Sample-efficient Strategies for Learning in the Presence of Noise

Nicolò Cesa-Bianchi
University of Milan, Milan, Italy
cesabian@dsi.unimi.it

Eli Dichterman
IBM Haifa Research Laboratory, Israel
eli@haifa.vnet.ibm.com

Paul Fischer
University of Dortmund, Germany
paulf@ls2.informatik.uni-dortmund.de

Eli Shamir
Hebrew University, Jerusalem, Israel
shamir@cs.huji.ac.il

Hans Ulrich Simon
Ruhr-Universität Dortmund, Germany
simon@lmi.ruhr-uni-bochum.de

March 13, 2001

Abstract

In this paper we prove various results about PAC learning in the presence of malicious noise. Our main interest is the sample size behaviour of learning algorithms. We prove the first nontrivial sample complexity lower bound in this model by showing that order of $\varepsilon/\Delta^2 + d/\Delta$ (up to logarithmic factors) 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). We also show that this result cannot be significantly improved in general by presenting efficient learning algorithms for the class of all subsets of d elements and the class of unions of at most d intervals on the real line. This is especially interesting as we can also show that the popular minimum disagreement strategy needs samples of size $d\varepsilon/\Delta^2$, hence is not optimal with respect to sample size. We then discuss the use of randomized hypotheses. For these the bound $\varepsilon/(1+\varepsilon)$ on the noise rate is no longer true and is replaced by $2\varepsilon/(1+2\varepsilon)$. In fact, we present a generic algorithm using randomized hypotheses which can tolerate noise rates slightly larger than $\varepsilon/(1+\varepsilon)$ while using samples of size d/ε as in the noise-free case. Again one observes a quadratic powerlaw (in this case $d\varepsilon/\Delta^2$, $\Delta = 2\varepsilon/(1+2\varepsilon) - \eta$) as Δ goes to zero. We show upper and lower bounds of this order.

Keywords: PAC learning, learning with malicious noise.

To appear in the *Journal of the ACM*.

1 Introduction

Any realistic learning algorithm should be able to cope with errors in the training data. A model of learning in the presence of *malicious noise* was introduced by [17] as an extension of his basic PAC framework for learning classes of $\{0,1\}$-valued functions. In this *malicious PAC model*, each training example given to the learner is independently replaced, with fixed probability η, by an adversarially chosen one (which may or may not be consistent with the $\{0,1\}$-valued target function). In their comprehensive investigation of malicious PAC learning, [8] show that a malicious noise rate $\eta \geq \varepsilon/(1+\varepsilon)$ can make statistically indistinguishable two target functions that differ on a subset of the domain whose probability measure is at least ε. This implies that with this noise rate no learner can generate hypotheses that are ε-good in the PAC sense, this irrespective to the sample size (number of training examples) and to the learner's computational power. In their work, Kearns and Li also analyze the performance of the minimum disagreement strategy in presence of malicious noise. They show that, for a sample of size[1] d/ε (where d is the VC dimension of the target class) and for a noise rate bounded by any constant fraction of $\varepsilon/(1+\varepsilon)$, the hypothesis in the target class having the smallest sample error is ε-good in the PAC sense.

We begin this paper by studying the behaviour of sample complexity in the case of deterministic hypotheses and a high malicious noise rate, i.e., a malicious noise rate arbitrarily close to the information-theoretic upper bound $\eta_{\text{det}} := \varepsilon/(1+\varepsilon)$. We prove the first nontrivial lower bound in this model (Theorem 3.4) 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 $\eta = \varepsilon/(1+\varepsilon) - \Delta$, every class of $\{0,1\}$-valued functions of VC dimension d. Our proof combines, in an original way, techniques from [5, 8, 15] 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 (Theorem 3.10 and Corollary 3.15) that there is an algorithm RMD (for Randomized Minimum Disagreement) that, for each d, learns both the class $\mathcal{C}_d$ of all subsets of d elements and the class $\mathcal{I}_d$ of unions of at most d intervals on the real line using a noisy sample whose size is of the same order as the size of our lower bound. Algorithm RMD makes essential use of the fact that the learning domain is small. Hence, for the class $\mathcal{I}_d$ we first discretize the universe in a suitable way. Then 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 (Theorem 3.9). This bound holds for any class of VC dimension $d \geq 3$ and for every noise rate η such that $\varepsilon/(1+\varepsilon) - \eta = \Delta = o(\varepsilon)$. This implies that, for high noise rate η and for every 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 for which choosing any minimum disagreement hypothesis from a fixed hypothesis class is provably worse, in terms of sample complexity, than a different learning strategy.

In the second part of the paper we consider the use of randomized hypotheses for learning with small sample sizes and high malicious noise rates. An easy modification of Kearns and Li's argument (Proposition 4.1) shows that no learner can output ε-good randomized hypotheses with a noise rate larger or equal to $\eta_{\text{rand}} := 2\varepsilon/(1+2\varepsilon)$. Given the gap between this bound and the corresponding bound $\varepsilon/(1+\varepsilon)$ for learners using deterministic hypotheses, we address the problem whether allowing randomized hypotheses helps in this setting. In fact, we present an

[1]All sample size orders in this section are given up to logarithmic factors.

algorithm (Theorem 4.2) that PAC learns any target class of VC dimension d using randomized hypotheses and d/ε training examples while tolerating any noise rate bounded by a constant fraction of $(7/6)\varepsilon/(1+(7/6)\varepsilon)$. The algorithm works by finding up to three functions in the target class that satisfy a certain independence condition defined on the sample. The value of the final hypothesis on a domain point is then computed by taking a majority vote over these functions (or by tossing a coin in case only two functions are found). Our investigation then moves on to consider the case of a noise rate close to the information-theoretic bound $2\varepsilon/(1+2\varepsilon)$ for randomized hypotheses. We show a strategy (Theorem 4.4) for learning the powerset of d elements using $d\varepsilon/\Delta^2$ training examples and tolerating a malicious noise rate of $\eta = 2\varepsilon/(1+2\varepsilon)-\Delta$, for every $\Delta > 0$. We also show (Theorem 4.5) that this sample size is optimal in the sense that *every* learner using randomized hypotheses needs at least $d\varepsilon/\Delta^2$ training examples for learning any target class of VC dimension d.

2 Definitions and notation

We recall the definitions of PAC learning and malicious PAC learning of a given *target class* $\mathcal{C}$, where $\mathcal{C}$ is a set of $\{0,1\}$-valued functions C, here called *concepts*, defined on some common domain X. We call *instance* every $x \in X$ and *labeled instance* or *example* every pair $(x, y) \in X \times \{0,1\}$. In Valiant's PAC learning model [17], the learning algorithm (or learner) gets as input a *sample*, i.e., a multiset $((x_1, C(x_1)), \ldots, (x_m, C(x_m)))$ of desired size $m < \infty$. Each instance x_t in the sample given to the learner must be independently drawn from the same distribution D on X and labeled according to the same target function $C \in \mathcal{C}$. Both C and D are fixed in advance and unknown to the learner. In the malicious PAC model, the input sample is corrupted by an adversary using noise rate $\eta > 0$ according to the following protocol. First, a sample $((x_1, C(x_1)), \ldots, (x_m, C(x_m)))$ of the desired size is generated exactly as in the noise-free PAC model. Second, before showing the sample to the learner, each example $(x_t, C(x_t))$ is independently marked with fixed probability η. Finally, the adversary replaces each marked example $(x_t, C(x_t))$ in the sample by a pair $(\hat{x}_t, \hat{y}_t)$ arbitrarily chosen from $X \times \{0,1\}$ and then feeds the corrupted sample to the learner. We call the collection of marked examples the *noisy part* of the sample and the collection of unmarked examples the *clean part* of the sample. Note that learning in this model is harder than with the definition of malicious PAC learning given by [8]. There, the examples were sequentially ordered in the sample and the adversary's choice for each marked example had to be based only on the (possibly marked) examples occurring *earlier* in the sample sequence.[2] We call *KL-adversary* this weaker type of adversary. We also consider a third type of malicious adversary, which we call "nonadaptive". Whereas the corruption strategy for our malicious adversary is a function from the set of samples to the set of corrupted samples, the corruption strategy for a nonadaptive adversary is a function from the set of examples to the set of corrupted examples. In other words, a nonadaptive adversary must decide, before seeing the sample, on a fixed rule for replacing each pair $(x, C(x))$, for $x \in X$, with a corrupted pair (x', y'). Note that learning with a KL-adversary is easier than learning with our malicious adversary, but harder than learning with a nonadaptive adversary.

