Noise-tolerant Learning Near the Information-theoretic Bound

Nicolò Cesa-Bianchi
DSI, Università di Milano
Via Comelico 39,
I-20135 Milano, Italy
cesabian@dsi.unimi.it

Eli Dichterman
Department of Computer Science
Technion
Haifa 32000, Israel
eli@cs.technion.ac.il

Paul Fischer **Hans Ulrich Simon**
Lehrstuhl Informatik II
Universität Dortmund
D-44221 Dortmund, Germany
{paulf,simon}@goedel.informatik.uni-dortmund.de

Abstract

We investigate the sample size necessary for PAC learning in the presence of malicious noise, a type of adversarial noise model introduced by Kearns and Li. We prove the first nontrivial sample complexity lower bound in this model by showing that order of $\varepsilon/\Delta^2 + d/\Delta$ examples are necessary for PAC learning any target class of $\{0,1\}$-valued functions of VC dimension d, where ε is the desired accuracy and $\eta = \varepsilon/(1+\varepsilon) - \Delta$ the malicious noise rate (it is well known that any nontrivial target class cannot be PAC learned with accuracy ε and malicious noise rate $\eta \geq \varepsilon/(1+\varepsilon)$, this irrespective to sample complexity.)

This result cannot be significantly improved in general. In fact, we show a learning algorithm for the same noise model that, for each d, learns the class of all subsets of d elements using a number of examples of the same order as that proven necessary by our lower bound (disregarding logarithmic factors). In contrast, we show that to learn any target class of VC-dimension d in the presence of a high malicious noise rate η (i.e. $\Delta = o(\varepsilon)$), the popular strategy choosing any hypothesis that minimizes disagreements on the sample needs $\Omega(d\varepsilon/\Delta^2)$ examples. This implies that, for high noise rate and for any target class of VC-dimension large enough, there are distributions over the target domain where the minimum disagreement strategy is outperformed by our algorithm.

STOC'96, Philadelphia PA,USA
© 1996 ACM 0-89791-785-5/96/05..$3.50

1 Introduction

There are two main extensions of the basic PAC learning framework that take noise into account. The first one is the *classification* noise model [AL88, Lai88, Kea93, AD93], where the boolean value of the target function is independently flipped with fixed probability in each example given to the learner. A second extension, less benign and perhaps more realistic than classification noise, was later introduced by Kearns and Li [KL93] and appropriately called *malicious* noise. In this model, each example of the target function given to the learner is independently replaced, with fixed probability η, by an adversarially chosen example (which may or may not be consistent with the target.) The sample size necessary and sufficient for learning classes of $\{0,1\}$-valued functions (also known as the "sample complexity" of the class) is quite well understood for both the noise-free PAC model and the PAC model with classification noise. Matching upper and lower bounds (up to logarithmic factors) for the noise-free model are given in [EHKV89, BEHW89, STAB93]. For learning in the presence classification noise, the upper bound of Laird [Lai88] is met by the lower bound of Simon [Sim93], again up to logarithmic factors. However, there is no such a satisfactory analysis for learning models where the sample given to the learner is corrupted by malicious noise.

As shown by Kearns and Li, PAC learning in the presence of malicious noise is intrinsically harder than in the presence of independent classification noise. Two target functions that differ on at least one domain point whose probability is ε can be made statistically indistinguishable by a malicious noise rate larger or equal than $\varepsilon/(1+\varepsilon)$, thus forbidding ε-accurate PAC learning on information-theoretic grounds, irrespective to sample size and to the learner's computational power. In contrast, it has been shown that choosing any hypothesis that has the fewest disagreements on the input sample is sufficient for PAC learning in presence of independent classification noise with rate arbitrarily close to 1/2 [AL88]. Using results from [AST94], it is not hard to show that this minimum disagreement strategy is a PAC learning algorithm also in the presence of a malicious noise rate $\eta = \varepsilon/(1+\varepsilon) - \Delta$, with a sufficient sample size of order $d\varepsilon/\Delta^2$ (disregarding logarithmic factors), where d is the VC-dimension of the hypothesis class and ε is the desired accuracy. If $\Delta = \Omega(\varepsilon)$, the above sample size reduces to d/ε, which is of the same order as the sample size sufficient for PAC learning in the noise-free case. As order of d/ε examples are necessary for learning without noise, we already have a tight estimate (within logarithmic factors) of the sample complexity for small rates of malicious noise.

In this paper we study the behaviour of the sample complexity in the case of a high malicious noise rate, i.e. a malicious noise rate arbitrarily close to the information-theoretic upper bound $\varepsilon/(1+\varepsilon)$. We prove the first nontrivial lower bound in this model by showing that at least order of $\varepsilon/\Delta^2 + d/\Delta$ examples are needed to PAC learn, with accuracy ε and tolerating a malicious noise rate $\eta = \varepsilon/(1+\varepsilon) - \Delta$, any class of $\{0,1\}$-valued functions of VC-dimension d. Our proof combines, in an original way, techniques from [Sim93, KL93, EHKV89] and uses some new estimates of the tails of the binomial distribution that may be of independent interest. We then prove that this lower bound cannot be improved in general. Namely, we show that there is an algorithm RMD (for Randomized Minimum Disagreement) that, for each d, learns the class $\mathcal{C}_d$ of all subsets of d elements using a noisy sample whose size is of the same order as the size of our lower bound (up to logarithmic factors.) Algorithm RMD uses a majority vote to decide the classification of those domain points which have a clear majority of one label, and tosses a fair coin to decide the classification of the remaining points. We also show a lower bound of order $d\varepsilon/\Delta^2$ for the sample size of the popular strategy of choosing any hypothesis that minimizes disagreements on the sample. This bound holds for any class of VC-dimension $d \geq 3$ and for any noise rate η such that $\varepsilon/(1+\varepsilon) - \eta = \Delta = o(\varepsilon)$. This implies that, for high noise rate η and for any target class of VC-dimension d large enough, there are distributions over the target domain where the minimum disagreement strategy is outperformed by algorithm RMD. To our knowledge, this is the first example of a natural PAC learning problem where choosing any minimum disagreement hypothesis from a fixed hypothesis class is provably worse, in terms of sample complexity, than a different learning strategy.

2 Basic definitions

We recall the definitions of PAC learning and PAC learning in presence of malicious noise of a given *target class* $\mathcal{C}$, where $\mathcal{C}$ is a set of $\{0,1\}$-valued functions C defined on some domain X. We call *instance* any $x \in X$ and *labeled instance* or *example* any pair $(x,y) \in X \times \{0,1\}$. In Valiant's PAC learning model [Val84], the learning algorithm, or learner, has access to a *noise-free oracle* returning on each call an example $(x, C(x))$, where $C \in \mathcal{C}$ is the target and the instance x is drawn from a distribution D on X. Both C and D are fixed in advance and unknown to the learner. In Kearns and Li's PAC model [KL93][1] with malicious noise the noise-free oracle is replaced by a *malicious oracle*. If the noise rate is η, on the t-th call the malicious oracle flips a coin with bias η for heads. If the outcome is heads, the oracle returns an example $(\hat{x}_t, \hat{y}_t)$ chosen from $X \times \{0,1\}$. If the outcome is tails, the oracle must behave exactly like the noise-free oracle returning the correctly labeled instance $(x_t, C(x_t))$, where C is the target and x_t is drawn from D. The malicious oracle's choice for the pair $(\hat{x}_t, \hat{y}_t)$ can depend, in an arbitrary way, on the current state of the learner and on the outcome of the oracle's previous random draws. In both PAC learning and PAC learning in the presence of malicious noise, after a polynomial number of calls to the oracle the learner must output an hypothesis H that with high probability is a close approximation of the target C. However, in the malicious model the learner receives examples corrupted by adversarial noise.

