2 mp/3}, \\ &\leq e^{-s^2 m/12p}, \\ &\leq e^{-s^2 m/12}, \text{ since } p \leq 1, \\ &\leq \delta, \text{ since } s^2 m/12 \geq \ln(1/\delta). \end{aligned}$$

This concludes the proof of Lemma 11. □

For the various bounds in the paper, we require several applications of this basic lemma, which are here summarized and proved.

Lemma 12 *Let N be a positive integer, $0 < \epsilon < 1$, $0 < \delta < 1$, and $0 \leq \eta \leq \eta_b < 1/2$. Define*

$$s = \epsilon(1 - 2\eta_b).$$

Then $0 < s < 1$. If

$$m \geq \frac{48}{\epsilon^2(1 - 2\eta_b)^2} \ln \frac{2N}{\delta},$$

then

$$GE(\eta, m, \eta + s/2) \leq \delta/2N,$$

and

$$LE(\eta + s, m, \eta + s/2) \leq \delta/2N.$$

Proof: We apply Lemma 11 with $s/2$ in place of s and $\delta/2N$ in place of δ to find the indicated lower bound on m. □

Lemma 13 *Let M be a positive integer, $0 < \epsilon < 1$, $0 < \delta < 1$, and $0 \leq \eta \leq \eta_b < 1/2$. Define*

$$s = \epsilon(1 - 2\eta_b)/2M,$$

and

$$Q_I = s/8M = \epsilon(1 - 2\eta_b)/16M^2.$$

Then $s/8 > Q_I/2$. If

$$m \geq \frac{KM^6}{\epsilon^3(1 - 2\eta_b)^3} \ln \frac{6M}{\delta},$$

where $K = 3 \cdot 2^{17}$, *then for any* p, t, *and* x *such that* $0 \leq p \leq 1$, $t \geq \lceil mQ_I/2 \rceil$, *and* $Q_I/2 \leq x < 1$, *we have*

$$LE(p, t, p - x) \leq \delta/6M,$$

and

$$GE(p, t, p + x) \leq \delta/6M.$$

Proof: This lemma is proved by applying Lemma 11 with $Q_I/2$ for s, $\delta/6M$ for δ, and $\lceil mQ_I/2 \rceil$ for m to obtain the indicated bound. It suffices if

$$mQ_I/2 \geq \frac{12}{(Q_I/2)^2} \ln \frac{6M}{\delta},$$

so it suffices if

$$m \geq \frac{96}{Q_I^3} \ln \frac{6M}{\delta}.$$

Substituting in the value of Q_I, we obtain

$$m \geq \frac{KM^6}{\epsilon^3(1 - 2\eta_b)^3} \ln \frac{6M}{\delta}.$$

for $K = 3 \cdot 2^{17}$. □

References

[1] D. Angluin. *Learning regular sets from queries and counter-examples.* Technical Report, Yale University Computer Science Dept. TR-464, 1986.

[2] D. Angluin and L. G. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. *J. Comput. Syst. Sci.*, 18:155–193, 1979.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. In *Proc. 18th Symposium on Theory of Computing*, pages 273–282, ACM, 1986.

[4] P. D. Laird. *Inductive inference by refinement.* Technical Report, Yale University Computer Science Dept. TR-376 (revised), 1986.

[5] L. G. Valiant. Learning disjunctions of conjunctions. In *Proceedings of IJCAI*, pages 560–566, IJCAI, 1985.

[6] L. G. Valiant. A theory of the learnable. *C. ACM*, 27:1134–1142, 1984.