To meet the PAC learning criterion, the learner, on the basis of a polynomially-sized sample (which is corrupted in case of malicious PAC learning), must output an hypothesis H that with high probability is a close approximation of the target C. Formally, an algorithm A is said to *PAC learn* a target class $\mathcal{C}$ using hypothesis class $\mathcal{H}$ if, for all distributions D on X, for all targets $C \in \mathcal{C}$, and for all $1 \geq \varepsilon, \delta > 0$, given as input a sample of size m, A outputs an hypothesis $H \in \mathcal{H}$ such

[2] All the results from [8] we mention here hold in our harder noise model as well.

that its *error probability*, $D(H \neq C)$, is strictly smaller than ε with probability at least $1-\delta$ with respect to the sample random draw, 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. We use $H \neq C$ to denote $\{x : H(x) \neq C(x)\}$. A hypothesis H is called ε-*good* (w.r.t. a distribution D) if it satisfies the condition $D(H \neq C) < \varepsilon$, otherwise it is called ε-*bad*. Similarly, an algorithm A is said to *learn* a target class $\mathcal{C}$ using hypothesis class $\mathcal{H}$ in the malicious PAC model with noise rate η if A learns $\mathcal{C}$ in the PAC model when the input sample is corrupted by any adversary using noise rate η. Motivated by the fact (shown in [8] and mentioned in the introduction) that a noise rate $\eta \geq \varepsilon/(1+\varepsilon)$ forbids PAC learning with accuracy ε, we allow the sample size m to depend polynomially also on $1/\Delta$, where $\Delta = \varepsilon/(1+\varepsilon) - \eta$.

We will occasionally use *randomized* learning algorithms that have a sequence of tosses of a fair coin as additional input source. In this case the definition of PAC learning given above is modified so that $D(C \neq H) < \varepsilon$ must hold with probability at least $1-\delta$ also with respect to the algorithm's randomization. Finally, we will also use *randomized hypotheses* or *coin rules*. A coin rule is any function $F : X \to [0,1]$ where $F(x)$ is interpreted as the probability that the boolean hypothesis defined by the coin rule takes value 1 on x. Coin rules are formally equivalent to *p-concepts*, whose learnability has been investigated by [9]. However, here we focus on a completely different problem, i.e., the malicious PAC learning of boolean functions using p-concepts as hypotheses. If a learner uses coin rules as hypotheses, then the PAC learning criterion $D(C \neq H) < \varepsilon$, where H is the learner's hypothesis, is replaced by $\mathbf{E}_{x \sim D}|F(x) - C(x)| < \varepsilon$, where F is the coin rule output by a learner and $\mathbf{E}_{x \sim D}$ denotes expectation with respect to the distribution D on X. Note that $|F(x) - C(x)|$ is the probability of misclassifying x using coin rule F. Thus $\mathbf{E}_{x \sim D}|F(x) - C(x)|$, which we call the *error probability* of F, is the probability of misclassifying a randomly drawn instance using coin rule F. Furthermore, since Proposition 4.1 in Section 4 shows that every noise rate larger or equal to $2\varepsilon/(1+2\varepsilon)$ prevents PAC learning with accuracy ε using randomized hypotheses, we allow the sample size of algorithms outputting randomized hypotheses to depend polynomially on $1/\Delta$, where $\Delta = 2\varepsilon/(1+2\varepsilon) - \eta$.

In addition to the usual asymptotical notations, let $\widetilde{O}(f)$ be the order obtained from $O(f)$ by dropping polylogarithmic factors.

3 Malicious Noise and Deterministic Hypotheses

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 $\Omega(\varepsilon/\Delta^2 + d/\Delta)$ that holds for any learning algorithm. In Subsection 3.3, we show that this bound cannot be significantly improved. Finally, Theorem 3.9 presents the stronger lower bound $\Omega(d\varepsilon/\Delta^2)$ that holds for the minimum disagreement strategy.

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

$$\Pr\{S_{N,p} \leq \lceil Np \rceil\} \geq \frac{1}{2} \quad \text{and} \quad \Pr\{S_{N,p} \geq \lfloor Np \rfloor\} \geq \frac{1}{2} \ .$$

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

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

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

Proof. The proof is in the Appendix. □

There are, in probability theory, various results that can be used to prove propositions of the form of Fact 3.2. A general but powerful tool is the Berry-Esseen theorem [4, Theorem 1, page 304] on the rate of convergence of the central limit theorem. This gives a lower bound on the left-hand side of (1) and (2) that is worse by approximately a factor of 2 than the bound $1/19$ proven in Fact 3.2. More specialized results on the tails of the Binomial distribution were proven by [1], [2], and [12]. We derive (1) and (2) by direct manipulation of the bound in [1], which is in a form suitable for our purposes.

Fact 3.3 *For every $0 < \beta < \alpha \leq 1$, for every random variable $S \in [0, N]$ with $\mathbf{E}S = \alpha N$, it holds that $\Pr\{S \geq \beta N\} > (\alpha - \beta)/(1 - \beta)$.*

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

$$\alpha N = \mathbf{E}S = \mathbf{E}[S \mid S < \beta N](1 - z) + \mathbf{E}[S \mid S \geq \beta N] z < \beta N(1 - z) + Nz \; .$$

□

3.1 A General Lower Bound on the Sample Size

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. [8] have shown 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 nonadaptive adversary will present x with the false label with probability $\eta_{\text{det}} := \varepsilon/(1+\varepsilon)$. Hence, x appears with the true label with probability $(1 - \eta_{\text{det}})\varepsilon$. As $(1 - \eta_{\text{det}})\varepsilon = \eta_{\text{det}}$, there is no chance to distinguish between C_0 and C_1.

Our first lower bound is based on a similar reasoning: For $\eta < \eta_{\text{det}}$, the targets C_0 and C_1 can be distinguished, but as η approaches η_{det}, the discrimination task becomes arbitrarily hard. This idea is made precise in the proof of the following result.

Theorem 3.4 *For every nontrivial target class $\mathcal{C}$, for every $0 < \varepsilon < 1$, $0 < \delta \leq 1/38$, 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 adversary that corrupts examples with probability η. Let A be a (possibly randomized) learning algorithm for $\mathcal{C}$ that uses sample size $m = m(\varepsilon, \delta, \Delta)$. For the purpose of contradiction, assume that A PAC learns $\mathcal{C}$ against the (nonadaptive) adversary that flips a fair coin to select target $C \in \{C_0, C_1\}$ and then returns a corrupted sample S' of size m whose examples in the noisy part are all equal to $(x_1, 1 - C(x_1))$.

Let $p_A(m)$ be $\Pr\{H \neq C\}$, where H is the hypothesis generated by A on input S'. Since $H(x_1) \neq C(x_1)$ implies that H is not an ε-good hypothesis, we have that $p_A(m) \leq \delta \leq 1/38$ must hold. For the above malicious adversary, 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, 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 the hypothesis returned by B is wrong. Based on the following two claims, we will show that whenever m is 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 $(1 - \eta - \Delta)\varepsilon m = (1 - \eta_{\text{det}})\varepsilon m = \eta_{\text{det}} m = (\eta + \Delta)m$ 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* $\Pr\{BAD_1(m)\} > 1/19$.