Formally, an algorithm A is said to *learn* a target class $\mathcal{C}$ in the PAC model if, for all distributions D on X, for all targets $C \in \mathcal{C}$, and for all $\varepsilon, \delta > 0$, after m calls to the noise-free oracle A outputs an hypothesis $H \in \mathcal{C}$ such that $D(H \neq C) < \varepsilon$ holds with probability at least $1-\delta$ with respect to the oracle's randomization, where $m = m(\varepsilon, \delta)$ is some polynomial in $1/\varepsilon$ and $\ln(1/\delta)$. We call ε the *accuracy* parameter and δ the *confidence* parameter. Similarly, an algorithm A is said to *learn* a target class $\mathcal{C}$ in the malicious PAC model with noise rate η if A learns $\mathcal{C}$ in the PAC model when the noise-free oracle is replaced by any malicious oracle for noise rate η We allow the number m of calls to the malicious oracle to depend polynomially also on $1/\Delta$, where $\Delta = \varepsilon/(1+\varepsilon) - \eta$. The reason for this choice will be made clear in a moment. When referred to the resources used by the learner, we will use the expressions "number of calls made to the oracle" and "sample size" interchangeably. We will occasionally use *randomized* learning algorithms that have a sequence of tosses of a fair coin as an additional input source. In this case the definition of PAC learning given above is modified so that $D(H \neq C) < \varepsilon$ must hold with probability at least $1-\delta$ also with respect to A's randomization.

In [KL93], it was shown that PAC learning (even with an unbounded number of calls to the malicious oracle) is not possible whenever the noise rate is close enough to ε. They prove that for any "nontrivial" target class $\mathcal{C}$, for each $\epsilon > 0$, and for each learning strategy A (even noncomputable) there is a target $C \in \mathcal{C}$ and a distribution D over the domain such that the hypothesis H output by A after *any* number of calls to the malicious oracle with noise rate $\eta \geq \varepsilon/(1+\varepsilon)$ satisfies $D(H \neq C) \geq \varepsilon$ with probability at least 1/2. Hence, it is reasonable to allow the sample size for learning in the presence of malicious noise to grow polynomially also as a function of $\Delta^{-1} = (\varepsilon/(1+\varepsilon) - \eta)^{-1}$, where ε is the desired accuracy and η is the malicious noise rate.

In addition to the usual asymptotical notations, let $\widetilde{O}(f)$ be equivalent to $\bigcup_{d \geq 0} O(f(\log f)^d)$ for some constant d.

3 Lower Bounds

This section presents three basic results concerning the sample size needed for PAC learning in the presence of malicious noise. Theorems 3.4 and 3.7 establish the general lower bound $\Omega(\eta/\Delta^2 + d/\Delta)$ that holds for any learning algorithm. Given the results of Section 4, this bound cannot be significantly improved. Theorem 3.8 presents the stronger lower bound $\Omega(d\varepsilon/\Delta^2)$ for the minimum disagreement strategy (and for a somewhat stronger malicious oracle).

We make use of the following definitions and facts from probability theory. Let $S_{N,p}$ be the random variable that counts the number of successes in N independent trials, each

[1] We use a 1-oracle variant of Kearns and Li's original 2-oracle learning model. All of our results could be translated to that model with minor differences.

trial with probability p of success. A real number s is called *median* of a random variable S if $\Pr\{S \leq s\} \geq 1/2$ and $\Pr\{S \geq s\} \geq 1/2$.

Fact 3.1 ([JS68]) *For all $0 \leq p \leq 1$ and all $N \geq 1$, the median of $S_{N,p}$ is $\lfloor Np \rfloor$ or $\lceil Np \rceil$.*
Thus, $\Pr\{S_{N,p} \leq \lceil Np \rceil\} \geq \frac{1}{2}$ and $\Pr\{S_{N,p} \geq \lfloor Np \rfloor\} \geq \frac{1}{2}$.

Fact 3.2 *Let $0 < p < 1$ and $q = 1 - p$. Then for all $N \geq 37/(pq)$,*

$$\Pr\left\{S_{N,p} \geq \lfloor Np \rfloor + \left\lfloor \sqrt{Npq - 1} \right\rfloor\right\} > \frac{1}{19} \qquad (1)$$

$$\Pr\left\{S_{N,p} \leq \lceil Np \rceil - \left\lfloor \sqrt{Npq - 1} \right\rfloor\right\} > \frac{1}{19}. \qquad (2)$$

The proof is given in Appendix A.

Fact 3.3 *For any random variable $S \in [0, N]$ with expectation αN, and for any $0 < \beta < \alpha \leq 1$,*
$$\Pr\{S \geq \beta N\} > (\alpha - \beta)/(1 - \beta).$$

Proof. It follows by setting $z = \Pr\{S \geq \beta N\}$ and solving

$$\begin{aligned} \alpha N &= E[S] = E[S \mid S < \beta N](1 - z) + E[S \mid S \geq \beta N]z \\ &< \beta N(1 - z) + Nz \end{aligned}$$

for z. □

Two $\{0,1\}$-valued functions C_0 and C_1 are called *disjoint* if there exists no $x \in X$ such that $C_0(x) = C_1(x) = 1$. A target class $\mathcal{C}$ is called *trivial* if any two targets $C_0, C_1 \in \mathcal{C}$ are either identical or disjoint. Kearns and Li have shown in [KL93] that nontrivial target classes cannot be PAC learned with accuracy ε if the malicious noise rate η is larger or equal than $\varepsilon/(1+\varepsilon)$. The proof is based on the statistical indistinguishability of two targets C_0 and C_1 that differ on some domain point x which has probability ε, but coincide on all other points with nonzero probability. The malicious oracle will present x with the false label with probability $\eta_0 = \varepsilon/(1+\varepsilon)$. Hence, x appears with the true label with probability $(1 - \eta_0)\varepsilon$. As $(1 - \eta_0)\varepsilon = \eta_0$, there is no chance to distinguish between C_0 and C_1.

Our first lower bound is based on a similar reasoning: For $\eta < \eta_0$, the targets C_0 and C_1 can be distinguished, but as η approaches η_0, the discrimination task becomes arbitrarily hard. These ideas are made precise in the proof of the following result.

Theorem 3.4 *For any nontrivial target class $\mathcal{C}$, any $0 < \varepsilon < 1$, $0 < \delta \leq 1/38$, and $0 < \Delta = o(\varepsilon)$, the sample size needed for PAC learning $\mathcal{C}$ with accuracy ε, confidence δ, and tolerating malicious noise rate $\eta = \varepsilon/(1+\varepsilon) - \Delta$, is greater than*
$$\frac{9\eta(1-\eta)}{37\Delta^2} = \Omega\left(\frac{\eta}{\Delta^2}\right).$$

Proof. For each nontrivial target class $\mathcal{C}$ there exist two points $x_0, x_1 \in X$ and two targets C_0, C_1 such that $C_0(x_0) = C_1(x_0) = 1$, $C_0(x_1) = 0$, and $C_1(x_1) = 1$. Let the distribution D be such that $D(x_0) = 1 - \varepsilon$ and $D(x_1) = \varepsilon$. We will use a malicious oracle that, with probability η, corrupts the current example by replacing it with x_1 labeled incorrectly. Let A be a (possibly) randomized learning algorithm for $\mathcal{C}$ that demands sample size $m = m(\varepsilon, \delta, \eta)$. Consider the following random experiment:

1. Flip a fair coin to select target $C \in \{C_0, C_1\}$ at random.
2. If A is randomized, draw a sequence of sufficiently many true random bits for A.
3. Each time A calls the oracle, draw x from D and flip a coin with bias η for heads. If heads, then return the corrupted example $(x_1, 1 - C(x_1))$; otherwise, return $(x, C(x))$.

