Electronic Colloquium on Computational Complexity, Report No. 63 (2000)

ECCC

On-line Learning of Rectangles in Noisy Environments

Peter Auer

Institute for Theoretical Computer Science
Technische Universität Graz
Klosterwiesgasse 32/2
A-8010 Graz, Austria

July 13, 2000

Abstract

We investigate the implications of noise in the equivalence query model. Besides some results for general target and hypotheses classes, we prove bounds on the learning complexity of d-dimensional rectangles (of size at most n^d) in the case where only rectangles are allowed as hypotheses. Our noise model assumes that a certain fraction of the examples is noisy. We show that d-dimensional rectangles are learnable if and only if the fraction of noisy examples is less than $1/(d+1)$, where learnable means that the learner can learn the target by a finite number of examples. Besides this structural result we present an algorithm which learns rectangles in $\mathrm{poly}(\frac{d \log n}{1-r(2d+1)})$ time using $\mathrm{O}(\frac{d^3 \log n}{(1-r(2d+1))^2})$ examples if the fraction of noise r is less than $\frac{1}{2d+1}$. As a related result we prove for the noise-free case that the number of examples necessary to learn is at least $\Omega(\frac{d^2}{\log d} \log n)$, where the best known upper bound on the learning complexity is $\mathrm{O}(d^2 \log n)$.

ISSN 1433-8092

1 Introduction

In the following we will deal with a generalization of the equivalence query model which introduces a notion of noise. The equivalence query model [Ang88, Lit88, MT92] assumes that the learner has to identify a target T from a target class $\mathcal{T}$ over some domain X, i.e. $T \in \mathcal{T} \subseteq 2^X$, where 2^X denotes the powerset of X. Hence the learner has to learn a classification of the elements in X and it knows in advance that only classifications from $\mathcal{T}$ are possible.

The learner proceeds by proposing hypotheses H_1, H_2, $\ldots \subseteq X$. After each hypothesis $H_q \neq T$ it receives a counterexample (CE), which is a misclassified point $x_q \in H_q \triangle T := (H_q \setminus T) \cup (T \setminus H_q)$. CEs x_q with $x_q \in H_q$ are called negative CEs since they do not belong to the target, CEs $x_q \notin H_q$ are called positive CEs. A natural question is then how many hypotheses the learner needs to identify the target.

Now consider the case that some of the CEs are noisy, i.e. some $x_q \notin H_q \triangle T$. This means that the learner has correctly classified x_q but is told that the classification is incorrect. Since the learner can be fooled by noisy CEs, it is intuitively clear that the learner will not be able to learn the target if the fraction of noisy CEs is too high. Therefore we are interested in determining the maximal fraction of noisy CEs such that the learner can still learn, and we are also interested in the number of questions the learner needs to learn, given that a certain fraction will be noisy.

Definition 1.1 (Learning complexity) *Let $\mathcal{T} \subseteq 2^X$ be the target class and $\mathcal{H} \subseteq 2^X$ the set of allowed hypotheses. For each sequence of CEs $x_1, \ldots, x_q$, $q \geq 0$, a learning algorithm $\mathcal{A}$ has to produce a hypothesis $H_{q+1} = \mathcal{A}(x_1, \ldots, x_q) \in \mathcal{H}$. If r, $0 \leq r \leq 1$, is the maximal fraction of noise then the maximal number of CEs is given as*

$$\begin{aligned} &\mathrm{LC}_{\mathcal{A}}(T, \mathcal{H}, r) \\ &\quad = \sup\{Q \geq 0 | \exists x_1, \ldots, x_Q : |\{1 \leq q \leq Q : x_q \notin \mathcal{A}(x_1, \ldots, x_{q-1}) \triangle T\}| \leq rQ\}, \\ &\mathrm{LC}(\mathcal{T}, \mathcal{H}, r) = \min_{\mathcal{A}} \sup_{T \in \mathcal{T}} LC_{\mathcal{A}}(T, \mathcal{H}, r). \end{aligned}$$

Thus $\mathrm{LC}_{\mathcal{A}}(T, \mathcal{H}, r)$ is the maximal number of CEs such that only a fraction r of them are noisy, and $\mathrm{LC}(\mathcal{T}, \mathcal{H}, r)$ is the learning complexity of the optimal learning algorithm if it has to learn the most difficult target. If $\mathrm{LC}(\mathcal{T}, \mathcal{H}, r) = \infty$ then the target class $\mathcal{T}$ is not learnable using hypotheses from $\mathcal{H}$ if the fraction of noise is r. Observe that r bounds the number of noisy CEs only in respect to the total number of CEs. It is <u>not</u> necessarily true that $|\{1 \leq p \leq q : x_p \notin H_p \triangle T\}| \leq rq$ for all $q \leq Q$.

Remark 1.2 *From the above definition it follows trivially that, for any fixed domain X, $\mathrm{LC}(\mathcal{T}_1, \mathcal{H}_1, r_1) \leq \mathrm{LC}(\mathcal{T}_2, \mathcal{H}_2, r_2)$ if $\mathcal{T}_1 \subseteq \mathcal{T}_2$, $\mathcal{H}_2 \subseteq \mathcal{H}_1$, $r_1 \leq r_2$. Furthermore $\mathrm{LC}(\mathcal{T}, \mathcal{H}, r) = \infty$ if $\mathcal{T} \setminus \mathcal{H} \neq \emptyset$. Therefore we assume $\mathcal{T} \subseteq \mathcal{H}$.*

In applying the above definitions we do a worst case analysis, worst case insofar as we consider the most difficult target and a malicious environment. Learning from noisy CEs in the PAC-learning model was among others investigated in [AL88, KL88].