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

$$\frac{1}{2}\sqrt{m\eta(1-\eta) - 1} \geq 3 \qquad \text{and} \qquad \frac{1}{2}\sqrt{m\eta(1-\eta) - 1} \geq \Delta m.$$

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 the Bayes strategy 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 [5] 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, for which $D(x) = \eta/(1-\eta)$. 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 observed very rarely in the sample. 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 every $0 < \varepsilon \leq 1/8$, $0 < \delta \leq 1/12$, and every $0 < \Delta < \varepsilon/(1+\varepsilon)$, 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 $X_0 = \{x_0, x_1, \ldots, x_t, x_{t+1}\}$ be a 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:

$$\begin{aligned} D(x_0) &= 1 - \frac{\eta}{1-\eta} - 8\left(\varepsilon - \frac{\eta}{1-\eta}\right), \\ D(x_1) = \cdots = D(x_t) &= \frac{8\left(\varepsilon - \frac{\eta}{1-\eta}\right)}{t} \\ D(x_{t+1}) &= \frac{\eta}{1-\eta}. \end{aligned}$$

(Note that $\varepsilon \leq 1/8$ implies that $D(x_0) \geq 0$.) We will use a nonadaptive adversary that, with probability η, corrupts each 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_{det} the probability to select a rare point approaches 0. Let A be a (possibly randomized) learning algorithm for $\mathcal{C}$ using sample size $m = m(\varepsilon, \delta, \Delta)$. Consider the adversary that flips a fair coin to select target $C \in \mathcal{C}$ and then returns a corrupted sample (of size m) whose examples in the noisy part are all equal to $(x_{t+1}, 1 - C(x_{t+1}))$.

Let e_A be the random variable denoting the error $\Pr\{H \neq C\}$ of A's hypothesis H. Then, by pigeonhole principle,

$$\Pr\{e_A \geq \varepsilon\} > \frac{1}{12} \tag{3}$$

implies the existence of a concept $C_0 \in \mathcal{C}$ such that the probability that A does not output an ε-good 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 adversary. In what follows, we call *unseen* the rare points that are not returned by the adversary. Given BAD_1, let BAD_2 be the event that the hypothesis H classifies incorrectly at least $t/8$ points among the set of $t/2$ unseen points with lowest indices. Finally, let BAD_3 be 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 $m \le \frac{t}{32\Delta(1+\varepsilon)}$, the examples returned by the adversary contain at most $t/4$ rare points in 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 $1/3$. Note that events BAD_1 and BAD_2 do not break the symmetry between the two possible labels for x_{t+1} (although they change the probability that sample point x_{t+1} is drawn at all). 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. □

Combining Theorems 3.4 and 3.7 we have

Corollary 3.8 *For any nontrivial target class $\mathcal{C}$ with, and for every $0 < \varepsilon \le 1/8$, $0 < \eta \le 1/38$, and every $0 < \Delta < \varepsilon/(1+\varepsilon)$, 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*

$$\Omega\left(\frac{d}{\Delta} + \frac{\eta}{\Delta^2}\right) = \Omega\left(\frac{d}{\Delta} + \frac{\varepsilon}{\Delta^2}\right)$$

Even though the lower bound $\Omega(\eta/\Delta^2 + d/\Delta)$ holds for the nonadaptive adversary, in Subsection 3.3 we show a matching upper bound (ignoring logarithmic factors) that holds for the (stronger) malicious adversary.

The lower bound proven in Theorem 3.4 has been recently improved to

$$\Omega\left(\frac{\eta}{\Delta^2} \ln \frac{1}{\delta}\right)$$

by [6], who introduced an elegant information-theoretic approach avoiding the analysis of the Bayes risk associated with the learning problem. It is not clear whether their approach can be applied to obtain also the other lower bound term, $\Omega(d/\Delta)$, that we prove in Theorem 3.7.

The next result is a lower bound on the sample size needed by the minimum disagreement strategy (called MDS henceafter.)

3.2 A Lower Bound on the Sample Size for MDS

Given any sample S' of corrupted training examples, MDS outputs the hypothesis $H \in \mathcal{C}$ with the fewest disagreements on S'.

Theorem 3.9 *For any target class $\mathcal{C}$ with VC dimension $d \geq 3$, every $0 < \varepsilon \leq 1/38$, $0 < \delta \leq 1/74$, and every $0 < \Delta = o(\varepsilon)$, the sample size needed by the Minimum Disagreement strategy for learning $\mathcal{C}$ 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 adversary would have no chance to fool MDS. However, if the sample size m is too small, the standard deviation σ for the number of true labels of a rare point x gets very big. Hence, with 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 malicious adversary 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 now 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 malicious adversary. 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 a set of points shattered by $\mathcal{C}$. 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

$$\begin{aligned} \Pr\{x_i \text{ is hidden}\} &= \Pr\{x_i \text{ is hidden} \mid BAD_1\} \Pr\{BAD_1\} \\ &+ \Pr\{x_i \text{ is hidden} \mid \neg BAD_1\} (1 - \Pr\{BAD_1\}) \end{aligned}$$

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 as follows

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

The number of false labels that the adversary 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-bad 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:

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

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

It is an open question whether a similar lower bound can be proven for the KL-adversary, or even the weaker nonadaptive adversary.

3.3 Learning the Powerset and k-Intervals with Deterministic Hypotheses

Let $\mathcal{C}_k$ be the class of all subsets over k points and let $\mathcal{I}_k$ be the class of unions of at most k intervals on the unit interval $[0,1]$. The Vapnik-Chervonenkis dimension of $\mathcal{I}_k$ is $2k$. In this subsection, we show that the lower bound proven in Subsection 3.1 cannot be improved in general. That is, we show that for each $d \geq 1$, the class $\mathcal{C}_d$ of all subsets over d points and the class $\mathcal{I}_k$, $k = d/2$, 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 $\Delta = \varepsilon/(1+\varepsilon) - \eta$.

According to Subsection 3.2, strategy MDS is not optimal for PAC learning powersets in the presence of malicious noise. Instead, we use a modification of MDS, called RMD-POW, which uses a majority vote on the sample (like MDS) to decide the labels of some of the domain points and tosses a fair coin to decide the labels of the remaining ones. The heart of RMD-POW is the subroutine which splits the domain into two appropriate subgroups.

It turns out that the result for k-intervals can be derived as a corollary to the result for powersets if the latter result is proven in a slightly generalized form. Instead of considering $\mathcal{C}_d$ and an arbitrary distribution D on $\{1,\ldots,d\}$, we consider $\mathcal{C}_N$, $N \geq d$, and a restricted subclass of distributions, where the restriction will become vacuous in the special case that $N = d$. Thus, although the result is distribution-specific in general, it implies a distribution-independent result for $\mathcal{C}_d$.

The restriction which proves useful later can be formalized as follows. Given parameters d, Δ and a distribution D on $X_N = \{1,\ldots,N\}$, we call point $i \in X_N$ *light* if $D(i) < \Delta/(3d)$, and *heavy* otherwise. We say that D is a *legal distribution* on X_N if it induces at most d light points. The main use of this definition is that even a hypothesis that is incorrect on all light points is only charged by less than $\Delta/3$ (plus its total error on heavy points, of course). Note furthermore that the existence of a legal distribution on X_N implies that $N \leq d + 3d/\Delta$, and every distribution is legal if $N = d$.

