Annals of Mathematics and Artificial Intelligence 23 (1998) 83–99

On-line learning with malicious noise and the closure algorithm

Peter Auer [a] and Nicolò Cesa-Bianchi [b]

[a] *IGI, Graz University of Technology, Klosterwiesgasse 32/2, A-8010 Graz, Austria*
E-mail: pauer@igi.tu-graz.ac.at

[b] *DSI, University of Milan, Via Comelico 39, I-20135 Milano, Italy*
E-mail: cesabian@dsi.unimi.it

We investigate a variant of the on-line learning model for classes of $\{0,1\}$-valued functions (concepts) in which the labels of a certain amount of the input instances are corrupted by adversarial noise. We propose an extension of a general learning strategy, known as "Closure Algorithm", to this noise model, and show a worst-case mistake bound of $m+(d+1)K$ for learning an arbitrary intersection-closed concept class $\mathcal{C}$, where K is the number of noisy labels, d is a combinatorial parameter measuring $\mathcal{C}$'s complexity, and m is the worst-case mistake bound of the Closure Algorithm for learning $\mathcal{C}$ in the noise-free model. For several concept classes our extended Closure Algorithm is efficient and can tolerate a noise rate up to the information-theoretic upper bound. Finally, we show how to efficiently turn any algorithm for the on-line noise model into a learning algorithm for the PAC model with malicious noise.

1. Introduction

In the on-line learning model introduced in [1,15] a learner has to identify a target chosen from a given class of concepts (i.e., subsets of a fixed set X) by seeing a sequence of labeled instances (i.e., elements of $X \times \{0,1\}$). Each instance is labeled according to whether it belongs or not to the target and the learner must exhibit some hypothesized target concept before seeing the next labeled instance. To evaluate a learner we look at the worst-case number of times (over all choices of targets and instance sequences) the current hypothesis misclassified the next instance.

In this paper we investigate an extension of the above framework to take into account the presence of adversarial noise. Namely, an adversary is allowed to choose the labels of a certain amount of the instances in the sequence presented to the learner. The learner's goal is to minimize the worst-case number of mistakes made over all noisy sequences. This approach can be compared to the ideas and results contained in [4,7,17] where general (nonefficient) "conversion strategies" to make an on-line learning algorithm robust to adversarial noise were proposed.

We consider a very general on-line strategy known as "Closure Algorithm" [10, 12,13,19] for learning intersection-closed concept classes in the noise-free model. We extend this strategy to our noisy learning setting and show a worst-case mistake bound

of $m+(d+1)K$ for learning an arbitrary intersection-closed concept class $\mathcal{C}$, where K is the number of noisy labels, d is a combinatorial parameter measuring $\mathcal{C}$'s complexity,[1] and m is the worst-case mistake bound of the Closure Algorithm for learning $\mathcal{C}$ in the noise-free model.

For several concept classes our extension is efficient and in some cases can tolerate a noise rate up to the information-theoretic upper bound for that class. Using results from [9,13] we show that the classes of monotone monomials, k-CNF functions, parity functions, integer lattices, conjunctions of counting functions, and k-ring-sum expansions can be efficiently learned on-line with adversarial noise. We also propose a general technique for showing upper bounds on the noise rate tolerated by any on-line learner disregarding computational constraints. This technique is applied to the classes of subspaces of a linear space, halfspaces in $\{0,1\}^n$, and to most of the above-mentioned classes.

Finally suppose that, for some positive m_0 and R, and for all integers $K \geqslant 0$, an on-line algorithm A makes at most $m_0 + RK$ mistakes for learning a concept class $\mathcal{C}$ using hypotheses from $\mathcal{H}$ when at most K labels are noisy. We then show that A can be efficiently turned into an algorithm for learning $\mathcal{C}$ by $\mathcal{H}$ in the malicious PAC model [14] with any accuracy ε and noise rate $\varepsilon/R - \alpha$ for any $\alpha > 0$.

2. Notation, terminology, and basic facts

Fix an arbitrary set X (the *instance domain*). A *concept class* over X is any collection of subsets of X. If C is a subset of X we will use the same symbol C also to denote the characteristic function of the subset. For any concept class $\mathcal{C}$ let $\overline{\mathcal{C}}$ the class of the complements $\overline{C}$ for all $C \in \mathcal{C}$.

Following Littlestone [15] we define the on-line learning process by a sequence of trials. On each trial the learner outputs a current hypothesis $H \in \mathcal{H}$ from some fixed concept class $\mathcal{H}$ (the *hypothesis class*). Afterwards, the next labeled instance $(x, C(x))$ is revealed, where $x \in X$ and C is some fixed target concept from the *target class* $\mathcal{C}$. The boolean label $C(x)$ is 1 if and only if x belongs to C. The learner makes a mistake on the trial if $H(x) \neq C(x)$, in this case we say that the instance x is a *counterexample* to hypothesis H. The counterexample x is *positive* if $C(x) = 1$ and *negative* otherwise. A "learner" in this model is thus defined by a mapping from finite sequences (possibly of zero length) of labeled instances to hypotheses $H \in \mathcal{H}$. In general, the mapping defining a learner needs not to be computable. When only computable mappings are considered we will use the term learning algorithm instead of learner.

Let $\mathcal{H}$ be the hypothesis class of learner A. We write $\mathrm{MB}(A, \mathcal{C}, \mathcal{H})$ to denote the worst-case number of mistakes made by A over all choices of the target $C \in \mathcal{C}$ and over all trial sequences labeled by C. Finally, let $\mathrm{MB}(\mathcal{C}, \mathcal{H})$ denote the minimum of

[1] For a particular implementation of our algorithm this combinatorial parameter is bounded from above by the VC dimension of $\mathcal{C}$.

MB($A, \mathcal{C}, \mathcal{H}$) over all learners A using hypothesis class $\mathcal{H}$. If $\mathcal{H} \equiv \mathcal{C}$, then we use the abbreviations MB($A, \mathcal{C}$) and MB($\mathcal{C}$).

A closely related on-line learning model was independently introduced by Angluin [1]. In this setting the learner receives *on each trial* a counterexample $x \in X$ to the current hypothesis H such that $H(x) \neq C(x)$. The learning process ends as soon as the learner's hypothesis H satisfies $H \equiv C$.

For any on-line learner A using hypothesis class $\mathcal{H}$, EQ($A, \mathcal{C}, \mathcal{H}$) is defined by the maximal length of a sequence of counterexamples received by A when the target is chosen from $\mathcal{C}$. Accordingly, EQ($\mathcal{C}, \mathcal{H}$) is the minimum of EQ($A, \mathcal{C}, \mathcal{H}$) over all learners A.

The following result (proven in [16]) relates Littlestone's MB model to Angluin's EQ model.

Fact 2.1. Any on-line learner A in the MB model is an on-line learner in the EQ model. Vice versa, any on-line learner A' in the EQ model is a conservative[2] learner in the MB model. Moreover, EQ($A, \mathcal{C}, \mathcal{H}$) $\leqslant$ MB($A, \mathcal{C}, \mathcal{H}$) and MB($A', \mathcal{C}, \mathcal{H}'$) $=$ EQ($A', \mathcal{C}, \mathcal{H}'$) for all target classes $\mathcal{C}$.