2 The general case

For general target classes only few results are known. The most important considers the case where arbitrary hypotheses are allowed. By the weighted majority algorithm [LW91] and simple considerations one easily obtains

Theorem 2.1 *For all target classes $\mathcal{T}$ with $|\mathcal{T}| \geq 2$ and $|X| < \infty$ we have for all $0 \leq r < 1/2$*

$$\begin{aligned} \mathrm{LC}(\mathcal{T}, 2^X, r) &\leq \frac{\log |\mathcal{T}|}{\log 2 + r \log r + (1-r)\log(1-r)}, \\ \mathrm{LC}(\mathcal{T}, 2^X, 1/2) &= \infty. \end{aligned}$$

Remark 2.2 *Observe that $-\log 2 < r \log r + (1-r)\log(1-r) < 0$ if $0 < r < 1/2$, $r \log r + (1-r)\log(1-r) \to 0$ if $r \to 0$, $r \log r + (1-r)\log(1-r) \sim -\log 2 + 2(1/2-r)^2$ if $r \to 1/2$.*

Therefore, if arbitrary hypotheses are allowed, any target class can be learned if and only if the fraction of noise is less then $1/2$. A drawback of the weighted majority algorithm is the fact that in general very complex hypotheses must be generated. First, the representation of these hypotheses might be very complex (relative to the representation of elements in $\mathcal{T}$, especially if $|\mathcal{T}| \ll 2^{|X|}$); secondly, the computation time to calculate the hypotheses might be infeasibly high. Furthermore it is often convenient to deal only with hypotheses from the target class. Thus, following [MT92], we assume $\mathcal{H} = \mathcal{T}$ and set

$$\mathrm{LC}(\mathcal{T}, r) := \mathrm{LC}(\mathcal{T}, \mathcal{T}, r).$$

The next theorem gives the maximal tolerable fraction of noise if only the size of the target class is known and the hypotheses must be from the target class.

Theorem 2.3 *For all target classes $\mathcal{T}$ we have for all $0 \leq r < 1/|\mathcal{T}|$ that*

$$\mathrm{LC}(\mathcal{T}, r) \leq \frac{|\mathcal{T}| - 1}{1 - r|\mathcal{T}|}.$$

Furthermore if $\mathcal{T}_n = \{\{1\}, \ldots, \{n\}\}$ then

$$\mathrm{LC}(\mathcal{T}_n, \frac{1}{n}) = \infty.$$

Corollary 2.4 *For all $n \in \mathbf{N}$ and $r \in [0,1]$ we have*

$$\{\forall \mathcal{T} : |\mathcal{T}| = n : [\mathrm{LC}(\mathcal{T}, r) < \infty]\} \Leftrightarrow r < 1/n.$$

3 Rectangles

In this section we study the natural geometric class of d-dimensional rectangles ([BEHW89, MT89, MT91, CM92]),

$$\mathcal{R}_n^d = \{\prod_{i=1}^{d} \{c_i^-, c_i^- + 1, \ldots, c_i^+\} : 0 \leq c_i^-, c_i^+ \leq n-1\},$$

$n \geq 2, d \geq 1$, where the rectangles are given by their "lower left" and "upper right" corners $c^- = (c_1^-, \ldots, c_d^-)$, $c^+ = (c_1^+, \ldots, c_d^+)$. This definition yields rectangles from the d-dimensional lattice space $\{0, \ldots, n-1\}^d$ with sides parallel to the coordinate axis, and furthermore $\emptyset \in \mathcal{R}_n^d$.

Concerning the learning complexity of $\mathcal{R}_n^d$ it is known that $\mathrm{LC}(\mathcal{R}_n^d, 0) = O(d^2 \log n)$ [CM92] and (by Theorem 2.1) $\mathrm{LC}(\mathcal{R}_n^d, 1/2) = \infty$. As trivial lower bound one can obtain $\mathrm{LC}(\mathcal{R}_n^d, 0) = \Omega(d \log n)$ [CM92]. For $d = 1$ the learning problem of rectangles is closely related to binary searching and there was previous research for example in [DGW92]. In the following we present our own results which give new and better bounds.

Theorem 3.1 *If $n \geq d \geq 2$ then*

$$\mathrm{LC}(\mathcal{R}_n^d, 0) = \Omega(\frac{d^2}{\log d} \log n).$$

If $n \leq d$ then

$$\mathrm{LC}(\mathcal{R}_n^d, 0) = \Omega(dn).$$

Theorem 3.2 *If $r \geq 1/(d+1)$ then*

$$\mathrm{LC}(\mathcal{R}_n^d, r) = \infty.$$

Theorem 3.3 *If $r < 1/(d+1)$ and $d \geq 2$ then*

$$\mathrm{LC}(\mathcal{R}_n^d, r) \leq \frac{dn}{1 - r(d+1)}.$$

Theorem 3.4 *There exists an algorithm* LR *such that for all $r < 1/(2d+1)$*

$$\mathrm{LC}_{\mathrm{LR}}(\mathcal{R}_n^d, r) \leq O(\frac{d^3 \log n}{(1 - r(2d+1))^2}),$$

and LR *runs in time $O(\frac{d^4 \log n}{(1-r(2d+1))^2} \log \frac{d^3 \log n}{(1-r(2d+1))^2})$.*

By Theorem 3.1 and [CM92] we have upper and lower bounds on $\mathrm{LC}(\mathcal{R}_n^d, 0)$ which are tight up to a factor $\log d$. By Theorems 3.2 and 3.3 we have for $d \geq 2$ that $\mathrm{LC}(\mathcal{R}_n^d, r) = \infty$ if and only if $r \geq 1/(d+1)$. For $d = 1$ and $n \geq 3$ one can prove that $\mathrm{LC}(\mathcal{R}_n^1, r) = \infty$ if and only

if $r \geq 1/3$ [1]. Theorem 3.4 gives a poly$(d \log n)$ learning algorithm which is robust up to a fraction of noise of $1/(2d+1)$. Apparently there is a trade-off between learning complexity and the maximal fraction of tolerable noise. E.g. the learning algorithm in [CM92] can be modified to yield an algorithm with learning complexity $O(d^2 \log n)$ which is robust up the a fraction of noise of $1/(16d^2)$.

If $n = 2$ then $\mathcal{R}_2^d$ is the class of conjunctions of d (negated) variables. Adopting Theorem 4.3.4 from [Lit89] one gets $\mathrm{LC}(\mathcal{R}_2^d, r) \leq \frac{Cd}{1-r(4d+1)}$ if $r < 1/(4d+1)$ where $C \approx 11$.