The rest of this subsection is structured as follows. We first prove a general result for powerset $\mathcal{C}_N$, $N \geq d$, and legal distributions with the obvious corollary for powerset $\mathcal{C}_d$ and arbitrary distributions. Afterwards, we present a general framework of partitioning the domain X of a concept class into so-called *bins* in such a way that labels can be assigned binwise without much damage. A bin $B \subseteq X$ is called *homogeneous* with respect to target concept C if C assigns the same label to all points in B. Otherwise, B is called *heterogeneous*. If there are at most d heterogeneous bins of total probability at most $\Delta/3$, we may apply the subroutine for $\mathcal{C}_N$ (working well for legal domain distributions), where heterogeneous bins play the role of light points and homogeneous bins the role of heavy (or another kind of unproblematic) points. From these general considerations, we derive the learnability result for k-intervals by showing that (for this specific concept class) an appropriate partition of the domain can be efficiently computed from a given sample with high probability of success.

Theorem 3.10 *There exists a randomized algorithm,* RMD-POW, *which achieves the following. For every $N \geq d \geq 1$ and every $1 \geq \varepsilon, \delta, \Delta > 0$,* RMD-POW *PAC learns the class $\mathcal{C}_N$ under legal distributions with accuracy ε, confidence δ, tolerating malicious noise rate $\eta = \varepsilon/(1+\varepsilon) - \Delta$, and using a sample of size $\widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$.*

Proof. Given parameters Δ, d and $N \leq d + 3d/\Delta$, a learning algorithm must deliver a good approximation to an unknown subset C of domain $X_N = \{1,\ldots,N\}$ with high probability of success. Let D denote the unknown domain distribution, and $p_i = D(i)$ for $i = 1,\ldots,N$. It is

Input: Parameters α, L, n. Domain size d, accuracy ε, confidence δ.

- Get sample $(i_1, y_1), \ldots, (i_m, y_m)$ of size $m = \widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$.
- For each point $i \in X_N = \{1, \ldots, N\}$.
  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-POW (see Theorem 3.10).

assumed that D induces at most d light points, i.e.,

$$I_{\text{light}} = \left\{ i \in X_N \,:\, p_i < \frac{\Delta}{3d} \right\}$$

contains at most d elements. Clearly, the total error (of any hypothesis) on light points is bounded by $D(I_{\text{light}}) < \Delta/3$. It is therefore sufficent to deliver, with probability at least $1 - \delta$ of success, a hypothesis whose total error on

$$I_{\text{heavy}} = \left\{ i \in X_N \,:\, p_i \geq \frac{\Delta}{3d} \right\}$$

is bounded by $\varepsilon - \Delta/3$. In the sequel, we describe the algorithm RMD-POW and show that it achieves this goal. RMD-POW is parameterized by three parameters α, L, n. We denote the logarithm to base β by $\log_\beta$ and assume that α, L, n are initialized as follows:

$$\alpha = \sqrt[3]{5/3} - 1 \tag{7}$$

$$L = \left\lceil \log_{1+\alpha} \frac{6d(1+\alpha)^2\varepsilon}{\Delta} \right\rceil \tag{8}$$

$$n = \lceil 50 \ln(4L/\delta) \rceil \tag{9}$$

An informal description of RMD-POW is given below. Its pseudo-code may be found in Figure 1. Algorithm RMD-POW can deliver any subset of X_N as hypothesis, and can therefore decide the label of each point in X_N 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.

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 $(i_1, y_1), \ldots, (i_m, y_m)$ drawn from the set $\{1, \ldots, N\} \times \{0, 1\}$, let $\nu_{0,i}$ and $\nu_{1,i}$ be the frequencies with which each point $i \in \{1, \ldots, N\}$ 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 $u_i = \max\{\nu_{0,i}, \nu_{1,i}\}$. We say that a domain point i is in *strong majority* (with respect to the sample and to parameter α defined in (7)) if $u_i > (1+\alpha)\ell_i$, and is in *weak majority* otherwise. We divide some of the points into L *bands,*, where L is the parameter defined in (8). A point i is in band k, for $k = 1, \ldots, L$, if i is in weak majority and $(1+\alpha)^{-k}\varepsilon < \ell_i \le (1+\alpha)^{1-k}\varepsilon$. We further divide the bands into *sparse bands*, containing less than n elements, and *dense bands*, containing at least n elements, where n is the parameter defined in (9). 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-POW'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

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

First, we upper bound in probability the sum $f_{\text{maj}} + f_{\text{sparse}} + f_{\text{dense}}$. Let $\widehat{\eta}$ be the frequency of corrupted examples in the sample. By using (25) from Lemma A.1 with $p = \eta$ and $\lambda = \Delta/(3\eta)$, we find that

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

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 incorrupted 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, $|I_{\text{heavy}}| \le N$, as there are at most N points. By using (23) from Lemma A.1 with $p = \Delta/(3d)$ and $\lambda = \alpha/(1+\alpha)$, we find that

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

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

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

Thus, there is a sample size with order $\widetilde{O}(\eta/\Delta^2 + d/\Delta)$ of magnitude, such that (10) and (11) simultaneously hold with probability at least $1 - \delta/2$. At this point recall that N linearly depends on d.

Let $I_{\text{wrong}} = \{i \in I_{\text{heavy}} : C(i) \ne H(i)\}$. Remember that $D(I_{\text{light}}) \le \Delta/3$ and that it suffices to bound in probability $D(I_{\text{wrong}}) + \Delta/3$ from above by ε. Claim 3.14 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

$$D(I_{\text{wrong}}) \le D(I_{\text{wrong}} \cap I_{\text{maj}}) + D(I_{\text{wrong}} \cap I_{\text{sparse}}) + D(I_{\text{wrong}} \cap I_{\text{dense}}). \tag{12}$$

Now, Claims 3.11–3.13 show how the three terms in the right-hand side of (12) can be simultaneously bounded. In the rest of this subsection, we prove Claims 3.11–3.14. We start by bounding the error made by H on heavy points $i \in I_{\mathrm{maj}}$.

Claim 3.11 (Strong majority) *If (11) holds, then*

$$D(I_{\mathrm{wrong}} \cap I_{\mathrm{maj}}) \le \frac{f_{\mathrm{maj}}}{1-\eta}.$$

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

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

Claim 3.12 (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 $D(I_{\mathrm{wrong}} \cap I_{\mathrm{sparse}}) \le \frac{f_{\mathrm{sparse}}}{1-\eta} + \frac{\Delta}{(1-\eta)3}$ 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

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

holds in probability. To show this, we use (23) from Lemma A.1 to write the following

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

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 $i \in I_{\mathrm{heavy}}$ and $t_i \ge \varepsilon$. As $\widehat{\eta} < \varepsilon$ by (10), $i \notin I_{\mathrm{wrong}}$. Hence, (10) and (11) imply that $p_i \le (1+\alpha)\varepsilon/(1-\eta)$ holds for all $i \in I_{\mathrm{wrong}}$. We then apply (14) to each $i \in I_{\mathrm{wrong}} \cap I_{\mathrm{sparse}}$. 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

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

This concludes the proof. □

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

Claim 3.13 (Dense bands) *If (11) holds, then*

$$D(I_{\text{wrong}} \cap I_{\text{dense}}) \leq \frac{f_{\text{dense}}}{1-\eta}$$

holds with probability at least $1 - \delta/4$ *with respect to the algorithm randomization.*

Proof. For each $k = 1, \ldots, L$, let I_k be the set of all heavy points in the k-th band. Furthermore, let $t^k_{\max} = \max\{t_i : i \in I_k \cap I_{\text{wrong}}\}$ and $f^k_{\min} = \min\{f_i : i \in I_k \cap I_{\text{wrong}}\}$. Since all points in I_k are in weak majority and by definition of bands, we have that $t^k_{\max} \leq (1+\alpha)^2 f^k_{\min}$ holds for each $k = 1, \ldots, L$. Furthermore, using (11), $p_j \leq \frac{1+\alpha}{1-\eta} t_j$, for each $j \in I_k$. As for each dense band $|I_k| \geq n \geq 50 \ln(4L/\delta)$, using (24) from Lemma A.1 we can guarantee that $|I_k \cap I_{\text{wrong}}| \leq \frac{3}{5}|I_k|$ holds simultaneously for all bands $k = 1, \ldots, L$ with probability at least $1 - \delta/4$. Combining everything we get

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

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

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

concluding the proof. □

Claim 3.14 *If (10) and (11) hold, then* $I_{\text{heavy}} \subseteq I_{\text{maj}} \cup I_{\text{sparse}} \cup I_{\text{dense}}$.