Assume for the purpose of contradiction that A PAC learns $\mathcal{C}$ against the above malicious oracle. Let $p_A(m)$ be $\Pr\{H \neq C\}$, where H is the hypothesis generated by A using a sample of size $m = m(\varepsilon, \delta, \eta)$. Since $H \neq C$ implies that H is not an ε-accurate hypothesis, we have that $p_A(m) \leq \delta \leq 1/38$ must hold. For the above malicious oracle, the probability that an example shows x_1 with the wrong label is η. The probability to see x_1 with the true label is a bit higher, namely $(1 - \eta)\varepsilon = \eta + \Delta + \varepsilon\Delta$. Let B be the Bayes strategy that outputs C_1 if the example $(x_1, 1)$ occurs more often in the sample than $(x_1, 0)$, and C_0 otherwise. It is easy to show that B minimizes the probability to output the wrong hypothesis. Thus $p_B(m) \leq p_A(m)$ for all m. We now show that $m \leq 9\eta(1-\eta)/(37\Delta^2)$ implies $p_B(m) > 1/38$. For this purpose, we define events $BAD_1(m)$ and $BAD_2(m)$ over runs of B that use sample size m as follows. $BAD_1(m)$ is the event that at least $\lceil(\eta + \Delta)m\rceil + 1$ examples are corrupted, $BAD_2(m)$ is the event that the true label of x_1 is shown at most $\lceil(\eta + \Delta)m\rceil$ times. Clearly, $BAD_1(m)$ implies that the false label of x_1 is shown at least $\lceil(\eta + \Delta)m\rceil + 1$ times. Thus, $BAD_1(m)$ and $BAD_2(m)$ together imply that B's hypothesis is wrong. Based on the following two claims, we will show that for m too small, $\Pr\{BAD_1(m) \wedge BAD_2(m)\} > 1/38$.

Claim 3.5 *For all $m \geq 1$,*
$$\Pr\{BAD_2(m) \mid BAD_1(m)\} \geq 1/2.$$

Proof of the claim. Given $BAD_1(m)$, there are less than $(1 - \eta - \Delta)m$ uncorrupted examples. Each uncorrupted example shows the true label of x_1 with probability ε. In the average, the true label is shown less than $(1 - \eta - \Delta)\varepsilon m = (1 - \eta_0)\varepsilon m = \eta_0 m = (\eta + \Delta)m$ times. The claim now follows from Fact 3.1. □

Claim 3.6 *If $\frac{37}{\eta(1-\eta)} \leq m \leq \frac{9\eta(1-\eta)}{37\Delta^2}$, then*
$$\Pr\{BAD_1(m)\} > \tfrac{1}{19}.$$

Proof of the claim. Let $S_{m,\eta}$ denote the number of corrupted examples. Fact 3.2 implies that for all $m \geq \frac{37}{\eta(1-\eta)}$,

$$\Pr\left\{S_{m,\eta} \geq \lfloor m\eta \rfloor + \left\lfloor \sqrt{m\eta(1-\eta) - 1} \right\rfloor\right\} > \frac{1}{19}.$$

The claim follows if

$$\lfloor m\eta \rfloor + \left\lfloor \sqrt{m\eta(1-\eta) - 1} \right\rfloor > \lceil \eta m + \Delta m \rceil + 1.$$

This condition is implied by

$$m\eta + \sqrt{m\eta(1-\eta) - 1} \geq \eta m + \Delta m + 3$$

which, in turn, is implied by

$$\frac{1}{2}\sqrt{m\eta(1-\eta) - 1} \geq 3$$

and
$$\frac{1}{2}\sqrt{m\eta(1-\eta)-1} \geq \Delta.$$
The latter two conditions easily follow from the lower and the upper bound on m specified in the statement of the claim. □

From these two claims we obtain that for $\frac{37}{\eta(1-\eta)} \leq m \leq \frac{9\eta(1-\eta)}{37\Delta^2}$, it holds that $p_B(m) > 1/38$. Note that $\Delta \leq \varepsilon/K$ for a sufficiently large constant K implies that the specified range for m contains at least one integer, i.e., the implication is not vacuous. As B is optimal, it cannot be worse than a strategy which ignores sample points, thus the error probability $p_B(m)$ does not increase with m. We may therefore drop the condition $m \geq \frac{37}{\eta(1-\eta)}$. This completes the proof. □

The proof of our next lower bound combines the technique from [EHKV89] for showing the lower bound on the sample size in the noise-free PAC learning model with the argument of statistical indistinguishability. Here the indistinguishability is used to force with probability 1/2 a mistake on a point x, with $D(x) = \eta/(1-\eta)$. To ensure that with probability greater than δ the learner outputs an hypothesis with error at least ε, we use t other points that are empirically seen very seldom. This entails that the learning algorithm must essentially perform a random guess on half of them.

Theorem 3.7 *For any target class $\mathcal{C}$ with VC-dimension $d \geq 3$, and for any $0 < \varepsilon \leq 1/8$, $0 < \delta \leq 1/12$, and any $0 < \Delta < \varepsilon/(1+\varepsilon)$, the sample size needed for PAC learning $\mathcal{C}$ with accuracy ε, confidence δ, and tolerating malicious noise rate $\eta = \varepsilon/(1+\varepsilon) - \Delta$, is greater than*
$$\frac{d-2}{32\Delta(1+\varepsilon)} = \Omega\left(\frac{d}{\Delta}\right).$$
Note that for $\Delta = \varepsilon/(1+\varepsilon)$, i.e. $\eta = 0$, this reduces to the known lower bound on the sample size for noise-free PAC learning.

Proof. Let $t = d-2$ and let $X_0 = \{x_0, x_1, \ldots, x_t, x_{t+1}\}$ the set of points shattered by $\mathcal{C}$. We may assume w.l.o.g. that $\mathcal{C}$ is the powerset of X_0. We define distribution D as follows:
$$D(x_0) = 1 - \frac{\eta}{1-\eta} - 8\left(\varepsilon - \frac{\eta}{1-\eta}\right),$$
$$D(x_1) = \cdots = D(x_t) = \frac{8\left(\varepsilon - \frac{\eta}{1-\eta}\right)}{t}, D(x_{t+1}) = \frac{\eta}{1-\eta}.$$
(Note that $\varepsilon \leq 1/8$ implies that $D(x_0) \geq 0$.) We will use a malicious oracle that, with probability η, corrupts the current example by replacing it with x_{t+1} labeled incorrectly. Therefore x_{t+1} is shown incorrectly labeled with probability η and correctly labeled with probability $(1-\eta)D(x_{t+1}) = \eta$. Thus, true and false labels for x_{t+1} are statistically indistinguishable. We will call $x_1, \ldots, x_t$ *rare points* in the sequel. Note that when η approaches η_0 the probability to select a rare point approaches 0. Let A be a (possibly) randomized learning algorithm for $\mathcal{C}$ which demands sample size $m = m(\varepsilon, \delta, \eta)$. Consider the following random experiment:

1. Flip a fair coin to select target $C \in \mathcal{C}$ at random.
2. If A is randomized, draw a sequence of sufficiently many true random bits for A.
3. Each time A calls the oracle, draw x from D and flip a coin with bias η for heads. If heads, then return the corrupted example $(x_{t+1}, 1-C(x_1))$; otherwise, return $(x, C(x))$.

There is a subtle way to use the statistical indistinguishability of the two labels for x_{t+1}. We obtain an equivalent random experiment when we modify step 3 by

3′. Determine a true random bit β by an independent coin flip and let $C(x_{t+1}) = \beta$.

Thus, the probability to misclassify x_{t+1} is always 1/2 (without any dependence on the other randomly chosen quantities). This observation will be used later in the analysis.

Let e_A be the random variable denoting the error $\Pr\{H \neq C\}$ of A's hypothesis H. Then, by pigeonhole,
$$\Pr\{e_A \geq \varepsilon\} > \frac{1}{12} \tag{3}$$
and this implies the existence of a concept $C_0 \in \mathcal{C}$ such that the probability that A does not output an ε-accurate hypothesis for C_0 is greater than $1/12 \geq \delta$. Let us assume for the purpose of contradiction that $m \leq t/(32\Delta(1+\varepsilon))$. It then suffices to show that (3) holds.