We now consider the following extensions to the MB and EQ models taking into account the presence of adversarial noise in the learning process. These extensions were respectively introduced in [17] and [3].

Again assume $\mathcal{H}$ is the hypothesis class of learner A. For any nonnegative integer K, let MB($A, \mathcal{C}, \mathcal{H}, K$) be the worst-case number of mistakes made by A over all sequences $(x_1, \ell_1), (x_2, \ell_2), \ldots$ of labeled instances such that there is some $C \in \mathcal{C}$ for which $C(x_t) \neq \ell_t$ holds for at most K indices t in the sequence. (In this model the learner makes a mistake if it predicts the next label incorrectly, i.e., if $H(x_t) \neq \ell_t$.) The quantity MB($\mathcal{C}, \mathcal{H}, K$) is defined analogously to MB($\mathcal{C}, \mathcal{H}$) before.

For any *noise rate* $0 \leqslant r < 1$ define EQ($A, \mathcal{C}, \mathcal{H}, r$) as the maximal length of a sequence of counterexamples received by A such that there is some $C \in \mathcal{C}$ for which $C(x_t) \neq \ell_t$ holds for at most a fraction r of the counterexamples in the sequence. The quantity EQ($\mathcal{C}, \mathcal{H}, r$) is defined as before. Observe that in the EQ model we measure the amount of noise by a relative noise rate while in the MB model we count the absolute number of noisy counterexamples.

The next result extends fact 2.1 by showing the relationships between the EQ model and the MB model in presence of adversarial noise.

Fact 2.2. Let A be an on-line learner with hypothesis class $\mathcal{H}$.

1. If MB($A, \mathcal{C}, \mathcal{H}, K$) $\leqslant M + RK$ for some $M, R > 0$ then for all $r < 1/R$ and $m \geqslant m_0 = M/(1 - rR)$, MB($A, \mathcal{C}, \mathcal{H}, rm$) $\leqslant m$. Furthermore, EQ($A, \mathcal{C}, \mathcal{H}, r$) $\leqslant m_0$.

[2] We say that a learner is conservative if it changes its hypothesis only when a mistake occurs.

2. If $\mathrm{EQ}(A, \mathcal{C}, \mathcal{H}, r) = m_0$ then there is an on-line learner A' with $\mathrm{MB}(A', \mathcal{C}, \mathcal{H}, K) \leqslant m_0 + RK$, where $R = (1 + 1/m_0)/r$.

Proof. For proving part 1 we have

$$\mathrm{MB}(A, \mathcal{C}, \mathcal{H}, rm) \leqslant M + Rrm = (1 - rR)m_0 + Rrm < m$$

for all $m > m_0$. Now assume that $\mathrm{EQ}(A, \mathcal{C}, \mathcal{H}, r) > m_0$. Then there is a sequence of counterexamples to the hypotheses of A of length $m > m_0$ such that at most rm of the counterexamples are noisy, contradicting $\mathrm{MB}(A, \mathcal{C}, \mathcal{H}, rm) < m$.

For proving part 2 let A' be the learner which runs A as subroutine until A makes $m_0 + 1$ mistakes. Then A' restarts A and runs A until it again makes $m_0 + 1$ mistakes. This continues for the whole sequence of trials. Observe that among the $m_0 + 1$ mistakes of one run of A there are at least $\lfloor rm_0 + 1 \rfloor$ noisy trials since $\mathrm{EQ}(A, \mathcal{C}, \mathcal{H}, r) = m_0$. Hence there are at most $K/\lfloor rm_0 + 1 \rfloor + 1$ runs of A where the last run makes at most m_0 mistakes, thus giving

$$\mathrm{MB}(A, \mathcal{C}, \mathcal{H}, K) \leqslant \frac{K}{\lfloor rm_0 + 1 \rfloor}(m_0 + 1) + m_0 \leqslant \frac{K}{rm_0}(m_0 + 1) + m_0$$

and concluding the proof. □

We close the section with some further definitions and notation. We use **N** to denote the nonnegative integers and **Z** to denote the integers. If S is an arbitrary set, P a distribution over S, and R a random variable over S, then the expectation of R with respect to P is denoted by $\mathbf{E}_{s \sim P}[R(s)]$. Finally, let log be the base 2 logarithm.

3. An extension of the Closure Algorithm

We begin by showing that whenever a target class is noise-free on-line learnable (i.e., on-line learnable with noise rate 0), then there exists a general (nonefficient) strategy such that $\mathcal{C}$ is on-line learnable for any noise rate $r < 1/2$.

Theorem 3.1. Fix a target class $\mathcal{C}$. Then for any concept class $\mathcal{H}$ such that $\mathrm{EQ}(\mathcal{C}, \mathcal{H}, 0) > 0$ and any $0 \leqslant r < 1/2$, there is an algorithm A that yields

$$\mathrm{EQ}\big(A, \mathcal{C}, 2^X, r\big) \leqslant 2\frac{\mathrm{EQ}(\mathcal{C}, \mathcal{H}, 0)}{1 - H(r)} \log \frac{2\,\mathrm{e} \cdot \mathrm{EQ}(\mathcal{C}, \mathcal{H}, 0)}{1 - H(r)},$$

where H is the binary entropy function

$$H(x) = -x \log(x) - (1 - x) \log(1 - x).$$

Proof. From [7, theorem 5], we know that, for any $m_0 = \mathrm{MB}(\mathcal{C}, \mathcal{H}, 0)$ and any $K \geqslant 0$, there exists a conservative on-line learner A such that

$$\mathrm{MB}\left(A, \mathcal{C}, 2^X, K\right) \leqslant \max\left\{q \in \mathbf{N}\colon\ q \leqslant \log \sum_{i=0}^{K} \binom{q}{i} + \log \sum_{j=0}^{m_0} \binom{q}{j}\right\}.$$

By using $\sum_{i=0}^{K} \binom{q}{i} \leqslant 2^{qH(K/q)}$ and $\sum_{j=0}^{m_0} \binom{q}{j} \leqslant (q\,\mathrm{e}/m_0)^{m_0}$, we get

$$\mathrm{MB}\left(A, \mathcal{C}, 2^X, K\right) \leqslant \max\left\{q \in \mathbf{N}\colon\ q \leqslant qH(K/q) + m_0 \log(q\,\mathrm{e}/m_0)\right\}.$$

If we run A (which is conservative) in the EQ model while assuming $K/q \leqslant r$, we find that

$$\begin{aligned}\mathrm{EQ}\left(A, \mathcal{C}, 2^X, r\right) &\leqslant \max\left\{q \in \mathbf{N}\colon\ q \leqslant qH(r) + m_0 \log(q\,\mathrm{e}/m_0)\right\} \\ &= \max\left\{q \in \mathbf{N}\colon\ q \leqslant \frac{m_0 \log(q\,\mathrm{e}/m_0)}{1 - H(r)}\right\}.\end{aligned}$$

It is then easy to verify that

$$q > 2\frac{m_0}{1 - H(r)} \log \frac{2\,\mathrm{e}m_0}{1 - H(r)}$$

implies

$$q > \frac{m_0 \log(q\,\mathrm{e}/m_0)}{1 - H(r)},$$

thus proving the theorem. □

We now move on to the description of the Closure Algorithm and its extension to the noisy on-line learning model. Some preliminary definitions are needed.