Proof. We have to show that each heavy point i in weak majority belongs to a band, or equivalently,

$$(1+\alpha)^{-L}\varepsilon < \ell_i \leq \varepsilon \text{ for all } i \in I_{\text{heavy}} \setminus I_{\text{maj}}.$$

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 $i \in I_{\text{heavy}} \setminus I_{\text{maj}}$,

$$\ell_i \geq \frac{u_i}{1+\alpha} \geq \frac{t_i}{1+\alpha} \geq \frac{1-\eta}{(1+\alpha)^2} p_i \geq \frac{(1-\eta)\Delta}{3d(1+\alpha)^2}.$$

Using the definition of L and the fact that $\eta < 1/2$, we can conclude the proof of the claim as follows:

$$(1+\alpha^{-L}\varepsilon < (1+\alpha)^{\log_{1+\alpha}(6d(1+\alpha)^2\varepsilon/\Delta)}\varepsilon \leq \frac{\Delta}{6d(1+\alpha)^2} < \frac{(1-\eta)\Delta}{3d(1+\alpha)^2}.$$

□

Putting (10), (11) and the preceding claims together, we find that, with probability at least $1-\delta$, a hypothesis H is delivered such that

$$\begin{array}{rcl} D(\{i : H(i) \neq C(i)\} & < & D(I_{\text{wrong}}) + \Delta/3 \\ & \leq & \dfrac{f_{\text{maj}} + f_{\text{sparse}} + f_{\text{dense}} + \Delta/3}{1-\eta} + \Delta/3 \\ & \leq & \dfrac{\eta + \Delta}{1-\eta} \\ & \leq & \dfrac{\eta_0}{1-\eta_0} \\ & = & \varepsilon. \end{array}$$

This concludes the proof. □

We want to use the algorithm for powerset $\mathcal{C}_N$ and legal distributions as a subroutine to learn other concept classes. For this purpose, we need the following observations concerning the proof of Theorem 3.10:

- The total error on heavy points is bounded above (in probability) by $\frac{f_{\text{heavy}}+\Delta/3}{1-\eta}$, where f_{heavy} denotes the sample frequency of corrupted examples associated with heavy points.
- The required sample size does not change its order of magnitude if heavy points i were only "heavy up to a constant", that is $D(i) \geq \Delta/(kd)$ for some fixed but arbitrary constant $k \geq 3$.

Imagine that the learning algorithm gets, as auxiliary information in addition to the corrupted sample, a partition of the domain X into bins $B_1, \ldots, B_N$ such that the following holds:

- $D(B_i) \leq \Delta/(3d)$ for each heterogeneous bin B_i. Here d denotes the VC dimension of the target class and D the domain distribution.
- There are at most d heterogeneous bins.
- Each homogeneous bin B_i satisfies at least one of the following properties:

  Almost Heavy Bin $D(B_i) \geq \Delta/(144d)$.

  Unprofitable Bin $f_i \geq D(B_i)$, where f_i denotes the sample frequency of corrupted examples associated with sample points hitting B_i.

  Nice Bin The corrupted sample shows the true label of B_i with a strong majority over the wrong label.

It should be clear from the proof of Theroem 3.10 that, given a sufficiently large corrupted sample and the auxiliary information, RMD-POW can be succesfully applied with heterogeneous bins in the role of light points and almost heavy homogeneous bins in the role of heavy points. For nice bins, RMD-POW will find the correct label. On unprofitable bins, RMD-POW might fail, but this does not hurt because the adversary's "investment" was too large. These considerations lead to the following

Corollary 3.15 *There exists a randomized algorithm,* RMD-kINV, *which achieves the following. For every $k \geq 1$ and every $1 \geq \varepsilon, \delta, \Delta > 0$,* RMD-$k$INV *PAC learns the class $\mathcal{I}_k$ with accuracy ε, confidence δ, tolerating malicious noise rate $\eta = \varepsilon/(1+\varepsilon) - \Delta$, and using a sample of size $\widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$ with $d = 2k$.*

Proof. Note that concept class $\mathcal{I}_k$ has VC dimension $d = 2k$. According to the preceding discussion, it suffices to show that an appropriate partition of $[0,1]$ into bins can be computed with probability at least $1 - \delta/2$ of success, using a (corrupted) sample of size $2m = \widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$.

The bin partition is computed from the given corrupted sample S in two stages. Stage 1 uses the first m sample points (subsample S_1) and Stage 2 the last m sample points (subsample S_2).

Stage 1 Each point $i \in [0,1]$ occuring in S_1 with a relative frequency of at least $\Delta/(24d)$ is put into a separate *point-bin*. Let $B_1, \ldots, B_{N_1}$ denote the resulting sequence of point-bins. The points in subdomain $X_2 := [0,1] \setminus (B_1 \cup \cdots \cup B_{N_1})$ are called *empirically light*.

Stage 2 In order to put the points from X_2 into bins, a sorted list S_2' (in increasing order), containing all points from X_2 occuring at least once in S_2, is processed from left to right. We keep track of a variable f which is initialized to zero and sums up the subsample frequencies (w.r.t. S_2) of the points from S_2' that are already processed. As soon as f reaches threshold $\Delta/(24d)$, we complete the next bin and reset f to zero. For instance, if f reaches this threshold for the first time when $z_1 \in S_2'$ is currently processed, we create bin $[0, z_1]$, reset f to zero, and proceed with the successor of z_1 in S_2'. When f reaches the threshold for the second time when $z_2' \in S_2'$ is currently processed, bin $(z_1, z_2]$ is created, and so on. The last bin B'_{N_2} has endpoint $z_{N_2} = 1$ (belonging to B'_{N_2} iff 1 is empirically light).[3]

Note that the bin partition does only depend on the instances within S and not on their labels. (Clearly, the labels are used when we apply procedure RMD-POW as a subroutine in order to determine the bin labels.)

If sample size $2m$ is appropriately chosen from $\widetilde{O}(\varepsilon/\Delta^2 + d/\Delta)$, the following holds with probabiility at least $1 - \delta/2$ of success:

Condition 1 $\hat{\eta} \le \eta + \Delta/3 < 1/2$, where $\hat{\eta} = \max\{\hat{\eta}_1, \hat{\eta}_2\}$ and $\hat{\eta}_i$ denotes the fraction of corrupted examples within subsample S_i.

Condition 2 $D(i) < \Delta/(6d)$ for all empirically light points i.

Condition 3 For every bin B created in Stage 2 we have $D(B) < \Delta/(3d)$.

For Condition 1, we can argue as in the proof of Theorem 3.10.
Loosely speaking, Condition 2 follows because the adversary cannot make points appearing substantially lighter than they actually are. A more formal argument is as follows. The class of singletons over domain $[0,1]$ has VC dimension 1. Assume $D(i) \ge \Delta/(6d)$. An easy application of Lemma A.2 shows (in probability) that i occurs in the uncorrupted subsample for Stage 1 (whose corruption yields S_1) more than $\Delta/(12d)$ times. Since $\hat{\eta}_1 < 1/2$, it occurs in the corrupted subsample S_1 more than $\Delta/(24d)$ times. This implies that each empirically light point has probability less than $\Delta/(6d)$.
Condition 3 can be seen as follows. The class of intervals over subdomain X_2 has VC dimension 2. Consider a bin B_j' created in Stage 2. Let $B_j'' := B_j' \setminus \{z_j\}$. Note that the relative frequency of hitting B_j'' with points from S_2 is smaller than $\Delta/(24d)$. Applying the same kind of reasoning as for empirically light points in Stage 1, it follows (in probability) that $D(B_j'') < \Delta/(6d)$. If z_j belongs to B_j' (which is always the case unless perhaps $j = N_2$ and $z_j = 1$), then z_j is empirically light. Applying Condition 2, we get $D(B_j') < \Delta/(3d)$.