Towards this end, we will define events BAD_1, BAD_2, and BAD_3, over runs of A that use sample size m, whose conjunction has probability greater than 1/12 and implies (3). BAD_1 is the event that at least $t/2$ rare points are not returned as examples by the oracle. In what follows, we call *unseen* the rare points that are not returned by the oracle. Given BAD_1, we denote by UP the set of $t/2$ unseen points with lowest indices and we define BAD_2 as the event that hypothesis H classifies at least $t/8$ points of UP incorrectly. Finally, BAD_3 is the event that hypothesis H classifies x_{t+1} incorrectly. It is easy to see that $BAD_1 \wedge BAD_2 \wedge BAD_3$ implies (3), because the total probability of misclassification adds up to
$$\frac{t}{8} \cdot \frac{8\left(\varepsilon - \frac{\eta}{1-\eta}\right)}{t} + \frac{\eta}{1-\eta} = \varepsilon.$$

We finally have to discuss the probabilities of the 3 events. Only noise-free examples potentially show one of the rare points. The probability that this happens is
$$8\left(\varepsilon - \frac{\eta}{1-\eta}\right)(1-\eta) = 8(\varepsilon(1-\eta)-\eta) = 8\Delta(1+\varepsilon).$$
Since $m \leq \frac{t}{32\Delta(1+\varepsilon)}$, the examples returned by the oracle contain at most $t/4$ rare points on the average. It follows from Markov inequality that the probability that these examples contain more than $t/2$ rare points is smaller than 1/2. Thus $\Pr\{BAD_1\} > 1/2$. For each unseen point there is a chance of 1/2 of misclassification. Thus $\Pr\{BAD_2 \mid BAD_1\}$ is the probability that a fair coin flipped $t/2$ times shows heads at least $t/8$ times. Fact 3.3 applied with $\alpha = 1/2$, and $\beta = 1/4$, implies that this probability is greater than 1/3. As the boolean labels of x_{t+1} are statistically indistinguishable, we get $\Pr\{BAD_3 \mid BAD_1, BAD_2\} = \Pr\{BAD_3\} = 1/2$. Thus
$$\begin{aligned}&\Pr\{BAD_1 \wedge BAD_2 \wedge BAD_3\}\\ &= \Pr\{BAD_1\} \cdot \Pr\{BAD_2 \mid BAD_1\} \cdot \Pr\{BAD_3\}\\ &> \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{12},\end{aligned}$$

which completes the proof. □

It follows from the proofs of the previous two theorems that the lower bound $\Omega(\eta/\Delta^2+d/\Delta)$ holds even for a "weak" malicious oracle whose corruption strategy does not depend on the past examples in the sample. The upper bound presented in Section 4 matches the lower bound (ignoring logarithmic factors), and holds even for a "strong" malicious oracle. Learning in the presence of *strong* malicious noise is performed according to the following protocol:

The learning algorithm issues all of the m oracle calls at the beginning. The oracle then draws $x_1, \ldots, x_m$ independently and according to D. Afterwards, a coin with bias η for heads is tossed m times, where heads on the t-th coin toss means that the oracle is allowed to corrupt the t-th example. Finally, the malicious oracle answers all calls (note that here the corruption strategy may depend on the whole sample) and the learner must compute its hypothesis without any more calls to the oracle.

The standard malicious oracle introduced in Section 2 lies somewhere in between the weak one and the strong one, as the corruption strategy may depend on the past examples, but not on the future ones. The following lower bound on the sample size needed by the minimum disagreement strategy (MDS henceafter) is shown for the strong malicious oracle. Given the whole sample S' of corrupted training examples, MDS will output the hypothesis $H \in C$ with the fewest disagreements on S'.

Theorem 3.8 *For any target class C with VC-dimension $d \geq 3$, any $0 < \varepsilon \leq 1/38$, $0 < \delta \leq 1/74$, and any $0 < \Delta = o(\varepsilon)$, the sample size needed by the Minimum Disagreement strategy for learning C with accuracy ε, confidence δ, and tolerating malicious noise rate $\eta = \frac{\varepsilon}{1+\varepsilon} - \Delta$, is greater than*

$$\frac{4(1-\eta)(1-\varepsilon)\lceil (d-1)/38 \rceil \varepsilon}{37(1+\varepsilon)^2\Delta^2} = \Omega\left(\frac{d\varepsilon}{\Delta^2}\right).$$

Proof. The proof uses d shattered points, where $d-1$ of them (the rare points) have a relatively small probability. The total probability of all rare points is $c\varepsilon$ for some constant c. Let μ be the mean and σ the standard deviation for the number of true labels of a rare point within the (corrupted) training sample. If the rare points were shown μ times, the malicious oracle would have no chance to fool MDS. The basic idea is however to argue as follows. If the sample size m is too small, it leads to a standard deviation σ for the number of true labels of a rare point x which is too big. Hence, with a constant probability, the number of true labels of a rare point is smaller than (roughly) $\mu - \sigma$. If this happens, we call x a hidden point. It follows that there is also a constant probability that a constant fraction of the rare points are hidden. This gives the strong malicous oracle the chance to present more false than true labels for each hidden point. We now make these ideas precise.

Our proof needs the following technical assumption:

$$m \geq \frac{37\lceil (d-1)/38 \rceil}{\varepsilon(1-\varepsilon)(1-\eta)}. \tag{4}$$

This condition can be forced by invoking the general lower bound $\Omega(d/\Delta)$ from Theorem 3.7 for $\Delta \leq \varepsilon/K$ and a sufficiently large constant K. For the purpose of contradiction, we assume that

$$m \leq \frac{4(1-\eta)(1-\varepsilon)\lceil (d-1)/38 \rceil \varepsilon}{37(1+\varepsilon)^2\Delta^2}. \tag{5}$$

Let BAD_1 be the event that at least $\lfloor \eta m \rfloor$ examples are corrupted by the strong malicious oracle. According to Fact 3.1, BAD_1 has probability at least 1/2. Let $t = d-1$ and let $X_0 = \{x_0, \ldots, x_t\}$ be the set of points shattered by C. Distribution D is defined by

$$D(x_0) = 1 - t\lceil t/38 \rceil^{-1}\varepsilon, D(x_1) = \cdots = D(x_t) = \lceil t/38 \rceil^{-1}\varepsilon.$$

Points $x_1, \ldots, x_t$ are called *rare*. Consider a fixed rare point x_i. Each example shows x_i with its true label with probability

$$p = \lceil t/38 \rceil^{-1}\varepsilon(1-\eta) = \lceil t/38 \rceil^{-1}(\eta + \Delta(1+\varepsilon)).$$

Let T_i denote the number of examples that present the true label of x_i. Call x_i *hidden* if

$$T_i \leq \lceil pm \rceil - \lfloor \sqrt{mp(1-p)-1} \rfloor.$$

Inequality (4) implies that $m \geq \frac{37}{p(1-p)}$. Thus, according to Fact 3.2, x_i is hidden with probability greater than 1/19. Using the fact that $\Pr\{x_i \text{ is hidden}\}$ is equal to

$$\Pr\{x_i \text{ is hidden} \mid BAD_1\} \Pr\{BAD_1\} \\ + \Pr\{x_i \text{ is hidden} \mid \neg BAD_1\} (1 - \Pr\{BAD_1\})$$

and

$$\Pr\{x_i \text{ is hidden} \mid BAD_1\} \geq \Pr\{x_i \text{ is hidden} \mid \neg BAD_1\},$$

it follows that

$$\Pr\{x_i \text{ is hidden} \mid BAD_1\} \geq \Pr\{x_i \text{ is hidden}\} > \frac{1}{19}.$$

Given BAD_1, let T be the (conditional) random variable which counts the number of hidden points. The expectation of T is greater than $t/19$. According to Fact 3.3 (with $\alpha = 1/19$ and $\beta = 1/38$), the probability that at least $t/38$ rare points are hidden is greater than 1/37. Thus with probability greater than $\delta = 1/74$, there are (at least) $\lfloor \eta m \rfloor$ corrupted examples and (at least) $\lceil t/38 \rceil$ hidden points. This is assumed in the sequel.