4 The algorithm LR (Proof of Theorem 3.4)

In this section we present a fast algorithm which is a substantial modification of the weighted majority algorithm [LW91]. The predictions of the next hypothesis for the $2d$ coordinates of the corners of the rectangle are calculated independently from each other. All possible values are weighted accordingly to the previous seen CEs. The next prediction is chosen such that the sums of weights corresponding to values less than and greater than the prediction are roughly proportional to 1 and $2d$ (This holds for the "upper right" coordinates. Greater and less must be interchanged for the "lower left" coordinates.). After receiving the next CE those weights which correspond to inconsistent coordinate values are multiplied by some β, $0 < \beta < 1$.

Formally denote the weights by

$$w_{ik}^-(q), w_{ik}^+(q), \quad i = 1, \ldots, d, \quad k = 0, \ldots, n-1, \quad q \geq 1,$$

set

$$W_i^-(q) = \sum_{k=0}^{n-1} w_{ik}^-(q), \; W_i^+(q) = \sum_{k=0}^{n-1} w_{ik}^+(q),$$

and let

$$H(q) = \prod_{i=1}^{d} \{h_i^-(q), \ldots, h_i^+(q)\},$$
$$x(q) = (x_1(q), \ldots, x_d(q)),$$

be the q-th hypothesis and the q-th CE.

Algorithm LR**:**

Initialization

- For all $i = 1, \ldots, d$, $k = 0, \ldots, n-1$ set $w_{ik}^-(1) := 1$, $w_{ik}^+(1) := 1$.
- Let $0 \leq r < 1/(2d+1)$ be the maximal fraction of noise. Set

$$\alpha := \frac{1 - r(2d+1)}{2d+1}, \quad \beta := 1 - \alpha.$$

[1] The difference between $d = 1$ and $d \geq 2$ is that for $d = 1$ it is more difficult to find a first point inside the rectangle (interval) than to determine the length of the sides, which is more difficult for $d \geq 2$.

Calculation of $H(q)$

- Select $h_i^-(q), h_i^+(q)$ such that

$$\sum_{k=h_i^-(q)+1}^{n-1} w_{ik}^-(q) \leq \frac{W_i^-(q)}{2d+1} \leq \sum_{k=h_i^-(q)}^{n-1} w_{ik}^-(q)$$

$$\sum_{k=0}^{h_i^+(q)-1} w_{ik}^+(q) \leq \frac{W_i^+(q)}{2d+1} \leq \sum_{k=0}^{h_i^+(q)} w_{ik}^+(q).$$

Update of weights

- If $x(q) \notin H(q)$, i.e. $x(q)$ is a positive CE, then choose <u>one</u> i such that $x_i(q) < h_i^-(q)$ or $x_i(q) > h_i^+(q)$.
  If $x_i(q) < h_i^-(q)$ then set $w_{ik}^-(q+1) := \beta w_{ik}^-(q)$ for all $k = h_i^-(q), \ldots, n-1$.
  If $x_i(q) > h_i^+(q)$ then set $w_{ik}^+(q+1) := \beta w_{ik}^+(q)$ for all $k = 0, \ldots, h_i^+(q)$.
- If $x(q) \in H(q)$, i.e. $x(q)$ is a negative CE, then <u>for all</u> $i = 1, \ldots, d$ do
  $w_{ik}^-(q+1) := \beta w_{ik}^-(q)$ for all $k = 0, \ldots, h_i^-(q)$,
  $w_{ik}^+(q+1) := \beta w_{ik}^+(q)$ for all $k = h_i^+(q), \ldots, n-1$.
- All other weights remain unchanged.

Observe that in the case of a positive CE the algorithm blames exactly one coordinate and updates the weights associated with not consistent values. In the case of a negative CE the hypothesis is too large. Since the algorithm does not know in which direction the rectangle should be shrunken, it "blames" all the coordinates as was done in [CM92], however the "punishment" is substantially different than theirs.

Remark 4.1 *Observe that the algorithm has to know an upper bound on the fraction of noise to calculate the update parameter β.*

Remark 4.2 *Using a balanced binary tree the hypothesis $H(q)$ and the updates of the weights can be calculated in $O(d \log q)$ time. Thus the entire algorithm takes*

$$O(\frac{d^4 \log n}{(1-r(2d+1))^2} \log \frac{d^3 \log n}{(1-r(2d+1))^2})$$

time, for fixed $r < 1/(2d+1)$.

To establish Theorem 3.4 we will prove

Lemma 4.3 *If $r < 1/(2d+1)$ then*

$$\mathrm{LC}_{\mathrm{LR}}(\mathcal{R}_n^d, r) \leq \frac{4d(2d+1)^2 \log n}{(1-r(2d+1))^2}. \tag{4.1}$$

4.1 Analysis and proof of Lemma 4.3

Let

$$T = \prod_{i=1}^{d} \{t_i^-, \ldots, t_i^+\}$$

be the target and $x(1), \ldots, x(Q)$ a sequence of CEs such that the number of noisy CEs is less than or equal to rQ. We have to prove that Q is bounded by the right hand side of (4.1). This is done by calculating lower and upper bounds on the weights.

We denote the number of positive and negative, correct and noisy CEs by

$$\begin{aligned} P^{(c)} &= |\{1 \le q \le Q : x(q) \notin H(q), x(q) \in T\}|, \\ P^{(n)} &= |\{1 \le q \le Q : x(q) \notin H(q), x(q) \notin T\}|, \\ N^{(c)} &= |\{1 \le q \le Q : x(q) \in H(q), x(q) \notin T\}|, \\ N^{(n)} &= |\{1 \le q \le Q : x(q) \in H(q), x(q) \in T\}|, \end{aligned}$$

and set

$$P = P^{(c)} + P^{(n)}, \quad N = N^{(c)} + N^{(n)},$$

$$W(q) = \prod_{i=1}^{d} W_i^-(q) W_i^+(q).$$

We calculate the effect of CE $x(q)$ on $W(q)$.

- If $x(q)$ is a positive CE then assume without loss of generality $x_i(q) < h_i^-(q)$. Thus by the choice of $h_i^-(q)$

$$\begin{aligned} W_i^-(q+1) &\le W_i^-(q)[1 - 1/(2d+1)] + \beta W_i^-(q)/(2d+1) \\ &\le W_i^-(q)[1 - (1-\beta)/(2d+1)] \end{aligned}$$

  and hence $W(q+1) \le W(q)[(1 - (1-\beta)/(2d+1)]$.