[3]If variable f is below its threshold when endpoint 1 is reached, the last bin must formally be treated like a heterogeneous bin. For the sake of simplicity, we ignore this complication in the sequel.

We conclude the proof by showing that these three conditions imply that the bin partition has the required properties. Clearly, all point-bins are homogeneous. Since the at most k intervals of the target concept have at most $d = 2k$ endpoints, at most d of the bins created in Stage 2 are heterogeneous. Thus, there are at most d heterogeneous bins altogether, and each of them has probability at most $\Delta/(3d)$. Each point-bin is hit by S_1 with a relative frequency of at least $\Delta/(24d)$. Similarly, each bin created in Stage 2 is hit by S_2 with a relative frequency of at least $\Delta/(24d)$. Since $|S_1| = |S_2| = m$, each homogeneous bin is hit by S with a relative frequency of at least $\Delta/(48d)$. Let B be a homogeneous bin. Remember that we consider B as almost heavy if $D(B) \geq \Delta/(144d)$. Assume B is not almost heavy. If the sample frequency of corrupted examples hitting B is at least $\Delta/(144d)$, then B is unprofitable. Assume this is not the case. But then S presents the true label of B with a relative frequency of at least $\Delta/(48d) - \Delta/(144d) = \Delta/(24d)$, twice as often as the wrong label. It follows that B is a nice bin. This concludes the proof of the corollary. $\square$

For class $\mathcal{I}_k$, we used a bin class of constant VC dimension with at most d heterogeneous bins. We state without proof that a partition using a bin class of VC dimension d_1 and at most d_2 heterogeneous bins leads to an application of algorithm RMD-POW requiring an additional term of order $\tilde{O}(d_1 d_2/\Delta)$ in the sample size. As long as $d_1 d_2 = O(d)$, the sample size has the same order of magnitude as in Corollary 3.15.

4 Malicious Noise and Randomized Hypotheses

In this section, we investigate the power of randomized hypotheses for malicious PAC learning. We start by observing that an easy modification of [8, Theorem 1] yields the following result. (Recall that a target class is nontrivial if it contains two functions C and C' and there exist two distinct points $x, x' \in X$ such that $C(x) = C'(x) = 1$ and $C(x') \neq C'(x')$).

Proposition 4.1 *For all nontrivial target classes $\mathcal{C}$ and all $\varepsilon < 1/2$, no algorithm can learn $\mathcal{C}$ with accuracy ε, even using randomized hypotheses, and tolerating malicious noise rate $\eta \geq 2\varepsilon/(1+2\varepsilon)$.*

Let $\eta_{\mathrm{rand}} = \eta_{\mathrm{rand}}(\varepsilon) = 2\varepsilon/(1+2\varepsilon)$ (we omit the dependence on ε when it is clear from the context.) As the corresponding information-theoretic bound $\eta_{\mathrm{det}} = \varepsilon/(1+\varepsilon)$ for learners using deterministic hypotheses is strictly smaller than η_{rand}, one might ask whether this gap is real, i.e., whether randomized hypotheses really help in this setting. In Subsection 4.1 we give a positive answer to this question by showing a general strategy that, using randomized hypotheses, learns any target class $\mathcal{C}$ tolerating any malicious noise rate η bounded by a constant fraction of $\frac{(7/6)\varepsilon}{1+(7/6)\varepsilon}$ and using sample size $\widetilde{O}(d/\varepsilon)$, where d is the VC dimension of $\mathcal{C}$. Note that $\frac{(7/6)\varepsilon}{1+(7/6)\varepsilon} > \eta_{\mathrm{det}}$, whereas no learner using deterministic hypotheses can tolerate a malicious noise rate $\eta \geq \eta_{\mathrm{det}}$, even allowing an *infinite* sample. Furthermore, the sample size used by our strategy is actually *independent* of η and is of the same order as the one needed in the noise-free case. Finally, in Subsection 4.2 we show an algorithm for learning the powerset of d points, for every $d \geq 1$, with malicious noise rates arbitrarily close to η_{rand}.

The problem of finding a general strategy for learning an arbitrary concept class with randomized hypotheses and malicious noise rate arbitrarily close to $2\varepsilon/(1+2\varepsilon)$ remains open.

4.1 A General Upper Bound for Low Noise Rates

We show the following result.

Theorem 4.2 *For every target class $\mathcal{C}$ with VC dimension d, every $0 < \varepsilon, \delta \le 1$, and every fixed constant $0 \le c < 7/6$, a sample size of order d/ε (ignoring logarithmic factors) is necessary and sufficient for PAC learning $\mathcal{C}$ using randomized hypotheses, with accuracy ε, confidence δ, and tolerating malicious noise rate $\eta = c\varepsilon/(1 + c\varepsilon)$.*

This, combined with Theorems 3.4 and 3.7 shows that, if the noise rate is about η_{det}, then a sample size of order $d/\Delta + \varepsilon/\Delta^2$ is only needed if the final hypothesis has to be deterministic.

The idea behind the proof of Theorem 4.2 is the following. A sample size $\tilde{O}(d/\varepsilon)$ is too small to reliably discriminate the target (or other ε-good concepts) from ε-bad concepts. (Actually, the adversary can make ε-bad hypotheses perform better on the sample than the target C.) It is, however, possible to work in two phases as follows. Phase 1 removes some concepts from $\mathcal{C}$. It is guaranteed that all concept with an error rate "significantly larger" than ε are removed, and that the target C is not removed. Phase 2 reliably checks whether two concepts are independent in the sense that they have a "small" joint error probability (the probability to produce a wrong prediction on the same randomly drawn example). Obviously, the majority vote of three pairwise independent hypotheses has error probability at most 3 times the "small" joint error probability of two independent concepts. If there is no independent set of size 3, there must be a maximal independent set $\mathcal{U}$ of size 2 (or 1). We will show that each maximal independent set contains a concept with error probability significantly smaller than ε. It turns out that, if either $\mathcal{U} = \{G\}$ or $\mathcal{U} = \{G, H\}$, then either G or the coin rule $\frac{1}{2}(G + H)$, respectively, is ε-good. The resulting algorithm, SIH (which stands for Search Independent Hypotheses), is described in Figure 2.

Proof. (Of Theorem 4.2.) The lower bound is obvious because it holds for the noise-free case [5]. For the upper bound, we begin with the following preliminary considerations. Let X be the domain of target class $\mathcal{C}$, D any distribution on X, and $C \in \mathcal{C}$ be the target. Given a hypothesis $H \in \mathcal{C}$, $E(H) = \{x : H(x) \neq C(x)\}$ denotes its *error set*, and $\mathrm{err}(H) = D(E(H))$ its error probability. The *joint error probability* of two hypotheses $G, H \in \mathcal{C}$ is given by $\mathrm{err}(G, H) = D(E(G) \cap E(H))$. Our proof will be based on the fact that (joint) error probabilities can be accurately empirically estimated. Let S be the sample. We denote the relative frequency of mistakes of H on the whole sample S by $\widehat{\mathrm{err}}(H)$. The partition of S into a clean and a noisy part leads to the decomposition $\widehat{\mathrm{err}}(H) = \widehat{\mathrm{err}}^c(H) + \widehat{\mathrm{err}}^n(H)$, where upper indices c and n refer to the clean and the noisy part of the sample, respectively. Note that term $\widehat{\mathrm{err}}^c(H)$ is an empirical estimation of $\mathrm{err}(H)$, whereas $\widehat{\mathrm{err}}^n(H)$ is under control of the adversary. The terms $\widehat{\mathrm{err}}(G, H)$, $\widehat{\mathrm{err}}^c(G, H)$, and $\widehat{\mathrm{err}}^n(G, H)$ are defined analogously.