The *closure* operator $\mathrm{Cl}_{\mathcal{C}} : 2^X \to 2^X$ is defined by the formula

$$\mathrm{Cl}_{\mathcal{C}}(S) = \bigcap_{\{C \in \mathcal{C}\colon\ S \subseteq C\}} C.$$

(If $\{C \in \mathcal{C}\colon\ S \subseteq C\} = \emptyset$ then $\mathrm{Cl}_{\mathcal{C}}(S) = X$.)

Notice that, if $\mathcal{C}$ is the class of all subspaces of a linear space X, then the closure operator $\mathrm{Cl}_{\mathcal{C}}(S)$ returns the subspace spanned by $S \subseteq X$.

A concept class $\mathcal{C}$ on domain X is *intersection-closed* if for all finite $S \subseteq X$, $\mathrm{Cl}_{\mathcal{C}}(S) \in \mathcal{C}$. In other words, the intersection-closedness property holds whenever the intersection of all concepts in $\mathcal{C}$ containing an arbitrary subset of the domain belongs to $\mathcal{C}$ as well.

Examples of intersection-closed concept classes include: axis-parallel n-dimensional rectangles, k-CNF boolean functions, subspaces of a linear space, integer lattices. However, notice that any concept class can be made intersection-closed by adding the set of all intersections of concepts in the class. The Closure Algorithm CA (sketched in figure 1) simply hypothesizes the closure of the set of all counterexamples received up

Algorithm CA.
Input: Hypothesis class $\mathcal{H}$.

- Initialize the state variable $S_0 := \emptyset$.
- **For** $t = 0, 1, \ldots$
  1. Let $H_t = \mathrm{Cl}_{\mathcal{H}}(S_t)$ be the current hypothesis.
  2. Read next labeled instance (x_{t+1}, ℓ_{t+1}).
  3. **If** $H_t(x_{t+1}) = 0$ **and** $\ell_{t+1} = 1$ **then** $S_{t+1} := S_t \cup \{x_{t+1}\}$.
     Else $S_{t+1} := S_t$.

Figure 1. A sketch of algorithm CA (the standard Closure Algorithm).

to the current trial. Due to the intersection-closedness property of the target class, the algorithm's hypotheses always are the smallest concepts consistent with all previously seen (positive) counterexamples, and thus in the noise-free case the Closure Algorithm will only receive positive counterexamples. For instance, let $\mathcal{C}$ be all subspaces of a d-dimensional linear space X. We then immediately have $\mathrm{MB}(\mathrm{CA}, \mathcal{C}) = d$, since the Closure Algorithm will receive only linearly independent counterexamples.

We now introduce a class of operators $\mathrm{Bas}_{\mathcal{C}}$ mapping subsets of X to subsets of X. A mapping $\mathrm{Bas}_{\mathcal{C}} : 2^X \to 2^X$ is a *basis operator* with respect to a concept class $\mathcal{C}$ if for all $S \subseteq X$ it holds that $\mathrm{Bas}_{\mathcal{C}}(S) \subseteq S$ and $\mathrm{Cl}_{\mathcal{C}}(\mathrm{Bas}_{\mathcal{C}}(S)) = \mathrm{Cl}_{\mathcal{C}}(S)$. (This definition of basis operator is analogous to that of *spanning set* for a set S as given in [10].) A trivial basis operator is the identity mapping. In the case $\mathcal{C}$ is the class of all subspaces of a linear space, a very natural basis operator maps each $S \subseteq X$ to a maximal subset $S' \subseteq S$ of linearly independent vectors.

We say that a basis operator $\mathrm{Bas}^*_{\mathcal{C}}$ is *minimal* if for all basis operators $\mathrm{Bas}_{\mathcal{C}}$ for $\mathcal{C}$ and for all $S \subseteq X$ it holds that $|\mathrm{Bas}^*_{\mathcal{C}}(S)| \leqslant |\mathrm{Bas}_{\mathcal{C}}(S)|$. Minimal basis operators enjoy the following property.

Lemma 3.2 [5,18]. For all intersection-closed concept classes $\mathcal{C}$ on a set X, if $\mathrm{Bas}_{\mathcal{C}}$ is minimal then for all $S \subseteq X$, $|\mathrm{Bas}_{\mathcal{C}}(S)|$ is at most the VC-dimension of $\mathcal{C}$.

Whenever clear from the context the subscript $\mathcal{C}$ will be dropped from $\mathrm{Cl}_{\mathcal{C}}$ and $\mathrm{Bas}_{\mathcal{C}}$. The Extended Closure Algorithm XCA (see figure 2) is designed to cope with noisy counterexamples. On each trial XCA chooses as current hypothesis the closure of the current set of positive counterexamples. When a (possibly noisy) positive counterexample x is received, the algorithm behaves like in the noiseless case adding x to the current set of positive counterexamples. However, if x was noisy, then a negative counterexample might be received in a later trial, since the new H will be too big containing at least the noisy x. Whenever that happens, that is XCA receives a negative counterexample, H is shrunk by removing from the current set S of positive counterexamples its basis (thus possibly all of S).

Algorithm XCA.
Input: Hypothesis class $\mathcal{H}$.

- Initialize the state variable $S_0 := \emptyset$.
- **For** $t = 0, 1, \ldots$
  1. Let $H_t := \mathrm{Cl}_{\mathcal{H}}(S_t)$ be the current hypothesis.
  2. Read next labeled instance (x_{t+1}, ℓ_{t+1}).
  3. **If** $H_t(x_{t+1}) = 0$ **and** $\ell_{t+1} = 1$ **then** $S_{t+1} := S_t \cup \{x_{t+1}\}$.
     If $H_t(x_{t+1}) = 1$ **and** $\ell_{t+1} = 0$ **then** $S_{t+1} := S_t \setminus \mathrm{Bas}_{\mathcal{H}}(S_t)$.
     Else $S_{t+1} := S_t$.

Figure 2. A sketch of algorithm XCA (the eXtended Closure Algorithm).

We are now ready to prove the main result of this section.

Theorem 3.3. Let $\mathcal{C}$ be a concept class and $\mathcal{H}$ be an intersection-closed concept class such that $\mathcal{C} \subseteq \mathcal{H}$. Then for any basis operator $\mathrm{Bas}_{\mathcal{H}}$, and for any $K \geqslant 0$,

$$\mathrm{MB}(\mathrm{XCA}, \mathcal{C}, \mathcal{H}, K) \leqslant \mathrm{MB}(\mathrm{CA}, \mathcal{C}, \mathcal{H}) + (d+1)K, \tag{1}$$

where

$$d = \max\left\{\left|\,\mathrm{Bas}_{\mathcal{H}}(S)\right|\colon\ S \subseteq X,\ |S| \leqslant \mathrm{MB}(\mathrm{CA}, \mathcal{C}, \mathcal{H})\right\}.$$

Moreover, if $\mathrm{Bas}_{\mathcal{H}}$ is minimal, then d is at most the VC-dimension of $\mathcal{H}$.

In the proof of the theorem we will assume without loss of generality that algorithm XCA does not receive supporting examples such that $\ell_{t+1} = H_t(x_{t+1})$. We will use the following lemma bounding the number of counterexamples kept in the state variable.