- If $x(q)$ is a negative CE then

$$W_i^-(q+1) \le W_i^-(q)[1 - \frac{2d}{2d+1}(1-\beta)]$$

  for all i and analogously for $W_i^+(q+1)$. Hence $W(q+1) \le W(q)[1 - \frac{2d}{2d+1}(1-\beta)]^{2d}$.

Thus

$$W(Q+1) \le n^{2d}[1 - (1-\beta)/(2d+1)]^P \cdot [1 - \frac{2d}{2d+1}(1-\beta)]^{2dN}.$$

Now we calculate lower bounds on the weights of the target coordinates $w_{i,t_i^-}^-(q), w_{i,t_i^+}^+(q)$. Let

$$W_T(q) = \prod_{i=1}^{d} (w_{i,t_i^-}^-(q) w_{i,t_i^+}^+(q))$$

and let us calculate the variation of $W_T(q)$.

- If $x(q)$ is a correct positive CE then $x_i(q) < h_i^-(q)$ implies $t_i^- < h_i^-(q)$ and therefore $W_T(q+1) = W_T(q)$.
- If $x(q)$ is a noisy positive CE then $W_T(q+1) \geq \beta W_T(q)$.
- If $x(q)$ is a correct negative CE then there is one coordinate such that $h_i^-(q) \leq x_i(q) < t_i^-$ or $t_i^+ < x_i(q) \leq h_i^+(q)$. Thus $W_T(q+1) \geq \beta^{2d-1} W_T(q)$.
- If $x(q)$ is a noisy negative CE then $W_T(q+1) \geq \beta^{2d} W_T(q)$.

Putting things together we get

$$\begin{aligned} \beta^{P^{(n)}+(2d-1)N^{(c)}+2dN^{(n)}} &\leq W_T(Q+1) \leq W(Q+1) \\ &\leq n^{2d}[1-(1-\beta)/(2d+1)]^P \cdot [1-\frac{2d}{2d+1}(1-\beta)]^{2dN}. \end{aligned}$$

Since $-x - x^2 - x^3/3 - x^4/2 \leq \log(1-x) \leq -x - x^2/2 - x^3/3$ if $0 \leq x \leq 1/3$ we have

$$\begin{aligned} 0 &\leq -[P^{(n)} + (2d-1)N^{(c)} + 2dN^{(n)}] \log \beta + 2d \log n + P \log[1-(1-\beta)/(2d+1)] \\ &\quad +2dN \log[1 - \frac{2d}{2d+1}(1-\beta)] \\ &\leq [P^{(n)} + (2d-1)N^{(c)} + 2dN^{(n)}] \cdot [\alpha + \alpha^2/2 + \alpha^3/3 + \alpha^4/2] + 2d \log n \\ &\quad +P[-\alpha/(2d+1)] + 2dN[-\frac{2d}{2d+1}\alpha - (\frac{2d}{2d+1})^2\alpha^2/2 - (\frac{2d}{2d+1})^3\alpha^3/3] \\ &\leq 2d \log n + \alpha[P^{(n)} + (2d-1)N^{(c)} + 2dN^{(n)} - P/(2d+1) - \frac{4d^2}{2d+1}N] \\ &\quad +\frac{\alpha^2}{2}[P^{(n)} + (2d-1)N^{(c)} + 2dN^{(n)} - \frac{8d^3}{(2d+1)^2}N] \\ &\quad +\frac{\alpha^3}{3}[P^{(n)} + (2d-1)N^{(c)} + 2dN^{(n)} - \frac{16d^4}{(2d+1)^3}N] \\ &\quad +\frac{\alpha^4}{2}[P^{(n)} + (2d-1)N^{(c)} + 2dN^{(n)}] \\ &\leq 2d \log n + \alpha[-Q/(2d+1) + P^{(n)} + N^{(n)}] \\ &\quad +\frac{\alpha^2}{2}[-Q/(2d+1) + P^{(n)} + N^{(n)} + P/(2d+1) + \frac{4d^2}{(2d+1)^2}N] \\ &\quad +\frac{\alpha^3}{3}[-Q/(2d+1) + P^{(n)} + N^{(n)} + P/(2d+1) + \frac{8d^3}{(2d+1)^3}N] \\ &\quad +\frac{\alpha^4}{2}[P^{(n)} + (2d-1)N^{(c)} + 2dN^{(n)}] \\ &\leq 2d \log n + \alpha[-1/(2d+1) + r]Q \\ &\quad +\frac{\alpha^2}{2}[0 + P/(2d+1) + \frac{4d^2}{(2d+1)^2}N] + \frac{\alpha^3}{3}[0 + P/(2d+1) + \frac{8d^3}{(2d+1)^3}N] + \frac{\alpha^4}{2}[2dQ] \\ &\leq 2d \log n - \alpha^2 Q[1 - \frac{2d^2}{(2d+1)^2} - \max\{\frac{1}{3(2d+1)^2}, \frac{8d^3}{3(2d+1)^4}\} - \frac{d}{(2d+1)^2}] \\ &\leq 2d \log n - \alpha^2 Q/2 \end{aligned}$$

which gives
$Q \leq 4d(2d+1)^2 \log n/[1-r(2d+1)]^2$.

5 A sharp bound on the maximal tolerable fraction of noise

5.1 Proof of Theorem 3.2

We give an adversary strategy for the subclass of rectangles $\mathcal{R}_n^d(\mathbf{0})$ given by

$$\mathcal{R}_n^d(\mathbf{0}) = \{\prod_{i=1}^{d}\{0,\ldots,c_i^+\} : 0 \leq c_i^+ \leq n-1\}.$$

This is the class of rectangles with lower left corner $c^- = \mathbf{0} = (0,\ldots,0)$. Since this is additional information for the learner we have[2]

$$\mathrm{LC}(\mathcal{R}_n^d(\mathbf{0}), r) \leq \mathrm{LC}(\mathcal{R}_n^d, r).$$

For $i = 1,\ldots,d$ let

$$e^{(i)} = (\delta_{i1},\ldots,\delta_{id}), \quad \delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases},$$

be the i-th base vector and let

$$e^{(0)} = (1,\ldots,1).$$

Now the CEs are constructed as

$$x(q) = \begin{cases} e^{(0)} & \text{if } \forall i \in \{1,\ldots,d\} : e^{(i)} \in H(q) \\ e^{(i)} & \text{for some } e^{(i)} \notin H(q) \text{ otherwise.} \end{cases}$$

Clearly $e^{(i)} \in H(q)$ for all $i = 1,\ldots,d$ implies $e^{(0)} \in H(q)$ and therefore $e^{(0)}$ is a negative CE where $e^{(i)}$, $i = 1,\ldots,d$ are positive CEs.