Let $\gamma, \lambda > 0$ denote two fixed constants to be determined by the analysis and let $\hat{\eta}$ denote the empirical noise rate. A standard application of the Chernoff-Hoeffding bound (23) and Lemma A.2 shows that, for a suitable choice of $m = \tilde{O}(d/\varepsilon)$, the following conditions are simultaneously satisfied with probability $1 - \delta$:

Condition 1 $\hat{\eta} \le (1 + \gamma)\eta$.

Condition 2 If $H \in \mathcal{C}$ satisfies $\mathrm{err}(H) \ge \lambda\varepsilon$, then $\widehat{\mathrm{err}}^c(H) \ge (1 - \gamma)(1 - \eta)\mathrm{err}(H)$.

Condition 3 If $G, H \in \mathcal{C}$ satisfy $\mathrm{err}(G, H) \ge \lambda\varepsilon$, then

$$\widehat{\mathrm{err}}^c(G, H) \ge (1 - \gamma)(1 - \eta)\mathrm{err}(G, H) .$$

We just mention that for proving Conditions 2 and 3 we use the fact that the VC dimensions of the classes of the error sets and the joint error sets are both $O(d)$.

Input: Target class $\mathcal{C}$, sample S. Accuracy and confidence parameters ε, δ, upper bound $\eta < \dfrac{(7/6)\varepsilon}{1+(7/6)\varepsilon}$ on true malicious noise.

Initialization: Let $\gamma = \dfrac{(7/6)-c}{(7/6)+c}$, where $c = \dfrac{\eta}{(1-\eta)\varepsilon}$.

Phase 1.

1. Remove from $\mathcal{C}$ all concepts H such that $\widehat{\text{err}}(H) > (1+\gamma)\eta$.
   Let $\mathcal{K}$ be the set of remaining concepts.
2. If $\mathcal{K}$ is empty, then output a default hypothesis.

Phase 2.

1. If there exists an independent set $\{F, G, H\} \subseteq \mathcal{K}$, then output the majority vote of F, G, and H.
2. Otherwise, pick a maximal independent set $\mathcal{U}$ of size 1 or 2.
   If $\mathcal{U} = \{H\}$, then output H. If $\mathcal{U} = \{G, H\}$, then output the coin rule $\frac{1}{2}(G+H)$.

Figure 2: A description of the randomized algorithm SIH used in the proof of Theorem 4.2.

We now describe the learning algorithm SIH illustrated in Figure 2.

In Phase 1, all concepts $H \in \mathcal{C}$ satisfying $\widehat{\text{err}}(H) > (1+\gamma)\eta$ are removed. Let

$$\mathcal{K} = \{H \in \mathcal{C} \,:\, \widehat{\text{err}}(H) \leq (1+\gamma)\eta\}$$

denote the set of remaining concepts. Note that $\widehat{\text{err}}(C) \leq \hat{\eta}$. Applying Condition 1, it follows that target C belongs to $\mathcal{K}$. Applying Condition 2 with constant $\lambda \leq c(1+\gamma)/(1-\gamma)$, it follows that all concepts $H \in \mathcal{K}$ satisfy:

$$\text{err}(H) \leq \frac{(1+\gamma)\eta}{(1-\gamma)(1-\eta)} \; . \tag{15}$$

We are now in position to formalize the notion of independence, which is central for Phase 2 of SIH. Let us introduce another parameter α whose value will also be determined by the analysis. We say that $G, H \in \mathcal{K}$ are *independent* if $\widehat{\text{err}}(G, H) \leq (\alpha+\gamma)\eta$. A subset $\mathcal{U} \subseteq \mathcal{K}$ is called independent if its hypotheses are pairwise independent.

Claim. If $\widehat{\text{err}}^c(H) \geq (1-\alpha)\eta$ then H and C are independent.

To prove the claim note that, since C is the target, $\widehat{\text{err}}(H, C) \leq \widehat{\text{err}}^n(H)$. The definition of $\mathcal{K}$ and the decomposition of $\widehat{\text{err}}$ into $\widehat{\text{err}}^c$ and $\widehat{\text{err}}^n$ imply that

$$\widehat{\text{err}}^n(H) = \widehat{\text{err}}(H) - \widehat{\text{err}}^c(H) \leq (1+\gamma)\eta - (1-\alpha)\eta = (\alpha+\gamma)\eta,$$

proving the claim.

From Claim and Condition 2 applied with $\lambda \leq \frac{c(1-\alpha)}{(1-\gamma)}$ we obtain the following facts:

Fact 1. If $\text{err}(H) \geq \frac{(1-\alpha)\eta}{(1-\gamma)(1-\eta)}$, then H and target C are independent.

Fact 2 Each maximal independent set $\mathcal{U} \subseteq \mathcal{K}$ contains at least one hypothesis whose error is smaller than $\frac{(1-\alpha)\eta}{(1-\gamma)(1-\eta)}$. In particular, if $\mathcal{U} = \{H\}$, then $\mathrm{err}(H) \leq \frac{(1-\alpha)\eta}{(1-\gamma)(1-\eta)}$. If $\mathcal{U} = \{G, H\}$, then one among the two quantities $\mathrm{err}(G)$ and $\mathrm{err}(H)$ is smaller than or equal to $\frac{(1-\alpha)\eta}{(1-\gamma)(1-\eta)}$.

We now move on to the description of Phase 2 (see Figure 2.) Note that Phase 2 of SIH either terminates with a deterministic hypothesis, or terminates with the coin rule $\frac{1}{2}(G+H)$. The following case analysis will show that the final hypothesis output by SIH is ε-good unless it is the default hypothsis.

Let us first consider the case that the final hypothesis is the majority vote $\mathrm{MAJ}_{F,G,H}$ of three independent hypothesis F, G, and H (Step 1 in Phase 2). Then an error occurs exactly on those instances x that are wrongly predicted by at least two hypotheses of F, G, H, that is

$$\mathrm{err}\left(\mathrm{MAJ}_{F,G,H}\right) \leq \mathrm{err}(F,G) + \mathrm{err}(F,H) + \mathrm{err}(G,H).$$

By definition of independence, we know that

$$\widehat{\mathrm{err}}^{c}(X,Y) \leq \widehat{\mathrm{err}}(X,Y) \leq (\alpha+\gamma)\eta$$

for each pair (X, Y) of distinct hypothesis in $\{F, G, H\}$. Then, observing that $c\varepsilon = \eta/(1-\eta)$ and applying Condition 3 to each such pair with $\lambda \leq c(\alpha+\gamma)/(1-\gamma)$, we get that

$$\mathrm{err}\left(\mathrm{MAJ}_{F,G,H}\right) \leq \frac{3(\alpha+\gamma)\eta}{(1-\gamma)(1-\eta)} \ . \tag{16}$$

If the final hypothesis is the coin rule $\frac{1}{2}(G+H)$ (from Step 2 in Phase 2), we may apply (15) and Fact 2 to bound the error probability as follows:

$$\begin{aligned} \mathrm{err}\left(\frac{1}{2}(G+H)\right) &= \frac{1}{2}\left(\mathrm{err}(G)+\mathrm{err}(H)\right) \\ &< \frac{1}{2}\left(\frac{(1-\alpha)\eta}{(1-\gamma)(1-\eta)} + \frac{(1+\gamma)\eta}{(1-\gamma)(1-\eta)}\right) \ . \end{aligned} \tag{17}$$