Lemma 3.4. After any sequence of counterexamples $x_1, \ldots, x_q$ the state variable S_q of algorithm XCA contains at most $\mathrm{MB}(\mathrm{CA}, \mathcal{C}, \mathcal{H})$ *correct* counterexamples.

Proof. Let $\emptyset = T_0, T_1, \ldots, T_m$ be a sequence of subsets of X such that for all $1 \leqslant i \leqslant m$

$$T_i = T_{i-1} \cup \{x_i\} \quad \text{for some } x_i \in X \setminus \mathrm{Cl}_{\mathcal{H}}(T_{i-1}). \tag{2}$$

Obviously $|T_i| = i$. Furthermore, if there is a concept $C \in \mathcal{C}$ such that $x_i \in C$ for all $1 \leqslant i \leqslant m$ then $m \leqslant \mathrm{MB}(\mathrm{CA}, \mathcal{C}, \mathcal{H})$ since the counterexamples $x_1, \ldots, x_m$ can be given to the closure algorithm when learning C with hypotheses from $\mathcal{H}$.

We prove the lemma by induction on q, showing that for any sequence of counterexamples $x_1, \ldots, x_q$ there is a sequence $\emptyset = T_0, T_1, \ldots, T_{m_q} \subseteq X$ with property (2)

such that T_{m_q} equals the subset $S_q^{(c)} \subseteq S_q$ of *correct* counterexamples of the state variable S_q of algorithm XCA. Since the state variable contains only positive counterexamples the correct counterexamples are elements of the target concept which implies the lemma.

The case $q = 0$ is trivial. Then assume that there exists a sequence $\emptyset = T_0, T_1, \ldots, T_{m_{q-1}}$ with property (2) and $T_{m_{q-1}} = S_{q-1}^{(c)}$. If x_q is a positive counterexample then $x_q \notin \mathrm{Cl}_{\mathcal{H}}(S_{q-1})$ and $S_q = S_{q-1} \cup \{x_q\}$. Thus $x_q \notin \mathrm{Cl}_{\mathcal{H}}(S_{q-1}^{(c)})$ and if x_q is a correct counterexample then $S_q^{(c)} = S_{q-1}^{(c)} \cup \{x_q\}$, otherwise $S_q^{(c)} = S_{q-1}^{(c)}$. If $S_q^{(c)} = S_{q-1}^{(c)}$ the same sequence $T_0, T_1, \ldots, T_{m_{q-1}} = S_{q-1}^{(c)} = S_q^{(c)}$ satisfies (2). If $S_q^{(c)} = S_{q-1}^{(c)} \cup \{x_q\}$ then $T_0, T_1, \ldots, T_{m_{q-1}} = S_{q-1}^{(c)}$, $T_{m_q} = S_q^{(c)}$ satisfies (2). If x_q is a negative counterexample then $S_q = S_{q-1} \setminus \mathrm{Bas}_{\mathcal{H}}(S_{q-1})$ and $S_q^{(c)} = S_{q-1}^{(c)} \setminus \mathrm{Bas}_{\mathcal{H}}(S_{q-1})$. Define $T_i' = T_i \setminus \mathrm{Bas}_{\mathcal{H}}(S_{q-1})$ for $i = 0, 1, \ldots, m_{q-1}$. If $x_i \in \mathrm{Bas}_{\mathcal{H}}(S_{q-1})$ then $T_i' = T_{i-1}'$, if $x_i \notin \mathrm{Bas}_{\mathcal{H}}(S_{q-1})$ then $T_i' = T_{i-1}' \cup \{x_i\}$ and $x_i \notin \mathrm{Cl}_{\mathcal{H}}(T_{i-1}')$ since $\mathrm{Cl}_{\mathcal{H}}(T_{i-1}') \subseteq \mathrm{Cl}_{\mathcal{H}}(T_{i-1})$. Hence, after removing duplicates from $T_0', T_1', \ldots, T_{m_{q-1}}'$ we have a sequence satisfying (2), which completes the proof of the lemma. □

Proof of theorem 3.3. Let $S = x_1, \ldots, x_q$ be the sequence of counterexamples presented by the adversary, S_q the subsequence of S which corresponds to the content of the state variable, and S' the subsequence of all other elements in S. Thus $q = |S_q| + |S'|$. According to lemma 3.4, $|S_q| \leqslant \mathrm{MB}(\mathrm{CA}, \mathcal{C}, \mathcal{H}) + K_q$, where K_q is the number of false counterexamples in S_q. Denoting by $K' \leqslant K - K_q$ the number of false counterexamples in S', it suffices to show that $|S'| \leqslant (d+1)K'$. Observe that S' consists of all negative counterexamples and all positive counterexamples in S which were removed from the state variable at some point. Consider a trial $t \in \{1, \ldots, q\}$ where a negative counterexample x_t was presented and a set P_t of at most d positive counterexamples was removed from the current state variable. Either x_t is a false negative counterexample or P_t contains a false positive counterexample since $x_t \in \mathrm{Cl}(P_t)$. Thus, we may remove x_t and the elements of P_t (at most $d+1$ elements altogether) from S and charge the false counterexample for that. Since no false counterexample is removed twice, at most $(d+1)K'$ elements of S are removed, i.e., $|S'| \leqslant (d+1)K'$. An application of lemma 3.2 concludes the proof. □

4. A general upper bound on the noise rate

In this section we introduce a general technique to prove upper bounds on the noise rate tolerable by any on-line learner (therefore disregarding computational issues).

Theorem 4.1. Let $\mathcal{C}$, $\mathcal{H}$ be (possibly identical) concept classes on domain X. Let $S = \{(x_1, \ell_1), \ldots, (x_s, \ell_s)\}$ be a subset of $X \times \{0, 1\}$ and, for all $1 \leqslant i \leqslant s$, let S_i be S where (x_i, ℓ_i) has been replaced by $(x_i, 1 - \ell_i)$. If the following hold:

(1) $x_i \neq x_j$ for $1 \leqslant i < j \leqslant s$,

(2) no $H \in \mathcal{H}$ is consistent with S,

(3) for all $1 \leqslant i \leqslant s$, S_i is consistent with some $C_i \in \mathcal{C}$,

then EQ($\mathcal{C}, \mathcal{H}, 1/s$) $= \infty$ and EQ($\overline{\mathcal{C}}, \overline{\mathcal{H}}, 1/s$) $= \infty$.

Furthermore, MB($\mathcal{C}, \mathcal{H}, K$) $\geqslant (s-1) + sK$ and MB($\overline{\mathcal{C}}, \overline{\mathcal{H}}, K$) $\geqslant (s-1) + sK$ for all $K \geqslant 0$.

Proof. Let A be an on-line learner for $\mathcal{C}$ using hypotheses from $\mathcal{H}$. For all $q \geqslant 0$, let $H_q \in \mathcal{H}$ be A's hypothesis after the adversary has returned q counterexamples. By definition of S, some $(x_j, \ell_j) \in S$ can be found such that $H_q(x_j) \neq \ell_j$. The adversary then returns the counterexample x_j.

We now show that after any number of counterexamples q there is a target $C \in \mathcal{C}$ such that at most q/s counterexamples disagree with C. Fix a $q \geqslant 0$. By the pigeonhole principle, after q counterexamples there is some $1 \leqslant i \leqslant s$ such that the adversary returned the counterexample x_i at most q/s times. Let C_i be any concept in $\mathcal{C}$ consistent with S_i. Notice that, by definition of S, C_i is consistent with all counterexamples x_j such that $j \neq i$. Thus C_i will disagree with at most q/s counterexamples.