To prove theorem 3.2 the following lemma is sufficient.

Lemma 5.1 *Assume the CEs are chosen as above. Then*

$$\forall q \geq 0\ \exists T(q) \in \mathcal{R}_n^d(\mathbf{0}) : |\{1 \leq p \leq q : x(p) \notin H(p) \triangle T(q)\}| \leq \frac{q}{d+1}.$$

Proof. *For $i = 0,\ldots,d$ let $\eta^i(q) = |\{1 \leq p \leq q : x(p) = e^{(i)}\}|$. Choose some $i \in \{0,\ldots,d\}$ with $\eta^i(q) \leq q/(d+1)$ and set for all $j = 1,\ldots,d$*

$$t_j^+(q) = \begin{cases} 1 & \text{if} \quad i = 0 \\ 1 - e_j^{(i)} & \text{if} \quad i \neq 0 \end{cases}.$$

Then $T(q)$ is consistent with all CEs $x(p) \neq e^{(i)}$ and the lemma follows. □

[2]This is not completely trivial since in general $\mathcal{T}_1 \subseteq \mathcal{T}_2$ does not imply $\mathrm{LC}(\mathcal{T}_1, r) \leq \mathrm{LC}(\mathcal{T}_2, r)$ because the learner's hypotheses are also restricted to $\mathcal{T}_1$. But clearly $\mathrm{LC}(\mathcal{T}_1, \mathcal{T}_2, r) \leq \mathrm{LC}(\mathcal{T}_2, \mathcal{T}_2, r)$ holds which in our case gives the statement since a "clever" learner will always include $\mathbf{0}$ in its hypotheses (otherwise it would receive $\mathbf{0}$ as a CE).

5.2 Proof of Theorem 3.3

In this section we present a conservative algorithm which is able to tolerate any fraction of noise less than $1/(d+1)$. The update of the coordinates $h_i^-(q), h_i^+(q)$ of the algorithm's hypothesis depends on the last CE $x(q)$ and the "median" of the positive CEs. Each coordinate is modified by at most 1. Let $P^{(c)}(q)$, $P^{(n)}(q)$, $N^{(c)}(q)$, $N^{(n)}(q)$, $P(q)$, $N(q)$ be the number of positive and negative, correct and noisy CEs up the q-th CE as defined in section 4.1. Then the median $m(q) = (m_1(q), \ldots, m_d(q))$ must satisfy

$$|\{1 \le p \le q : x(p) \notin H(q), x_i(p) \le m_i(q)\}| \ge P(q)/2,$$

$$|\{1 \le p \le q : x(p) \notin H(q), x_i(p) \ge m_i(q)\}| \ge P(q)/2$$

for all $i = 1, \ldots, d$. Among the $m(q)$ satisfying this condition the algorithm may choose arbitrarily. Intuitively the above condition says that the median is somewhere in the middle of the positive CEs, and therefore that even if some of them are noisy, $m(q)$ should be part of the target. Hence the median can be used to decide in which directions the hypothesis should be shrunken if a negative CE is received (instead of "shrinking" in all $2d$ possible directions as is done by the update rule of algorithm LR).

Algorithm OPT:

Initialization

- For all $i = 1, \ldots, d$ set $h_i^-(1) := 1, h_i^+(q) := 0$.

Update

- If $x(q) \notin H(q)$ then choose <u>one</u> i such that $x_i(q) < h_i^-(q)$ or $h_i^+(q) < x_i(q)$.
  If $x_i(q) < h_i^-(q)$ then set $h_i^-(q+1) := h_i^-(q) - 1$
  else (thus $x_i(q) > h_i^+(q)$) set $h_i^+(q+1) := h_i^+(q) + 1$.
- If $x(q) \in H(q)$ then <u>for all</u> $i = 1, \ldots, d$ do:
  If $x_i(q) < m_i(q)$ then $h_i^-(q+1) := h_i^-(q) + 1$
  else (thus $x_i(q) \ge m_i(q)$) $h_i^+(q+1) := h_i^+(q) - 1$.
- All other coordinates remain unchanged.

Remark 5.2 *After a positive CE the hypothesis is enlarged by 1 in exactly one direction, and after a negative CE the hypothesis is reduced by 1 in d directions. To decide whether the upper or the lower coordinate should be modified, the algorithm uses the median of the positive CEs. Observe that the algorithm* LR *does not use the median because for algorithm* LR *we were not able to prove something like Proposition 5.7 below. Therefore the fraction of noise must be less than* $1/(2d+1)$ *for algorithm* LR *but must be only less than* $1/(d+1)$ *for algorithm* OPT.

Remark 5.3 *Observe that algorithm* OPT *does not have to know a bound on the fraction of noise.*

To establish Theorem 3.3 we prove

Lemma 5.4 *If* $d \geq 2$ *and* $r < 1/(d+1)$ *then* $\mathrm{LC}_{\mathrm{OPT}}(\mathcal{R}_n^d, r) \leq \frac{dn}{1-r(d+1)}$.