The total probability of $\lceil t/38 \rceil$ hidden points (measured by D) is exactly $\lceil t/38 \rceil \lceil t/38 \rceil^{-1}\varepsilon = \varepsilon$. It suffices therefore to show that there are enough corrupted examples to present each of the $\lceil t/38 \rceil$ hidden points with more false than true labels. The total number of true labels for $\lceil t/38 \rceil$ hidden points can be bounded from above:

$$\lceil t/38 \rceil \cdot \left(\lceil pm \rceil - \lfloor \sqrt{mp(1-p)-1} \rfloor\right) \\ \leq \eta m + \Delta(1+\varepsilon)m + 2\lceil t/38 \rceil \\ - \lceil t/38 \rceil \sqrt{\frac{m\varepsilon(1-\eta)(1-\varepsilon)}{\lceil t/38 \rceil} - 1}.$$

The number of false labels that the oracle can use is greater than $\eta m - 1$ and should exceed the number of true labels by at least $\lceil t/38 \rceil$. The oracle can therefore force an ε-inaccurate hypothesis of MDS if

$$\eta m - 1 \geq \eta m + \Delta(1+\varepsilon)m + 3\lceil t/38 \rceil \\ - \lceil t/38 \rceil \sqrt{\frac{m\varepsilon(1-\eta)(1-\varepsilon)}{\lceil t/38 \rceil} - 1}$$

or equivalently if

$$\lceil t/38\rceil\sqrt{\frac{m\varepsilon(1-\eta)(1-\varepsilon)}{\lceil t/38\rceil}}-1 \geq \Delta(1+\varepsilon)m+3\lceil t/38\rceil+1. \tag{6}$$

We will develop a sufficient condition which is easier to handle. The right-hand side of (6) contains the three terms

$$z_1 = 3\lceil t/38\rceil,\ z_2 = 1,\ z_3 = \Delta(1+\varepsilon)m.$$

Splitting the left-hand side Z of (6) in three parts, we obtain the sufficient condition $Z/2 \geq z_1, Z/6 \geq z_2, Z/3 \geq z_3$, which reads (after some algebraic simplifications) in expanded form as follows:

$$\begin{aligned}
\sqrt{\frac{m\varepsilon(1-\eta)(1-\varepsilon)}{\lceil t/38\rceil}}-1 &\geq 6\\
\lceil t/38\rceil\sqrt{\frac{m\varepsilon(1-\eta)(1-\varepsilon)}{\lceil t/38\rceil}}-1 &\geq 6\\
\lceil t/38\rceil\sqrt{\frac{m\varepsilon(1-\eta)(1-\varepsilon)}{\lceil t/38\rceil}}-1 &\geq 3\Delta(1+\varepsilon)m
\end{aligned}$$

An easy computation shows that these three conditions are implied by (4) and (5). This completes the proof. □

It is an open question whether a similar lower bound can be proven for the standard (or even the weak) oracle.

4 A tight upper bound for the class of all subsets over d points

In this section we show that the lower bound proven in Section 3 cannot be improved in general. That is, we show that for each $d \geq 1$, the class C_d of all subsets over d points can be PAC learned with accuracy $\varepsilon > 0$ and malicious noise rate $\eta < \varepsilon/(1+\varepsilon)$ using a sample of size $\widetilde{O}(\eta/\Delta^2 + d/\Delta)$, where $\Delta = \varepsilon/(1+\varepsilon) - \eta$. The learning algorithm uses a majority vote on the sample to decide the labels of some of the domain points and tosses a fair coin to decide the labels of the remaining ones.

4.1 The algorithm RMD

In this section we prove the following result for the randomized algorithm RMD informally described below. The pseudo-code may be found in Figure 1.

Theorem 4.1 *For any $d \geq 1$ and any $1 \geq \varepsilon, \delta, \Delta > 0$, algorithm* RMD, *with input parameters $\alpha = \sqrt[3]{5/3} - 1$, $L = \left\lceil \log \frac{6d(1+\alpha)^2\varepsilon}{\Delta} \right\rceil$, and $n = \lceil 50\ln(4L/\delta)\rceil$, PAC learns the class C_d with accuracy ε, confidence δ, tolerating malicious noise rate $\eta = \varepsilon/(1+\varepsilon) - \Delta$, and using a sample of size $\widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$.*

As C_d includes all concepts on $\{1, \dots, d\}$, algorithm RMD can choose its hypothesis by deciding the label of each point in the domain independently. This is done as follows: based on the sample, the domain is divided into two main groups. The label of each point i in the first group is decided by taking a majority vote on the occurrences of $(i, 0)$ and $(i, 1)$ in the sample. The labels of the points in the second group are instead chosen in a random way.

```
Algorithm RMD.
Input: Parameters $\alpha, L, n$. Domain size $d$, accuracy $\varepsilon$, confidence $\delta$.
  • Make $m = m(d, \varepsilon, \delta, \alpha, L, n)$ calls to the malicious oracle, where $m = \widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$, and get sample $(i_1, y_1), \dots, (i_m, y_m)$.
  • For each point $i \in \{1, \dots, d\}$.
      1. If $i$ is in strong majority or belongs to a sparse band, then let $H(i)$ be the most frequent label with which $i$ appears in the sample;
      2. else, let $H(i)$ be a random label.
  • Output the hypothesis $H$.
```

Figure 1: Pseudo-code for the randomized algorithm RMD (see Theorem 4.1).

To bound the total error of the hypothesis chosen by the algorithm, we divide each of the two above groups into subgroups, and then separately bound the contributions of each subgroup to the total error. Within each such subgroup, we approximately bound the total probability of the domain points for which the algorithm chooses the wrong label by the total frequency of corrupted examples of points in the subgroup. Since, for a large enough sample, the sample frequency of corrupted examples is very close to the actual noise rate η, and since the noise rate is bounded away from the desired accuracy ε, we can show that the total probability of the points labeled incorrectly, summed over all subgroups, is at most the desired accuracy ε.

Given a sample $(i_1, y_1), \dots, (i_m, y_m)$ drawn from the set $\{1, \dots, d\} \times \{0, 1\}$, let $\nu_{0,i}$ and $\nu_{1,i}$ be the frequencies with which each point $i \in \{1, \dots, d\}$ appears in the sample with label respectively 0 and 1. For each i, we define $\ell_i = \min\{\nu_{0,i}, \nu_{1,i}\}$ and $u_i = \max\{\nu_{0,i}, \nu_{1,i}\}$. For some fixed $\alpha > 0$, a domain point i is in *strong majority* (with respect to the sample) if $u_i > (1+\alpha)\ell_i$, and is in *weak majority* otherwise. We divide some of the points into L *bands,* for fixed integer $L \geq 1$. A point i is in band k, for $k = 1, \dots, L$, if i is in weak majority and $(1+\alpha)^{-k}\varepsilon < \ell_i \leq (1+\alpha)^{1-k}\varepsilon$. We further divide the bands in *sparse bands*, containing less than n elements, and *dense bands*, containing at least n elements, where n is some other fixed positive integer.

4.2 Proof of Theorem 4.1

Let D be any distribution assigning probability p_i to each $i \in \{1, \dots, d\}$. We say that a point i is *heavy* if $p_i \geq \Delta/3d$. Let I_{heavy} be the set of all heavy points and I_{light} its complement with respect to $\{1, \dots, d\}$. Let I_{maj}, I_{sparse}, and I_{dense} be the sets of all domain points respectively in strong majority, sparse bands and dense bands. For fixed choice of input parameters, we denote RMD's hypothesis by H.

For simplicity, for each point i we will write t_i and f_i to denote, respectively, $\nu_{C(i),i}$ and $\nu_{1-C(i),i}$. That is, t_i and f_i are the sample frequencies of, respectively, clean and corrupted examples associated with each point i. We define

$$f_{\text{maj}} = \sum_{i \in I_{\text{maj}}} f_i,\ \ f_{\text{sparse}} = \sum_{i \in I_{\text{sparse}}} f_i,\ \ f_{\text{dense}} = \sum_{i \in I_{\text{dense}}} f_i.$$

We now state some classical Chernoff-Hoeffding inequalities (see e.g. [Lit95]) we will repeatedly use throughout the proof.