By flipping the labels of S one can apply the same argument to the concept classes $\overline{\mathcal{C}}$, $\overline{\mathcal{H}}$. The bound for the MB model is derived similarly. □

5. Applications

In this section we give some applications of theorems 3.3 and 4.1. The first one is a simple upper bound on the tolerable noise rate when learning subspaces of an arbitrary linear space.

Corollary 5.1. Let $\mathcal{V}$ be the class of all subspaces of a d-dimensional linear space V. Then EQ($\mathcal{V}, 1/(d+1)$) $= \infty$ and MB($\mathcal{V}, K$) $=$ MB(XCA, $\mathcal{V}, K$) $= d + (d+1)K$ for all $K \geqslant 0$.

Proof. We fix d linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ in $\mathcal{V}$ and set $\mathbf{u} = \sum_{i=1}^{d} \mathbf{v}_i$. It is then easy to see that the set $\{(\mathbf{v}_1, 1), \ldots, (\mathbf{v}_d, 1), (\mathbf{u}, 0)\}$ fulfills the conditions of theorem 4.1. To prove the upper bound on MB(XCA, $\mathcal{V}, K$) we use the identity basis operator for the class $\mathcal{V}$. Then the corollary follows immediately from theorem 3.3 since MB(CA, $\mathcal{V}$) $= d$. □

Let $\mathbf{0} = (0, \ldots, 0)$, $\mathbf{1} = (1, \ldots, 1)$, and $\mathbf{e}_1, \ldots, \mathbf{e}_n$ the unit vectors of $\{0,1\}^n$, where n is made clear from the context.

Let MON$_n$ be the concept class of all the boolean functions that can be expressed as monotone monomials (that is monomials containing only unnegated variables) over n boolean variables. Let k-CNF$_n$ be the concept class of all boolean functions over $\{0,1\}^n$ that can be expressed in conjunctive normal form using clauses with at most

k literals (k-clauses). Notice that both classes are intersection-closed. An easy result is the following.

Corollary 5.2. For any $K \geqslant 0$ the class MON_n is on-line learnable with $\text{MB}(\text{MON}_n, K) = \text{MB}(\text{XCA}, \text{MON}_n, K) = n + (n+1)K$. Furthermore, XCA runs in time polynomial in n and K.

Proof. Notice that the closure of a set S of positive counterexamples is the longest monomial M satisfying all counterexamples. Thus all hypotheses in MON_n are representable with $O(n)$ bits and their values are computable in linear time. Also, each positive counterexample added to S shortens M by dropping at least one variable. Thus $\text{EQ}(\text{CA}, \text{MON}_n) \leqslant n$. Consider now the Extended Closure Algorithm using MON_n as hypothesis class and the identity basis operator for the class MON_n. By theorem 3.3 we immediately conclude $\text{MB}(\text{XCA}, \text{MON}_n, K) \leqslant n + (n+1)K$. To prove the lower bound on $\text{MB}(\text{MON}_n, K)$ let S be the set $\{(\bar{\mathbf{e}}_1, 1), \ldots, (\bar{\mathbf{e}}_n, 1), (\mathbf{1}, 0)\}$. Clearly, all the instances are distinct and no monotone monomial is consistent with S (the empty monomial has constant value 1 on all of $\{0,1\}^n$). Moreover, we can easily find a monotone monomial consistent with the set S' obtained by flipping the label of any single instance in S. An application of theorem 4.1 then concludes the proof. □

Corollary 5.2 allows us to prove a second result.

Corollary 5.3. Let $N = \sum_{i=0}^{k} \binom{n}{i} 2^i$. Then for any $K \geqslant 0$ the class $k\text{-CNF}_n$ is on-line learnable with $\text{MB}(k\text{-CNF}_n, K) \leqslant N + (N+1)K$ and in time polynomial in N and K.

Proof. Observe that $N = \sum_{i=0}^{k} \binom{n}{i} 2^i$ equals the number of satisfiable k-clauses over n variables $x_1, \ldots, x_n$. Let $y_1, \ldots, y_N$ be a set of boolean variables where each y_i ($1 \leqslant i \leqslant N$) is uniquely associated to a satisfiable k-clause. Then any $\mathbf{x} \in \{0,1\}^n$ is mapped to a $\mathbf{y}_\mathbf{x} \in \{0,1\}^N$ (notice that the image of $\{0,1\}^n$ under this mapping is a strict subset of $\{0,1\}^N$). Therefore, each k-CNF formula F on $x_1, \ldots, x_n$ will be mapped to a monomial M_F on $y_1, \ldots, y_N$ such that $F(\mathbf{x}) = M_F(\mathbf{y}_\mathbf{x})$. The algorithm A for learning $k\text{-CNF}_n$ uses the Extended Closure Algorithm, applied to the class MON_N, as a subroutine. Each time a counterexample $\mathbf{x}$ is received A maps it to the corresponding $\mathbf{y}_\mathbf{x}$ and feeds it to XCA. In response, XCA outputs a concept $C \in \text{MON}_N$. A then translates C into a conjunction of satisfiable k-clauses H which becomes A's new current hypothesis. An application of corollary 5.2 then yields $\text{MB}(k\text{-CNF}_n, K) \leqslant N + (N+1)K$. The computation time spent by algorithm A on each trial is clearly polynomial in N. □

For all $n \geqslant 1$ let PARITY_n be the class of parity functions over all subsets of $\{x_1, \ldots, x_n\}$. The following observation legitimates the use of the Extended Closure Algorithm to learn PARITY_n.

Lemma 5.4 [9]. Each $C \in \text{PARITY}_n$ is a linear subspace of $\{0,1\}^n$ with respect to the addition modulo 2 and the usual scalar product over $\{0,1\}$.

Let SUB_n be the class of all linear subspaces $\{0,1\}^n$ with respect to the operations defined in the statement of lemma 5.4.

Corollary 5.5. For all $K \geqslant 0$ the class PARITY_n is on-line learnable with $\text{MB}(\text{PARITY}_n, \text{SUB}_n, K) = \text{MB}(\text{XCA}, \text{PARITY}_n, \text{SUB}_n, K) = n + (n+1)K$. Furthermore, XCA runs in time polynomial in n and K.

Proof. By lemma 5.4 we have $\text{PARITY}_n \subseteq \text{SUB}_n$. We run the Extended Closure Algorithm using the identity basis operator Bas_I for SUB_n. Since SUB_n is the class of all linear subspaces of an n-dimensional linear space we immediately have $\text{EQ}(\text{CA}, \text{SUB}_n) \leqslant n$ and therefore theorem 3.3 implies $\text{MB}(\text{XCA}, \text{PARITY}_n, \text{SUB}_n, K) \leqslant n + (n+1)K$. Finally, all hypotheses $H \in \text{SUB}_n$ can be represented using $\text{O}(n^2)$ bits and computing the value of any H (i.e., testing for linear independence a set of at most n boolean vectors over $\{0,1\}^n$) takes time polynomial in n (see, e.g., [20]). Thus XCA spends polynomial time (in n) on each trial.