5.3 Analysis and proof of Lemma 5.4

Let T be the target and $x(1), \ldots, x(Q)$ a sequence of CEs such that $|\{1 \leq q \leq Q : x(q) \notin H(q) \triangle T\}| \leq rQ$. It is easy to see that the update rules of the algorithm are meaningful provided that $m(q)$ is an element of the target T, but there is some problem if $m(q) \notin T$. Therefore we distinguish between two learning phases. During the first phase $m(q)$ will occasionally not be in T, during the second phase we will always have $m(q) \in T$. Note that the algorithm does not know whether it is in phase 1 or 2. This is only recognized by an external observer. Let

$$Q_1 = \max\{1 \leq q \leq Q : m(q) \notin T\}, \quad Q_2 = Q - Q_1.$$

If $m(q) \in T$ for all q we set $Q_1 = 0$. Let $P_1^{(c)} = P^{(c)}(Q_1)$, $P_1^{(n)} = P^{(n)}(Q_1)$, $N_1^{(c)} = N^{(c)}(Q_1)$, $N_1^{(n)} = N^{(n)}(Q_1)$, $P_1 = P_1^{(c)} + P_1^{(n)}$, $N_1 = N_1^{(c)} + N_1^{(n)}$, $P_2^{(c)} = P^{(c)}(Q) - P_1^{(c)}$, $P_2^{(n)} = P^{(n)}(Q) - P_1^{(n)}$, $N_2^{(c)} = N^{(c)}(Q) - N_1^{(c)}$, $N_2^{(n)} = N^{(n)}(Q) - N_1^{(n)}$, $P_2 = P_2^{(c)} + P_2^{(n)}$, $N_2 = N_2^{(c)} + N_2^{(n)}$.

Now we bound Q_1 and Q_2 in respect to the number of noisy CEs in the first and second learning phase, respectively.

Proposition 5.5 $N_1 \leq P_1/d$.

Proof. *If* $x(q)$ *is a negative CE then clearly* $H(q) \neq \emptyset$. *Furthermore a negative CE reduces the sum* $\sum_{i=1}^{d}[h_i^+(q) - h_i^-(q)]$ *by* d *and a positive CE enlarges it by 1. Since* $\sum_{i=1}^{d}[h_i^+(1) - h_i^-(1)] = -d$ *and* $H(q) \neq \emptyset$ *implies that the sum is not negative, it follows that* $0 \leq -d + P(q-1) - dN(q-1)$ *if* $x(q)$ *is a negative CE. Hence* $N(q) = N(q-1) + 1 \leq P(q-1)/d = P(q)/d$. □

Proposition 5.6 $P_1^{(c)} \leq P_1^{(n)}$.

Proof. *If more than a half of the positive CEs have been correct then* $m(Q_1)$ *would be an element of the target.* □

Proposition 5.7 $P_1^{(n)} + N_1^{(n)} \geq Q_1/(d+1)$ *for* $d \geq 2$.

Proof. *By Propositions 5.5 and 5.6 we have* $Q_1 \leq P_1 + N_1 \leq (1+1/d)P_1 \leq 2(1+1/d)P_1^{(n)} \leq (d+1)P_1^{(n)}$. *Observe that Proposition 5.7 does not hold for* $d = 1$. □

Proposition 5.8

$$Q_2 \leq dn + (d+1)(P_2^{(n)} + N_2^{(n)}).$$

Proof. *Remember that the median is element of the target. We observe how the following sum varies depending on the CEs:*

$$\zeta(q) = \sum_{i=1}^{d} \max\{0, t_i^+ - h_i^+(q)\} + \sum_{i=1}^{d} \max\{0, h_i^-(q) - t_i^-\}.$$

- *Since a correct positive CE reduces the distance of one coordinate we have* $\zeta(q+1) = \zeta(q) - 1$.
- *For a noisy positive CE we have* $\zeta(q+1) \leq \zeta(q)$.
- *A correct negative CE corrects one noisy positive CE and enlarges at most* $d-1$ *distances. Thus* $\zeta(q+1) \leq \zeta(q) + d - 1$.
- *For a noisy negative CE we have* $\zeta(q+1) \leq \zeta(q) + d$.

Since $\zeta(Q_1) \leq dn$ *we get* $0 \leq \zeta(Q_2+1) \leq dn - P_2^{(c)} + (d-1)N_2^{(c)} + dN_2^{(n)}$. *Furthermore* $N_2^{(c)} \leq P_2^{(n)}$ *since a correct negative CE always corrects a noisy positive CE. Thus*

$$\begin{aligned} & Q_2 \\ &= P_2^{(c)} + P_2^{(n)} + N_2^{(c)} + N_2^{(n)} \\ &\leq dn + (d-1)N_2^{(c)} + dN_2^{(n)} + P_2^{(n)} + N_2^{(c)} + N_2^{(n)} \\ &\leq dn + (d+1)P_2^{(n)} + (d+1)N_2^{(n)}. \end{aligned}$$

□

Now Lemma 5.4 follows immediately from Propositions 5.7 and 5.8.

6 A lower bound on the learning complexity in the noise-free case (Proof of Theorem 3.1)

We give an adversary strategy for the subclass of rectangles $\mathcal{R}_n^d(\mathbf{0})$. The adversary dynamically updates a cube of possible "upper right" corners. It tries to find CEs such that the volume of this cube decreases by a factor of at most $1-a$ where a is close to 0. Since a positive CE may eliminate at most a fraction a from the cube, the adversary has to give a negative CE if the learner's hypothesis is big. But after a negative CE the set of possible corners is not a cube. Thus after a negative CE the adversary gives $(d-1)$ positive CEs such that the remaining set again is a cube. As can be seen from Figure 1 (some generalization

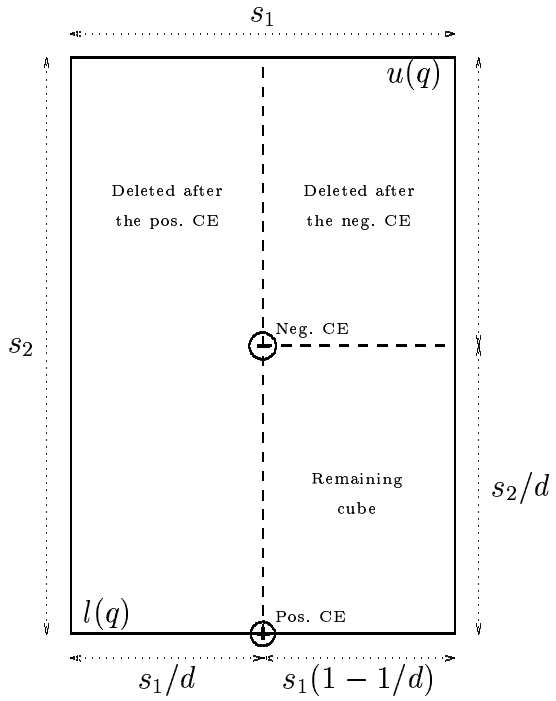

Figure 1: The effect of positive and negative CEs.

for large d is needed) for $a = 1/d$ the volume of the remaining cube is $\frac{1}{d}(1-\frac{1}{d})^{d-1}$ times the volume of the original cube.