Let $S_{m,p}$ and $S'_{m,p'}$ be the sums of successes in a sequence of m Bernouilli trials each succeding with probability respectively at least p and at most p'. Then, for all $0 < \lambda < 1$,

$$\begin{aligned} \Pr\{S_{m,p} \le (1-\lambda)mp\} &\le e^{-\lambda^2 mp/2} & (7)\\ \Pr\{S_{m,p} \le m(p-\lambda)\} &\le e^{-2\lambda^2 m}, & (8)\\ \Pr\{S'_{m,p} \ge (1+\lambda)mp'\} &\le e^{-\lambda^2 mp'/3}. & (9)\end{aligned}$$

First, we upper bound in probability the sum $f_{\text{maj}} + f_{\text{sparse}} + f_{\text{dense}}$. Let $\hat{\eta}$ be the frequency of corrupted examples in the sample. By using (9) with $p = \eta$ and $\lambda = \Delta/(3\eta)$, we find that

$$f_{\text{maj}} + f_{\text{sparse}} + f_{\text{dense}} \le \hat{\eta} \le \left(1 + \frac{\Delta}{3\eta}\right)\eta = \eta + \frac{\Delta}{3} \qquad (10)$$

holds with probability at least $1 - \delta/4$ whenever the sample size is at least $(27\eta/\Delta^2)\ln(4/\delta) = \widetilde{O}(\eta/\Delta^2)$.

Second, we lower bound in probability the sample frequency t_i of uncorrupted examples for each $i \in I_{\text{heavy}}$. Note that the probability that a point i appears uncorrupted in the sample is at least $(1-\eta)p_i$. Also, $|I_{\text{heavy}}| \le d$, as there are at most d points. By using (7) with $p = \Delta/(3d)$ and $\lambda = \alpha/(1+\alpha)$, we find that

$$t_i \ge \frac{1-\eta}{1+\alpha}p_i = \left(1 - \frac{\alpha}{1+\alpha}\right)(1-\eta)p_i \quad \forall i \in I_{\text{heavy}} \qquad (11)$$

holds with probability at least $1 - \delta/4$ whenever the sample size is at least

$$\frac{6(1+\alpha)^2 d}{(1-\eta)\alpha^2\Delta}\ln\frac{4d}{\delta} = \widetilde{O}\left(\frac{d}{\Delta}\right).$$

Thus, if the sample size is $\widetilde{O}(\eta/\Delta^2 + d/\Delta)$, then (10) and (11) simultaneously hold with probability at least $1 - \delta/2$.

Let $I_{\text{wrong}} = \{i : C(i) \ne H(i)\}$. Claim 4.5 shows that, if (10) and (11) hold, then all heavy points are in the set $I_{\text{maj}} \cup I_{\text{sparse}} \cup I_{\text{dense}}$. Thus

$$\begin{aligned} D(I_{\text{wrong}}) \le\ & D(I_{\text{wrong}} \cap I_{\text{maj}}) & (12)\\ & + D(I_{\text{wrong}} \cap I_{\text{sparse}}) + D(I_{\text{wrong}} \cap I_{\text{dense}}) \\ & + D\left(I_{\text{light}} \setminus (I_{\text{maj}} \cup I_{\text{sparse}} \cup I_{\text{dense}})\right). \end{aligned}$$

Now, Claims 4.2–4.4 show how the first three terms in the right-hand side of (12) can be simultaneously bounded. In the rest of this section, we prove Claims 4.2–4.5. We start by bounding the error made by H on points $i \in I_{\text{maj}}$.

Claim 4.2 (Strong majority) *If (11) holds, then*
$D(I_{\text{wrong}} \cap I_{\text{maj}}) \le \frac{f_{\text{maj}}}{1-\eta} + D(I_{\text{maj}} \cap I_{\text{light}})$.

Proof. Recall that, for each $i \in I_{\text{maj}}$, $H(i) \ne C(i)$ if and only if $t_i = \ell_i$. Hence, if (11) holds, we find that for any $i \in I_{\text{wrong}} \cap I_{\text{maj}} \cap I_{\text{heavy}}$, $(1-\eta)p_i \le (1+\alpha)t_i = (1+\alpha)\ell_i \le \frac{1+\alpha}{1+\alpha}u_i = f_i$. As $\sum_{i \in I_{\text{wrong}} \cap I_{\text{maj}} \cap I_{\text{heavy}}} f_i \le f_{\text{maj}}$, the proof is concluded. □

We now bound the error occurring in the sparse bands by proving the following.

Claim 4.3 (Sparse bands) *Let the sample size be at least*

$$\frac{6(1+\alpha)\varepsilon L^2 n^2}{\Delta^2}\ln\frac{4d}{\delta} = \widetilde{O}\left(\frac{\varepsilon}{\Delta^2}\right).$$

Then (10) and (11) together imply that $D(I_{\text{wrong}} \cap I_{\text{sparse}}) \le \frac{f_{\text{sparse}}}{1-\eta} + \frac{\Delta}{(1-\eta)3}$ holds with probability at least $1 - \delta/4$ with respect to the sample random draw.

Proof. Recall that there are L bands and each sparse band contains at most n elements. We first prove that

$$t_i \ge (1-\eta)p_i - \frac{\Delta}{3Ln} \qquad \text{for all } i \in I_{\text{wrong}} \cap I_{\text{sparse}} \qquad (13)$$

holds in probability. To show this, we use (7) to write the following

$$\begin{aligned} \Pr\{S_m \le (p-\lambda)m\} &= \Pr\left\{S_m \le \left(1 - \frac{\lambda}{p}\right)mp\right\} \\ &\le \exp\left(-\frac{\lambda^2 m}{2p}\right) \le \exp\left(-\frac{\lambda^2 m}{2p'}\right), \qquad (14) \end{aligned}$$

where the last inequality holds for all $p' \ge p$ by monotonicity. Now assume (10) and (11) both hold and choose i such that $p_i > (1+\alpha)\varepsilon/(1-\eta)$. Then $i \in I_{\text{heavy}}$ and $t_i \ge \varepsilon$. As $\hat{\eta} < \varepsilon$ by (10), $i \notin I_{\text{wrong}}$. Hence, (10) and (11) imply that $p_i \le (1+\alpha)\varepsilon/(1-\eta)$ holds for all $i \in I_{\text{wrong}}$. We then apply (14) to each $i \in I_{\text{wrong}} \cap I_{\text{sparse}}$. Setting $p = (1-\eta)p_i$, $p' = (1+\alpha)\varepsilon \ge p$, and $\lambda = \Delta/(3Ln)$ we find that (13) holds with probability at least $1 - \delta/4$ whenever the sample size is at least

$$\frac{18(1+\alpha)\varepsilon L^2 n^2}{\Delta^2}\ln\frac{4d}{\delta} = \widetilde{O}\left(\frac{\varepsilon}{\Delta^2}\right).$$

Finally, from (13) we get that

$$\begin{aligned} D(I_{\text{wrong}} \cap I_{\text{sparse}}) &\le \sum_{I_{\text{wrong}} \cap I_{\text{sparse}}} \left(\frac{t_i}{1-\eta} + \frac{\Delta}{(1-\eta)3Ln}\right) \\ &\le \sum_{I_{\text{wrong}} \cap I_{\text{sparse}}} \frac{f_i}{1-\eta} + \frac{\Delta}{(1-\eta)3} = \frac{f_{\text{sparse}}}{1-\eta} + \frac{\Delta}{(1-\eta)3}. \end{aligned}$$

This concludes the proof. □

We move on to bounding the error made on points in dense bands.