The lower bound on $\text{MB}(\text{PARITY}_n, \text{SUB}_n, K)$ can be established analogously to corollary 5.1 if the $\mathbf{v}_i$ are chosen as the unit vectors. □

Notice that $\overline{\text{PARITY}}_n$ is the concept class $\{C_I\colon\ I \subseteq \{1,\ldots,n\}\}$, where $C_I(\mathbf{x}) = \bigoplus_{i \in I} x_i$ for all $\mathbf{x} \in \{0,1\}^n$ and $\bigoplus$ denotes addition modulo 2. A generalization of $\overline{\text{PARITY}}_n$ is the class of *ring-sum expansions* over n variables whose learnability in the PAC model was studied in [9]. A ring-sum expansion is a boolean function $C_{\mathcal{M}}(\mathbf{x}) = \bigoplus_{M \in \mathcal{M}} M(\mathbf{x})$ for an arbitrary $\mathcal{M} \subseteq \text{MON}_n$. A well-known fact states that any boolean function can be represented as a ring-sum expansion. By insisting that at most k variables (for $k \leqslant n$) appear in each monomial one obtains the class of k-ring-sum expansions (k-RSE).

Corollary 5.6. Let $N = \sum_{i=0}^{k} \binom{n}{i}$. Then for all $K \geqslant 0$ the class $k\text{-RSE}_n$ of k-ring-sum expansions over $\{0,1\}^n$ is on-line learnable with $\text{MB}(k\text{-RSE}_n, \mathcal{H}_{k,n}, K) \leqslant N + (N+1)K$ and in time polynomial in N and K, where $\mathcal{H}_{k,n}$ is a concept class that contains $k\text{-RSE}_n$, is evaluatable in polynomial time, and such that its complement $\overline{\mathcal{H}}_{k,n}$ is intersection-closed.

Proof. The statement is proven in much the same way as corollary 5.3. We consider a new set of $N = \sum_{i=0}^{k} \binom{n}{i}$ boolean variables and a one-to-one mapping between the monotone monomials and this set of variables. Let $\mathbf{y}_{\mathbf{x}}$ be the unique element of $\{0,1\}^N$ to which $\mathbf{x} \in \{0,1\}^n$ gets mapped. Then for each $C \in k\text{-RSE}_n$ there is a $H_C \in \overline{\text{PARITY}}_N$ such that $C(\mathbf{x}) = H_C(\mathbf{y}_{\mathbf{x}})$. Let A be the algorithm that runs the Extended Closure Algorithm on PARITY_N using SUB_N as hypothesis class. A's initial hypothesis is the complement of XCA's initial hypothesis. Each time a new counterexample $\mathbf{x}$ is

received, A computes $\mathbf{y}_{\mathbf{x}}$ (in time polynomial in n) and presents it to XCA. The complement of XCA's new hypothesis is then A's new current hypothesis. By the above considerations and corollary 5.5 we easily obtain MB(k-RSE$_n$, $\mathcal{H}_{k,n}$, K) $\leqslant N + (N+1)K$ where $\mathcal{H}_{k,n}$ is the polynomial-time evaluatable class $\overline{\text{SUB}}_N$ in which each variable y_i ($1 \leqslant i \leqslant N$) has been replaced by its associated monomial. Finally, notice that the time spent by A on each trial is polynomial in n. □

Another application of theorem 3.3 yields the learnability of *integer lattices* in the presence of noise. An integer lattice $\mathcal{L}^k$ is a subset of $\mathbf{Z}^k$ closed with respect to the operations of addition and multiplication by an integer. Let $\mathcal{L}^k(n)$ be the restriction of $\mathcal{L}^k$ on $\{-n, \ldots, -1, 0, 1, \ldots, n\}^k$. Notice that $\mathcal{L}^k(n)$ is intersection-closed.

In [13] it is shown that $\mathcal{L}^k(n)$ is noise-free on-line learnable by the Closure Algorithm in time polynomial in $\log n$ and k. We can show the following.

Corollary 5.7. Let $g = \lfloor k \log n + k(\log k)/2 \rfloor + k$. Then:

(1) for all $K \geqslant 0$ the class $\mathcal{L}^k(n)$ is on-line learnable in time polynomial in g and K with MB(XCA, $\mathcal{L}^k(n)$, K) $\leqslant g + (g+1)K$;

(2) EQ($\mathcal{L}^k(n), r$) $= \infty$ for any $r \geqslant (1 - \text{o}(1))\log\log n/(k \log n)$ where o(1) $\to 0$ as $n \to \infty$.

Proof. To prove part (1) we apply the Extended Closure Algorithm with the identity basis operator. Checking for membership in the closure of a set S of counterexamples is computable in polynomial time (see, e.g., [20]). The bound of the number of mistakes is then obtained by applying theorem 3.3 to the bound MB(CA, $\mathcal{L}^k(n)$) $\leqslant g$ proven in [13].

To prove part (2) let $p_1, \ldots, p_m \in \mathbf{Z}$ be distinct primes with $\prod_{i=1}^m p_i \leqslant 2n$. Denote by $\mathbf{e}_1, \ldots, \mathbf{e}_k$ the unit vectors of $\mathbf{Z}^k$. For $1 \leqslant \ell \leqslant k$, $1 \leqslant i \leqslant m$, let $\mathbf{x}_{\ell i} = (\prod_{j \neq i} p_j)\mathbf{e}_\ell$, and $\mathbf{x}_0 = (1, \ldots, 1)$.

No $C \in \mathcal{L}^k(n)$ is consistent with $S = \{(\mathbf{x}_0, 0), (\mathbf{x}_{1,1}, 1), \ldots, (\mathbf{x}_{k,m}, 1)\}$. Furthermore, the set $\{(\mathbf{x}_0, 1), (\mathbf{x}_{1,1}, 1), \ldots, (\mathbf{x}_{k,m}, 1)\}$ is consistent with $\{-n, \ldots, -1, 0, 1, \ldots, n\}^k \in \mathcal{L}^k(n)$.

Now assume that $(x_{\ell,i}, 1)$ is replaced by $(\mathbf{x}_{\ell,i}, 0)$ giving $S_{\ell,i}$. Then $S_{\ell,i}$ is consistent with Cl($\{\mathbf{e}_1, \ldots, \mathbf{e}_{\ell-1}, p_i\mathbf{e}_\ell, \mathbf{e}_{\ell+1}, \ldots, \mathbf{e}_k\}$). Using [13, equation (1), p. 245] for all $\varepsilon > 0$ there exists n_ε such that for all $n \geqslant n_\varepsilon$ we can choose $m > (1+\varepsilon)\log n/(\log\log n)$ primes satisfying $\prod_{i=1}^m p_i \leqslant 2n$ and thus proving the result. □

As another application consider the target class defined as follows. For all positive integers m let $\mathbf{Z}_m$ be the class of residues modulo m. If n is a positive integer, $k < m$ a nonnegative integer, and $\mathbf{w}$ a vector in $\mathbf{Z}_m^n$, we define the *k-counting function* $M_{\mathbf{w},k} : \mathbf{Z}_m^n \to \{0, 1\}$ by

$$M_{\mathbf{w},k}(\mathbf{x}) = \begin{cases} 0 & \text{if } \sum_i w_i x_i \equiv k \bmod m, \\ 1 & \text{otherwise.} \end{cases}$$

Let k-COUNT$_n$ be the class $\{M_{\mathbf{w},k}\colon\ \mathbf{w} \in \mathbf{Z}_m^n\}$ of k-counting functions over $\mathbf{Z}_m^n$ and k-DCOUNT$_n$ the class of all disjunctions of functions in k-COUNT$_n$. In [8], an algorithm using the Closure Algorithm as subroutine is shown to learn k-DCOUNT$_n$ over $\mathbf{Z}_p^n$ for any prime p with at most $n+1$ mistakes in polynomial time. Moreover, the algorithm generates hypotheses that are disjunctions of at most n k-counting functions. By applying theorem 3.3 to this result we can easily get the following.