A formal treatment of this construction reveals another technical difficulty: We are operating in a discrete domain, hence we have to argue about the number of points in the cube instead about its volume. We overcome this difficulty by arguing only about cubes with relatively large edges such that the volume very closely approximates the number of points. This is done by considering only coordinates with large edges.

We denote the adversary's cube by $\prod_{j=1}^{d}[l_j(q), u_j(q)] \subseteq \{0,\ldots,n-1\}^d$ and the set of "active" coordinates by $D(q) \subseteq \{1,\ldots,d\}$. The hypotheses of the learner are denoted by $h^+(q)$, the CEs by $x(q)$. We assume without loss of generality that the learner's hypotheses are consistent with all previously seen CEs.

Adversary strategy $\mathcal{S}$:

1. Set $q := 1$ and for all $j = 1,\ldots,d$ let $l_j(1) := 0$, $u_j(1) := n-1$.
2. $D(q) := \{i : u_i(q) - l_i(q) \geq d\}$.
3. If $D(q) = \emptyset$ then STOP.
4. For all $j = 1,\ldots,d$ set
   $y_j(q) := \lfloor l_j(q) + (u_j(q) - l_j(q))/d \rfloor$.
5. If $\exists i \in D(q) : h_i^+(q) < y_i(q)$ then fix that i and set for all $j = 1,\ldots,d$

$$x_j(q) := \begin{cases} y_i(q) & \text{if} \quad j = i \\ 0 & \text{if} \quad j \neq i \end{cases}$$

(This is a positive CE.)

$$l_j(q+1) := \begin{cases} y_i(q) & \text{if} \quad j = i \\ l_j(q) & \text{if} \quad j \neq i \end{cases}$$

$u_j(q+1) := u_j(q)$

$q := q+1$, GOTO 2

6. If $\forall i \in D : h_i^+(q) \geq y_i(q)$ then

   (a) $$x_j(q) := \begin{cases} y_j(q) & \text{if} \quad j \in D(q) \\ 0 & \text{if} \quad j \notin D(q) \end{cases}$$

   (This is a negative CE.)

   (b) For $p = 1, \ldots, |D(q)| - 1$ choose some $i \in D(q)$ with $h_i^+(q+p) < y_i(q)$ and set for all $j = 1, \ldots, d$

   $$x_j(q+p) := \begin{cases} y_i(q) & \text{if} \quad j = i \\ 0 & \text{if} \quad j \neq i \end{cases}$$

   (This is a positive CE. For $p = 1, \ldots, |D(q)| - 1$ such an i exists by the assumption that the learner's hypotheses are consistent.)

   (c) For the $i \in D(q)$ with $l_i(q) < y_i(q)$ *(there exists exactly one)* set

   $$l_j(q + |D(q)|) := \begin{cases} y_j(q) & \text{if} \quad j \in D(q),\ j \neq i \\ l_i(q) & \text{if} \quad j = i \\ l_j(q) & \text{if} \quad j \notin D(q) \end{cases}$$

   $$u_j(q + |D(q)|) := \begin{cases} u_j(q) & \text{if} \quad j \in D(q),\ j \neq i \\ y_i(q) - 1 & \text{if} \quad j = i \\ u_j(q) & \text{if} \quad j \notin D(q) \end{cases}$$

   (d) $q := q + |D(q)|$, GOTO 2.

Assume the adversary stops with $q = Q$. It is easy to see from the definition of the adversary strategy that there is a target $t^+ \in \prod_{j=1}^d [l_j(Q), u_j(Q)]$ consistent with all given CEs. To bound the number of CEs $Q - 1$ from below we distinguish between step 5 and 6.

Proposition 6.1 *If the condition of step 5 is satisfied for q then $\prod_{j=1}^d [u_j(q+1) - l_j(q+1) + 1] \geq (1 - 1/d) \prod_{i=j}^d [u_j(q) - l_j(q) + 1]$.*

Proof. *This follows from $u_i(q+1) - l_i(q+1) + 1 = u_i(q) - \lfloor l_i(q) + (u_i(q) - l_i(q))/d \rfloor + 1 \geq (u_i(q) - l_i(q))(1 - 1/d) + 1$.* □

Proposition 6.2 *If the condition of step 6 is satisfied for q then $\prod_{j=1}^d [u_j(q + |D(q)|) - l_j(q + |D(q)|) + 1] \geq \frac{1}{2d}(1 - 1/d)^{|D(q)|-1} \prod_{j=1}^d [u_j(q) - l_j(q) + 1]$.*

Proof. *For all $i \in D(q)$ which appear in step 6b we have $u_i(q + |D(q)|) - l_i(q + |D(q)|) + 1 \geq (1 - 1/d)[u_i(q) - l_i(q) + 1]$. For the i of step 6c we have $u_i(q + |D(q)|) - l_i(q + |D(q)|) + 1 \geq \lfloor l_i(q) + (u_i(q) - l_i(q))/d \rfloor - l_i(q) \geq (u_i(q) - l_i(q) + 1)/(2d)$.* □

Lemma 6.3 *If $n \geq d^2$, $d \geq 2$, and the adversary strategy $\mathcal{S}$ stops with $q = Q$ then*

$$Q \geq \frac{d^2 \log n}{40 \log d}.$$

Proof. *Let $Q_0 = \min\{q \leq Q : |D(q)| < d/2\}$. Thus for $q < Q_0$ we have $d/2 \leq |D(q)| \leq d$. Since $(1-1/d)^{|D(q)|} \geq \frac{1}{2d}(1-1/d)^{|D(q)|-1}$ we have by Propositions 6.1 and 6.2 and by induction on q that*

$$\begin{aligned}
&\prod_{j=1}^{d}[u_j(Q_0) - l_j(Q_0) + 1] \\
&\quad \geq \prod_{j=1}^{d}[u_j(1) - l_j(1) + 1] \\
&\qquad \cdot \left[\frac{1}{2d}(1-1/d)^{d-1}\right]^{Q_0/|D(Q_0-1)|} \\
&\quad \geq n^d \left[\frac{1}{2d}(1-1/d)^{d-1}\right]^{2Q_0/d}
\end{aligned}$$

and

$$\begin{aligned}
\prod_{j=1}^{d}[u_j(Q_0) - l_j(Q_0) + 1] &\leq n^{|D(Q_0)|} d^{d-|D(Q_0)|} \\
&\leq n^{d/2} d^{d/2}.
\end{aligned}$$