Claim 4.4 (Dense bands) *If (11) holds, then*
$D(I_{\text{wrong}} \cap I_{\text{dense}}) \le \frac{f_{\text{dense}}}{1-\eta} + D(I_{\text{dense}} \cap I_{\text{light}})$
holds with probability at least $1 - \delta/4$ with respect to the algorithm randomization.

Proof. For each $k = 1, \ldots, L$, let B_k be all points in the k-th band. Furthermore, let $t^k_{\max} = \max\{t_i : i \in B_k \cap I_{\text{wrong}}\}$ and $f^k_{\min} = \min\{f_i : i \in B_k \cap I_{\text{wrong}}\}$. Since all points in B_k are in weak majority and by definition of bands, we have that $t^k_{\max} \le (1+\alpha)^2 f^k_{\min}$ holds for each $k = 1, \ldots, L$. Furthermore, using (11), $p_j \le \frac{1+\alpha}{1-\eta}t_j$, for each $j \in B_k \cap I_{\text{heavy}}$. As for each dense band $|B_k| \ge n \ge 50\ln(4L/\delta)$, using (8) we can guarantee that $|B_k \cap I_{\text{wrong}}| \le \frac{3}{5}|B_k|$ holds simultaneously for all bands $k = 1, \ldots, L$ with probability at least $1 - \delta/4$. Combining everything we get

$$\sum_{B_k \cap I_{\text{wrong}} \cap I_{\text{heavy}}} p_i \leq \frac{1+\alpha}{1-\eta} \sum_{B_k \cap I_{\text{wrong}} \cap I_{\text{heavy}}} t_i$$
$$\leq \frac{1+\alpha}{1-\eta} \sum_{B_k \cap I_{\text{wrong}}} t_i \leq \frac{1+\alpha}{1-\eta} \cdot \frac{3}{5} |B_k| t^k_{\max}$$
$$\leq \frac{(1+\alpha)^3}{1-\eta} \cdot \frac{3}{5} |B_k| f^k_{\min} \leq \frac{(1+\alpha)^3}{1-\eta} \cdot \frac{3}{5} \sum_{B_k} f_i.$$

By choosing $\alpha = (5/3)^{1/3} - 1$ so that $(3/5)(1+\alpha)^3 = 1$, we get

$$D(I_{\text{dense}} \cap I_{\text{wrong}} \cap I_{\text{heavy}}) = \sum_k \sum_{B_k \cap I_{\text{wrong}} \cap I_{\text{heavy}}} p_i$$
$$\leq \sum_k \sum_{B_k} \frac{f_i}{1-\eta} = \frac{f_{\text{dense}}}{1-\eta}$$

concluding the proof. □

Claim 4.5 *If (10) and (11) hold, then*
$$I_{\text{heavy}} \subseteq I_{\text{maj}} \cup I_{\text{sparse}} \cup I_{\text{dense}}.$$

Proof. If (10) holds, then $\ell_i \leq f_i \leq \eta + \frac{\Delta}{3} < \varepsilon$ for each point i. Also, if (11) holds, then, for each $i \in I_{\text{heavy}} \cap I_{\text{maj}}$, $\ell_i \geq \frac{u_i}{1+\alpha} \geq \frac{t_i}{1+\alpha} \geq \frac{1-\eta}{(1+\alpha)^2} p_i \geq \frac{(1-\eta)\Delta}{3d(1+\alpha)^2}$. Thus, by the choice of L (and recalling that $\eta < 1/2$), i belongs to a band. Hence, if (10) and (11) both hold, then all points not in $I_{\text{maj}} \cup I_{\text{sparse}} \cup I_{\text{dense}}$ are light points. □

To finish the proof of the theorem recall that $D(I_{\text{light}}) \leq \Delta/3$ by definition, and $\eta = O(\varepsilon)$. Combining the above we find from (12) that $D(I_{\text{wrong}}) \leq \frac{\eta+\Delta}{1-\eta} < \frac{\eta_0}{1-\eta_0} = \varepsilon$ holds with probability at least $1-\delta$ for a sample of size $\widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$, as desired.

Acknowledgements

The first author gratefully acknowledges the support of Progetto Vigoni and Esprit Working Group N. 8556 "Neural and Computational Learning". The third and fourth author gratefully acknowledge the support of Bundesministerium für Forschung und Technologie grant 01IN102C/2, the support of Deutsche Forschungsgemeinschaft grant Si 498/3-1, the support of Deutscher Akademischer Austauschdienst grant 322-vigoni-dr and the support of Deutsche Forschungsgemeinschaft research grant We 1066/6-2.

The authors take the responsibility for the contents.

References

[AD93] J. A. Aslam and S. E. Decatur. General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. In *Proc. 35th Annual IEEE Sympos. Found. Comput. Sci.*, November 1993.

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

[AST94] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1994.

[Bah60] R.R. Bahadur. Some approximations to the binomial distribution function. *Annals of Mathematical Statistics*, 31:43–54, 1960.

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

[EHKV89] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

[JS68] Kumar Jogdeo and S. M. Samuels. Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. *The Annals of Mathematical Statistics*, 39(4):1191–1195, 1968.

[Kea93] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proc. 25th Annual ACM Sympos. Theory Comput.*, pages 392–401. ACM Press, New York, NY, 1993.

[KL93] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22:807–837, 1993.

[Lai88] Philip D. Laird. Learning from good and bad data. In *Kluwer international series in engineering and computer science.* Kluwer Academic Publishers, Boston, 1988.

[Lit95] N. Littlestone. On the derivation and quality of Chernoff bounds. Submitted for publication, 1995.

[Sim93] Hans Ulrich Simon. General bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the 6th Annual Workshop on Computational Learning Theory*, pages 402–412. ACM Press, 1993. To appear in Journal of Computer and System Sciences.

[STAB93] John Shawe-Taylor, Martin Anthony, and Norman Biggs. Bounding sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 41:65–73, 1993.

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

A Proof of Fact 3.2

Proof. In this extended abstract we prove inequality (1) (the proof of (2) is similar). We proceed by establishing a series of inequalities. We shall also use Stirling's formula

$$\sqrt{2\pi N}\left(\frac{N}{e}\right)^N < N! < \sqrt{2\pi N}\left(\frac{N}{e}\right)^N e^{\frac{1}{12N}}\,. \qquad (15)$$

Using (15) one can lower bound the binomial coefficient $\binom{N}{Np}$ as follows (assuming that and N is a multiple of $1/p$, which will be justified later in the proof)

$$\begin{aligned}\binom{N}{Np} &= \frac{N!}{(Np)!(Nq)!}\\ &> \frac{\sqrt{2\pi N}\left(\frac{N}{e}\right)^N}{\sqrt{2\pi Np}\left(\frac{Np}{e}\right)^{Np}e^{\frac{1}{12Np}}\sqrt{2\pi Nq}\left(\frac{Nq}{e}\right)^{Nq}e^{\frac{1}{12Nq}}}\\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{Npq}}\frac{1}{p^{Np}\,q^{Nq}}\,e^{-\frac{1}{12Npq}}\,.\end{aligned}$$

This leads to

$$\sqrt{Npq} > \frac{1}{\sqrt{2\pi}p^{Np}\,q^{Nq}}\binom{N}{Np}^{-1}e^{-\frac{1}{12Npq}}\,. \qquad (16)$$

Bahadur [Bah60] proved the following lower bound on the tail of the binomial distribution, where $0 \le k \le N$,

$$\begin{aligned}\Pr\{S_{N,p} \ge k\} \;\ge\; & \binom{N}{k}p^k q^{(N-k)}\cdot\frac{q(k+1)}{k+1-p(N+1)}\\ & \cdot\left(1+\frac{Npq}{(k-Np)^2}\right)^{-1}. \qquad (17)\end{aligned}$$

In order to be able to apply (16) we remove the first "floors" in (1). To this end we replace p by $p' = p - \gamma$ (and q by $q' = q + \gamma$) such that $\lfloor Np\rfloor = Np'$. Then Np' is integer and $p' > p - \frac{1}{N}$. We shall also need the following observation.