Corollary 5.8. For all $K \geqslant 0$ and for all primes p the class k-DCOUNT$_n$ over $\mathbf{Z}_p^n$ is on-line learnable in time polynomial in n, p and K with MB(k-DCOUNT$_n, K) \leqslant n + 1 + (n+2)K$.

We conclude the section by proving an upper bound on the noise rate tolerable by any on-line learner for the class HALFSPACES$_n$ of all linearly separable boolean functions over $\{0,1\}^n$.

Corollary 5.9. For all $n > 1$, EQ(HALFSPACES$_n, 1/(n+2)) = \infty$ and

$$\text{MB(HALFSPACES}_n, K) \geqslant n + 1 + (n+2)K$$

for all $K \geqslant 0$.

Proof. Let S be the set $\{(\mathbf{0}, 0), (\mathbf{e}_1, 1), \ldots, (\mathbf{e}_n, 1), (\mathbf{1}, 0)\}$. Clearly, no halfspace is consistent with S. It is also easy to see that by flipping the label of either $(0, \ldots, 0)$ or $(1, \ldots, 1)$ we can find consistent halfspaces. Finally, choose $1 \leqslant i \leqslant n$ and let S_i be S with $(\mathbf{e}_i, 0)$ in place of $(\mathbf{e}_i, 1)$. Consider the halfspace $\{(v_1, \ldots, v_n)\colon \sum_{i=1}^n w_i v_i \geqslant 1\}$, where $w_j = 1$ for $j \neq i$ and $w_i = 1-n$. It is easy to see that this halfspace is consistent with S_i. Thus theorem 4.1 can be applied and the result immediately follows. □

Remark. In the above applications we did not use the full generality of algorithm XCA because we only used the identity mapping as basis operator. Nevertheless, the analysis of XCA is not much easier in this case, and in fact there are concept classes where other basis operators can be used, e.g., the class of axis-parallel rectangles (see [3]).

6. From on-line to PAC learning in the presence of noise

In this section we show that any learning algorithm developed for the on-line model with noise can be canonically and efficiently turned into an algorithm for the PAC model with malicious noise.

In the standard PAC model introduced by Valiant [21] the learner has access to an oracle returning on each query some labeled instance $(x, C(x))$, where C is some fixed concept belonging to a given target class $\mathcal{C}$ and x is randomly drawn from a fixed distribution D over the domain X. Both C and D are unknown to the learner and each random draw of x is independent on the outcomes of the other draws.

In the malicious variant of the PAC model introduced by Kearns and Li [14] (the reader is referred to that paper for motivations) on each query the oracle is allowed

Algorithm A_{pac}.
Input: A labeled sample $(x_1, \ell_1), \ldots, (x_m, \ell_m)$.

1. Initialize algorithm A.
2. Remove from the sample a counterexample (x_i, ℓ_i) to A's current hypothesis and present it to A until all examples have been removed from the sample or no further counterexamples can be found.
3. Output A's final hypothesis $H \in \mathcal{H}$.

Figure 3. A sketch of the PAC learning algorithm A_{pac} using the on-line learning algorithm A as subroutine.

to flip a coin with fixed bias η for heads. If the outcome is heads, the oracle returns some labeled instance (x, ℓ) adversarially chosen from $X \times \{0, 1\}$. If the outcome is tails, the oracle is forced to behave exactly like in the standard model returning the correctly labeled instance $(x, C(x))$ where $x \sim D$.

In both the standard and the malicious PAC model the learner's goal on all inputs $\varepsilon, \delta > 0$ is to output some hypothesis $H \in \mathcal{H}$ (where $\mathcal{H}$ is the learner's fixed hypothesis class) by querying the oracle at most m times for some $m = m(\varepsilon, \delta)$ in the standard model and for some $m = m(\varepsilon, \delta, \eta)$ in the malicious model. For all targets $C \in \mathcal{C}$ and distributions D, the hypothesis H of the learner must satisfy $\mathbf{E}_{x \sim D}[H(x) \neq C(x)] \leqslant \varepsilon$ with probability at least $1 - \delta$ with respect to the oracle's randomization. We will call ε and δ respectively the accuracy and the confidence parameter.

We now present a conversion of an on-line learning algorithm A to a learning algorithm A_{pac} (see figure 3) for the malicious PAC model. The following lemma will be used.[3]

Lemma 6.1 [2]. For all target classes $\mathcal{C}$ and hypothesis classes $\mathcal{H}$ on a domain X, for all targets $C \in \mathcal{C}$, for all distributions D on X, and for all $\varepsilon, \delta, \gamma > 0$. Given a sample of m instances independently drawn from D and labeled by C, where

$$m \geqslant \frac{8}{\gamma^2 \varepsilon}\left(d \ln \frac{48}{\gamma^2 \varepsilon} + \ln \frac{4}{\delta}\right)$$

and d is the VC-dimension of $\mathcal{H}$, the probability that there exists $H \in \mathcal{H}$ making at most $(1 - \gamma)\varepsilon m$ mistakes on the sample and such that $\mathbf{E}_{x \sim D}[H(x) \neq C(x)] > \varepsilon$ is at most δ with respect to the random sample draw.

Theorem 6.2. Choose a target class $\mathcal{C}$, an hypothesis class $\mathcal{H}$ and suppose A is an on-line algorithm such that $\text{MB}(A, \mathcal{C}, \mathcal{H}, K) \leqslant m_0 + RK$ for some positive m_0, R and

[3] The result proven in [2] is more general. We specialize it for our purposes.

for all nonnegative K. Then for all $\alpha > 0$ and all $\varepsilon, \delta > 0$, given a sample of size

$$m = \max\left\{\frac{9}{2\alpha^2}\ln\frac{2}{\delta}, \frac{3m_0}{\alpha R}, \frac{72\varepsilon}{\alpha^2 R^2}\left(d\ln\frac{432\varepsilon}{\alpha^2 R^2} + \ln\frac{8}{\delta}\right)\right\},$$

where d is the VC-dimension of $\mathcal{H}$, the algorithm A_{pac} learns $\mathcal{C}$ using hypothesis class $\mathcal{H}$ in the PAC learning model with malicious noise rate $\varepsilon/R - \alpha$, accuracy ε, and confidence δ.