Since $\log(1-x) \geq -2x$ for $0 \leq x \leq 1/2$ we get

$$2Q_0/d \geq \frac{\frac{d}{2}[\log n - \log d]}{\log(2d) + 2(d-1)/d} \geq \frac{d \log n}{20 \log d}.$$

□

Proof of Theorem 3.1. For $n \geq d^2$ the first statement of the theorem follows from Lemma 6.3. Since the second statement of the theorem yields $\mathrm{LC}(\mathcal{R}_n^d, 0) = \Omega(d^2)$ for $n \geq d$ and $d^2 \geq \frac{d^2 \log n}{2 \log d}$ for $d \leq n \leq d^2$, the first statement holds also in this case.

The second statement of the theorem can be established much more easily. We reuse the adversary strategy $\mathcal{S}$, replacing only step 2 by

$$D(q) := \{i : u_i(q) - l_i(q) \geq 1\}$$

and step 4 by

$$y_i(q) := l_i(q) + 1.$$

Analogously to Propositions 6.1, 6.2 we get

$$\sum_{j=1}^{d}[u_j(q+1) - l_j(q+1) + 1]$$
$$\geq \sum_{i=j}^{d}[u_j(q) - l_j(q) + 1] - 1,$$
$$\sum_{j=1}^{d}[u_j(q+|D(q)|) - l_j(q+|D(q)|) + 1]$$
$$\geq \sum_{j=1}^{d}[u_j(q) - l_j(q) + 1] - (n-1) - (|(D(q)| - 1),$$

respectively. A modification of the proof of Lemma 6.3 then gives

$$dn - (2Q_0/d)(n+d-2)$$
$$\leq \textstyle\sum_{j=1}^{d}[u_j(Q_0) - l_j(Q_0) + 1] \leq dn/2 + d/2$$

and thus

$$2Q_0/d \geq \frac{d(n-1)/2}{n+d-2} \geq \frac{n}{8}.$$

□

7 Proof of Theorem 2.3

The following trivial algorithm learns any target from the target class $\mathcal{T} = \{T_1, \ldots, T_{|\mathcal{T}|}\}$ if $r < 1/|\mathcal{T}|$: The hypotheses of the algorithm are

$$H_1 = T_1, H_2 = T_2, \ldots, H_{|\mathcal{T}|} = T_{|\mathcal{T}|},$$
$$H_{|\mathcal{T}|+1} = T_1, \ldots, H_q = T_{(q \bmod |\mathcal{T}|)+1}, \ldots$$

Let T_i be the target and $x_1, \ldots, x_Q$ a sequence of CEs with $|\{1 \leq q \leq Q : x_q \notin T_i \triangle H_q\}| \leq rQ$. Since $|\{1 \leq q \leq Q : H_q = T_i\}| \geq \frac{Q+1}{|\mathcal{T}|} - 1$ we get $\frac{Q+1}{|\mathcal{T}|} - 1 \leq rQ$ and $Q \leq \frac{|\mathcal{T}|-1}{1-r|\mathcal{T}|}$.

To prove the second statement of the theorem let the adversary construct only negative CEs by

$$x_q = i \quad \text{if} \quad H_q = \{i\}.$$

If we set $\eta_i(q) = |\{1 \leq p \leq q : x_q = i\}|$ then for all $q \geq 0$ there is a target $\{i\}$ with $\eta_i(q) \leq q/n$.

8 Conclusion

We investigated the implications of noise in the equivalence query model for the target class of d-dimensional rectangles. Assuming a noise model where only the fraction of noisy examples

is bounded, we archived a characterization of the maximal tolerable noise rate. Furthermore we presented a reasonably efficient, robust learning algorithm. Besides some results about general target classes, we also obtained a lower bound on the learning complexity of rectangles in the noise-free case.

9 Acknowledgement

I want to thank Wolfgang Maass for drawing my attention to the problem of noise-robust on-line learning of rectangles. I am also very grateful to Phil Long for valuable suggestions and discussions.

References

[AL88] D. Angluin and P.D. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

[Ang88] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

[CM92] Zhixiang Chen and Wolfgang Maass. On-line learning of rectangles. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 16–28. ACM Press, 1992.

[DGW92] Aditi Dhagat, Peter Gacs, and Peter Winkler. On playing "Twenty Questions" with a liar. In *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 16–22, 1992.

[KL88] M. Kearns and M. Li. Learning in the presence of malicious errors. *Proceedings of the 20th ACM Symposium on the Theory of Computation*, pages 267–279, 1988.

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

[Lit89] N. Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms*. PhD thesis, UC Santa Cruz, 1989.

[LW91] N. Littlestone and M.K. Warmuth. The weighted majority algorithm. Technical Report UCSC-CRL-91-28, UC Santa Cruz, 1991. An extended abstract appeared in: *Proceedings of the 30th Annual Symposium on the Foundations of Computer Science.*.

[MT89] W. Maass and G. Turán. On the complexity of learning from counterexamples. In *30th Annual IEEE Symposium on Foundations of Computer Science*, pages 262–267, 1989.

[MT91] Wolfgang Maass and György Turan. Algorithms and lower bounds for on-line learning of geometrical concepts. IIG-Report 316, Technische Universität Graz, TU Graz, Austria, 1991.

[MT92] Wolfgang Maass and György Turan. Lower bound methods and separation results for on-line learning models. *Machine Learning*, pages 107–145, 1992.

ECCC ISSN 1433-8092
http://www.eccc.uni-trier.de/eccc
ftp://ftp.eccc.uni-trier.de/pub/eccc
ftpmail@ftp.eccc.uni-trier.de, subject 'help eccc'