$$\begin{aligned}pq = (p'+\gamma)(q'-\gamma) &= p'q' + \gamma(q'-p') - \gamma^2\\ &= p'q' + \gamma(q'-p'-\gamma)\\ &= p'q' + \gamma(q-p')\\ &< p'q' + \gamma < p'q' + \frac{1}{N}. \qquad (18)\end{aligned}$$

Then (18) and $N \ge 37/(pq)$ imply that

$$Np'q' > Npq - 1 \ge 36. \qquad (19)$$

Hence (1) can be lower bounded as follows

$$\begin{aligned}&\Pr\left\{S_{N,p} \ge \lfloor Np\rfloor + \left\lfloor\sqrt{Npq-1}\right\rfloor\right\}\\ &\ge \Pr\left\{S_{N,p'} \ge Np' + \left\lfloor\sqrt{Npq-1}\right\rfloor\right\}\\ &\ge \Pr\left\{S_{N,p'} \ge Np' + \left\lfloor\sqrt{N(p'q'+\frac{1}{N})-1}\right\rfloor\right\}\\ &\ge \Pr\left\{S_{N,p'} \ge Np' + \left\lfloor\sqrt{Np'q'}\right\rfloor\right\}. \qquad (20)\end{aligned}$$

In order to bound (20) we apply inequality (17) with $k = Np' + \lfloor\sqrt{Np'q'}\rfloor$ and p and q being replaced by p' and q', respectively. The three factors in the right-hand side of (17), denoted by F_1, F_2 and F_3, are separately bounded as follows.

$$F_1 = \binom{N}{Np' + \lfloor\sqrt{Np'q'}\rfloor}p'^{Np'}q'^{Nq'}\left(\frac{p'}{q'}\right)^{\lfloor\sqrt{Np'q'}\rfloor}. \qquad (21)$$

$$\begin{aligned}F_2 &= \frac{q'(Np' + \lfloor\sqrt{Np'q'}\rfloor + 1)}{Np' + \lfloor\sqrt{Np'q'}\rfloor + 1 - Np' - p'}\\ &= \frac{Np'q' + q'(\lfloor\sqrt{Np'q'}\rfloor + 1)}{\lfloor\sqrt{Np'q'}\rfloor + q'} > \sqrt{Np'q'}\\ &> \frac{1}{\sqrt{2\pi}\,p'^{Np'}\,q'^{Nq'}}\binom{N}{Np'}^{-1}e^{-\frac{1}{12Np'q'}}. \qquad (22)\end{aligned}$$

(The inequality (22) follows from (16).)

$$\begin{aligned}F_3 &= \left(1 + \frac{Np'q'}{\left(Np' + \lfloor\sqrt{Np'q'}\rfloor - Np'\right)^2}\right)^{-1}\\ &= \left(1 + \frac{Np'q'}{\lfloor\sqrt{Np'q'}\rfloor^2}\right)^{-1}\\ &= \frac{\lfloor\sqrt{Np'q'}\rfloor^2}{\lfloor\sqrt{Np'q'}\rfloor^2 + Np'q'}\\ &> \frac{\left(\sqrt{Np'q'}-1\right)^2}{\left(\sqrt{Np'q'}+1\right)^2 + Np'q'}\\ &= \frac{Np'q' - 2\sqrt{Np'q'} + 1}{2Np'q' - 2\sqrt{Np'q'} + 1}\\ &> \frac{Np'q' - 2\sqrt{Np'q'}}{2Np'q' - 2\sqrt{Np'q'}}\\ &= \frac{1}{2}\left(1 - \frac{\sqrt{Np'q'}}{Np'q' - \sqrt{Np'q'}}\right)\\ &= \frac{1}{2}\left(1 - \frac{1}{\sqrt{Np'q'}-1}\right). \qquad (23)\end{aligned}$$

The following calculation shows how the product of (21) and (22) can be lower bounded. For notational convenience let $T = \left(e^{\frac{1}{12Np'q'}}\sqrt{2\pi}\right)^{-1}$ and let $K = \lfloor\sqrt{Np'q'}\rfloor$.

$$\begin{aligned}F_1 \cdot F_2 &> \frac{\binom{N}{Np'+K}\left(\frac{p'}{q'}\right)^K}{\sqrt{2\pi}\binom{N}{Np'}e^{\frac{1}{12Np'q'}}}\\ &= T\left(\frac{p'}{q'}\right)^K\frac{N!(Np')!(N-Np')!}{N!(Np'+K)!(N-Np'-K)!}\\ &= T\left(\frac{p'}{q'}\right)^K\frac{(Nq'-K+1)\cdots(Nq')}{(Np'+1)\cdots(Np+K)}\\ &= T\left(\frac{p'}{q'}\right)^K\left(\prod_{i=1}^{K}\frac{Nq'-K+i}{Np'+i}\right) \qquad (24)\end{aligned}$$

$$\begin{aligned}
&> T \left(\frac{p'}{q'}\right)^{K} \left(\frac{Nq'}{Np'+K}\right)^{K} && (25)\\
&= T\left(1+\frac{K}{Np'}\right)^{-K} &&\\
&\geq T\left(1+\frac{\sqrt{Np'q'}}{Np'}\right)^{-\sqrt{Np'q'}} &&\\
&= T\left(1+\frac{q'}{\sqrt{Np'q'}}\right)^{-\sqrt{Np'q'}} && (26)\\
&\geq Te^{-q'} && (27)\\
&= \frac{1}{\sqrt{2\pi}\, e^{\frac{1}{12Np'q'}}} e^{-q'} && (28)\\
&\geq \frac{1}{\sqrt{2\pi}\, e^{\frac{1}{12\cdot 36}}} e^{-1} \geq 0.14642\ldots && (29)
\end{aligned}$$

In (28) and (29) we used that $Np'q' > 36$, by (19). For the step from (24) to (25) we assume that $Nq' - k \geq Np'$. If $Nq' - k < Np'$ the steps from (25) in the above calculation are replaced by the following:

$$\begin{aligned}
T \left(\frac{p'}{q'}\right)^{K} &\left(\prod_{i=1}^{K} \frac{Nq'-K+i}{Np'+i}\right) && (30)\\
&> T \left(\frac{p'}{q'}\right)^{K} \left(\frac{Nq'-K}{Np'}\right)^{K} && (31)\\
&= T\left(\frac{Nq'-K}{Nq'}\right)^{K} &&\\
&= T\left(\frac{Nq'-\lfloor\sqrt{Np'q'}\rfloor}{Nq'}\right)^{\lfloor\sqrt{Np'q'}\rfloor} &&\\
&\geq T\left(\frac{Nq'-\sqrt{Np'q'}}{Nq'}\right)^{\sqrt{Np'q'}} &&\\
&= T\left(1-\frac{p'}{\sqrt{Np'q'}}\right)^{\sqrt{Np'q'}} && (32)\\
&\geq \frac{1}{\sqrt{2\pi}\, e^{\frac{1}{12Np'q'}}}\ \frac{10}{11}e^{-p'} && (33)\\
&\geq \frac{1}{\sqrt{2\pi}\, e^{\frac{1}{12\cdot 36}}}\ \frac{10}{11}e^{-1} \geq 0.133112\ldots && (34)
\end{aligned}$$

The step from (32) to (33) follows from an elementary analysis of the function $(1-\frac{a}{b})^b - \frac{10}{11}e^{-a}$. Using (34) (which is less than the bound in (29)) and (23) we can lower bound the product $F_1F_2F_3$ as follows:

$$\begin{aligned}
F_1 \cdot F_2 \cdot F_3 &> 0.133112 \cdot \frac{1}{2} \cdot \left(1-\frac{1}{\lfloor\sqrt{Np'q'}\rfloor - 1}\right) &&\\
&\geq 0.066556 \cdot \left(1-\frac{1}{\lfloor\sqrt{36}\rfloor - 1}\right) && (35)\\
&> 0.05324\ldots > \frac{1}{19}\ . && (36)
\end{aligned}$$

For the step from (35) to (36) we again used inequality (19). □