Proof. In the sample $S = \langle(x_t, \ell_t)\rangle_{1\leqslant t\leqslant m}$ returned by the malicious oracle, let K be the number of examples which were subject to malicious noise, i.e., those examples (x, ℓ), where x and ℓ have been arbitrarily chosen by the oracle. Let S' be the sample obtained from S by replacing each noisy example (x, ℓ) with $(x', C(x))$, where x' is independently drawn from D and $C \in \mathcal{C}$ is the target. The proof is based on the following observation: The total number of mistakes made by the final hypothesis H on the "clean" sample S' will be at most the number of counterexamples presented to A while run on S plus the number of remaining noisy examples in S that were not given to A as counterexamples. By applying standard PAC learning results we can then bound the expected error of H in terms of empirical error on S'.

Observe that K is the sum of m independent Bernoulli trials each with probability of success at most $\varepsilon/R - \alpha$. Thus, by standard Hoeffding bounds, for all $0 < \tau < 1$ the inequality $K \leqslant m(\varepsilon/R - \alpha + \tau)$ holds with probability at least $1 - \delta/2$ whenever $m \geqslant (1/2\tau^2)\ln(2/\delta)$.

Let $K_A \leqslant K$ be the number of noisy examples in S which were presented as counterexamples to A during the run of A_{pac}. Then the total number of counterexamples presented to A is bounded by $\text{MB}(A, \mathcal{C}, \mathcal{H}, K_A) \leqslant m_0 + RK_A$. Hence the number of examples in S' that are misclassified by A_{pac}'s final hypothesis is at most

$$m_0 + RK_A + (K - K_A) \leqslant m_0 + RK \leqslant m_0 + m\varepsilon - mR(\alpha - \tau),$$

where the last inequality holds with probability at least $1 - \delta/2$. If we now choose $\tau = \alpha/3$ and $\gamma = \alpha R/(3\varepsilon)$ then

$$m_0 + m\varepsilon - mR(\alpha - \tau) \leqslant (1 - \gamma)m\varepsilon \quad \text{for } m \geqslant \frac{3m_0}{\alpha R}.$$

Applying lemma 6.1 we find that $\mathbf{E}_{x\sim D}[H(x) \neq D(x)] > \varepsilon$ with probability at most $\delta/2$, yielding the theorem. □

We can apply theorem 6.2 to efficiently learn each of the concept classes MON_n, $k\text{-CNF}_n$, $k\text{-RSE}_n$, $\mathcal{L}^k(n)$ and $k\text{-DCOUNT}_n$ in the malicious PAC model. However, it should be also noted that there exist techniques to efficiently turn any learning algorithm for the *noise-free* PAC model into PAC algorithms tolerating a certain rate of malicious noise. In [14, theorem 11, p. 824] it is shown that any PAC learning algorithm using sample size m can be efficiently turned into an algorithm tolerating a malicious noise rate of $(\ln m)/m$.

Via different techniques, using [11, corollary 8, p. 10] and a result from [6], it can be shown how to efficiently use any PAC learning algorithm with *finite* hypothesis space of VC-dimension d to learn in presence of a malicious noise rate arbitrarily close to $\varepsilon/(7d+1+\varepsilon)$. It is likely that, via a more careful analysis, the constant in front of d might be improved.

7. Conclusions and open problems

In this paper we have introduced a new on-line algorithm (a simple variant of the popular "Closure Algorithm") for learning intersection-closed concept classes while tolerating a bounded fraction of adversarial noise. In several natural cases the running time of our algorithm has been shown to be polynomial in the problem's parameters. To our knowledge, this is the first example of a quite general and efficient on-line strategy for learning in presence of noise.

An open problem is whether the sample size bounds for converting an on-line algorithm to a malicious PAC learning algorithm can be substantially improved or, alternatively, the general upper bound of ε/R on the noise tolerance brought closer to the information-theoretic upper bound $\varepsilon/(1+\varepsilon)$.

A second open problem is whether randomized variants of our algorithm can be shown to have good bounds on the expected number of mistakes.

Acknowledgements

This work was partially supported by the ESPRIT NeuroCOLT project No. 8556. The first author was also partially supported by grant J01028-MAT from the Fonds zur Förderung der wissenschaftlichen Forschung. The second author wishes to thank the Institute for Theoretical Computer Science (IGI) at the Graz University of Technology that he visited during the academical year 1993–1994. Finally we want to thank an anonymous referee for valuable comments.

References

[1] D. Angluin, Queries and concept learning, Machine Learning 2(4) (1988) 319–342.
[2] M. Anthony and J. Shawe-Taylor, A result of Vapnik with applications, Discrete Applied Mathematics 47 (1994) 207–217.
[3] P. Auer, On-line learning of rectangles in noisy environments, in: *Proceedings of the 6th Annual ACM Workshop on Computational Learning Theory* (ACM Press, 1993) pp. 253–261.
[4] P. Auer and P.M Long, Simulating access to hidden information while learning, in: *Proceedings of the 26th ACM Symposium on the Theory of Computing* (ACM Press, 1994) pp. 263–272.
[5] S. Boucheron, Learnability from positive examples in the Valiant framework, Manuscript (1988).
[6] N. Cesa-Bianchi, Models of learning with noise, Unpublished manuscript (1994).
[7] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold and M.K. Warmuth, On-line prediction and conversion strategies, in: *Proceedings of the First Euro-COLT Workshop*, The Institute of Mathematics and

its Applications Conference Series – New Series Number 53 (Clarendon Press, Oxford, 1994) pp. 205–216.
[8] Z. Chen and S. Homer, On learning counting functions with queries, in: *Proceedings of the 7th Annual ACM Workshop on Computational Learning Theory* (ACM Press, 1994) pp. 218–227.
[9] P. Fischer and H.U. Simon, On learning ring-sum-expansions, SIAM Journal on Computing 21 (1992) 181–192.
[10] D. Haussler, N. Littlestone and M.K. Warmuth, Predicting $\{0,1\}$-functions on randomly drawn points, Information and Computation 115(2) (1994) 248–292.
[11] D.P. Helmbold and P.M Long, Tracking drifting concepts by minimizing disagreements, Machine Learning 14(1) (1994) 27–45.
[12] D.P. Helmbold, R. Sloan and M.K. Warmuth, Learning nested differences of intersection-closed concept classes, Machine Learning 5(2) (1990) 165–196.
[13] D.P. Helmbold, R. Sloan and M.K. Warmuth, Learning integer lattices, SIAM Journal on Computing 21(2) (1992) 240–266.
[14] M.J. Kearns and M. Li, Learning in the presence of malicious errors, SIAM Journal on Computing 22(4) (1993) 807–837.
[15] N. Littlestone, Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm, Machine Learning 2(4) (1988) 285–318.
[16] N. Littlestone, Mistake bounds and logarithmic linear-threshold learning algorithms, Ph.D. thesis, University of California at Santa Cruz (1989).
[17] N. Littlestone and M.K. Warmuth, The weighted majority algorithm, *Information and Computation* 108 (1994) 212–261.
[18] B.K. Natarajan, On learning boolean functions, in: *Proceedings of the 19th ACM Symposium on the Theory of Computing* (ACM Press, 1987) pp. 296–304.
[19] B.K. Natarajan, *Machine Learning: A Theoretical Approach* (Morgan Kaufmann, San Mateo, CA, 1991).
[20] A. Schrijver, *Theory of Linear and Integer Programming* (Wiley, New York, 1986).
[21] L. Valiant, A theory of the learnable, Communications of the ACM 27(11) (1984) 1134–1142.