If the final hypothesis is H (from Step 2), then $\mathrm{err}(H) < \frac{(1-\alpha)\eta}{(1-\gamma)(1-\eta)}$ by Fact 2. This error bound is smaller than the bound (17). We now have to find choices for the parameters α, γ, λ such that (16) and (17) are both upper bounded by ε and the previously stated conditions

$$\lambda \leq c(1+\gamma)/(1-\gamma) \ , \ \lambda \leq c(1-\alpha)/(1-\gamma) \ , \ \lambda \leq c(\alpha+\gamma)/(1-\gamma)$$

on λ hold. Equating bounds (16) and (17) and solving for α gives $\alpha = \frac{2}{7} - \frac{5}{7}\gamma$. Substituting this into (16), setting the resulting formula to ε and solving for γ yields $\gamma = \frac{(7/6)-c}{(7/6)+c}$. (Observe that the choice of $\eta = \frac{c\varepsilon}{1+c\varepsilon}$ implies the equation $\frac{\eta}{1-\eta} = c\varepsilon$.) This in turn leads to the choice $\alpha = 3\frac{c-(1/2)}{c+(7/6)}$. According to these settings one can finally choose $\lambda = 1/3$. □

Remark 4.3 *The result of Theorem 4.2 gives rise to a challenging combinatorial problem: Given a target class, find 3 independent hypotheses, or alternatively, a maximal independent set of less than 3 hypotheses. This replaces the "consistent hypothesis" paradigm of noise-free PAC learning and the "minimizing disagreement" paradigm of agnostic learning. There are examples of concept classes such that, for certain samples, one can find three or more independent hypotheses.*

Initialization
Input: A sample $(x_1, l_1), \ldots, (x_m, l_m)$, $x_i \in \{1, \ldots, d\}$, $l \in \{0, 1\}$.
For $i := 1, \ldots, d$ **Compute**

$$p_0(i) := \frac{|\{j \mid x_j = i,\ l_j = 0\}|}{m}$$

$$p_1(i) := \frac{|\{j \mid x_j = i,\ l_j = 1\}|}{m}$$

$$h(i) := \frac{(p_1(i))^2}{(p_0(i))^2 + (p_1(i))^2}$$

Prediction
Input: Some $i \in \{1, \ldots, d\}$
Output: Label 1 with probability $h(i)$ and label 0 with probability $1 - h(i)$.

Figure 3: Pseudo-code for Algorithm SQ-RULE. In the initialization phase some parameters are computed from the sample. These are then used in the working phase to make randomized predictions.

4.2 An Almost Optimal Coin Rule for the Powerset

In this subsection, we introduce and analyze a simple algorithm called Square Rule (SQ-RULE) for learning with coin rules the powerset $\mathcal{C}_d$ of d elements in presence of a malicious noise rate arbitrarily close to η_{rand} and using almost optimal sample size. (See also Figure 3.) Algorithm SQ-RULE works as follows. Let $H(p,q) = q^2/(p^2+q^2)$. For a given sample S, let $\widehat{p}_0(x)$ and $\widehat{p}_1(x)$, respectively, denote the relative frequency of $(x,0)$ and $(x,1)$. On input S, SQ-RULE outputs the coin rule[4] $F(x) = H(\widehat{p}_0(x), \widehat{p}_1(x)) = \widehat{p}_1(x)^2/(\widehat{p}_0(x)^2 + \widehat{p}_1(x)^2)$. We now show a derivation of the coin rule F. Consider a single point x and let $\widehat{p}_0 = \widehat{p}_0(x)$ and $\widehat{p}_1 = \widehat{p}_1(x)$. Now, if the true label of x is 0, we say that the "empirical return" of the adversary is $\widehat{p}_0$. Likewise, the adversary's "investment" for the false label 1 is $\widehat{p}_1$. Also, as F incorrectly classifies x with probability $H(\widehat{p}_0, \widehat{p}_1)$, the "empirical return to investment ratio", which we denote by ρ, is $H(\widehat{p}_0, \widehat{p}_1) \cdot \widehat{p}_0/\widehat{p}_1$. Similarly, if the true label of x is 1, then ρ is $(1 - H(\widehat{p}_0, \widehat{p}_1)) \cdot \widehat{p}_1/\widehat{p}_0$. The function F that minimizes ρ over all choices of $\widehat{p}_0$ and $\widehat{p}_1$ is found by letting $H(\widehat{p}_0, \widehat{p}_1) \cdot \widehat{p}_0/\widehat{p}_1 = (1 - H(\widehat{p}_0, \widehat{p}_1)) \cdot \widehat{p}_1/\widehat{p}_0$ and solving for H to obtain $H(\widehat{p}_0, \widehat{p}_1) = \widehat{p}_1^2/(\widehat{p}_0^2 + \widehat{p}_1^2)$. (A plot of the functions H and ρ is shown in Figure 4. In the same figure we also plot return to investment ratio, showing that the best strategy for the adversary is to balance the labels whence this ratio becomes 1/2.) Note that, as our final goal is to bound the quantity $\mathbf{E}_{x \sim D}|F(x) - C(x)|$, we should actually choose F so to minimize the *expected* return to investment ratio, i.e., the ratio $|H(\widehat{p}_0, \widehat{p}_1) - C(x)| \cdot D(x)/\widehat{p}_{1-C(x)}$. However, as we will show in a moment, an estimate of the unknown quantity $D(x)$ will suffice for our purposes.

Theorem 4.4 *For every $d \geq 1$ and every $0 < \varepsilon, \delta, \Delta \leq 1$, algorithm* SQ-RULE *learns the class $\mathcal{C}_d$ with accuracy ε, confidence δ, tolerating malicious noise rate $\eta = 2\varepsilon/(1+2\varepsilon) - \Delta$, and using a*

[4]The function H was also used by [10] in connection with agnostic learning.

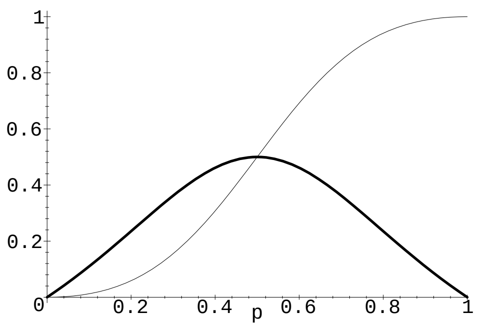

Figure 4: Curve of the coin rule $H(p,q) = q^2/(p^2+q^2)$ (thin) and of the return to investment ratio $H(p,q)\cdot p/q$ (thick); the curves are scaled to $p+q=1$.

sample of size $\widetilde{O}(d\varepsilon/\Delta^2)$.

Proof. We start the proof with the following preliminary considerations. Let $X = \{1,\ldots,d\}$, let D be any distribution on X, and let C be the target function. Let $t(x) = (1-\eta)D(x)$, i.e., $t(x)$ denotes the probability that x is presented by the adversary with the true label $C(x)$. Fix a sample S. The relative frequency of $(x, C(x))$ in S is denoted by $\widehat{t}(x)$. We assume w.l.o.g. that the adversary does never present an instance with its true label in noisy trials. (The performance of the coin rule F gets better in this case.) We denote the relative frequency of $(x, 1-C(x))$ in S by $\widehat{f}(x)$. The relative frequency of noisy trials in S is denoted by $\widehat{\eta}$. Clearly, $\widehat{\eta} = \sum_{x\in X}\widehat{f}(x)$. Applying Lemmas A.1 and A.2, it is not hard to show that there exists a sample size $m = \widetilde{O}(d\varepsilon/\Delta^2)$ such that with probability $1-\delta$ the sample S satisfies the following conditions:

$$\widehat{\eta} \;\leq\; \eta + \frac{\Delta}{2}, \tag{18}$$

$$\forall x\in X : t(x) \geq \frac{\Delta}{24d} \;\Rightarrow\; \widehat{t}(x) \geq \frac{t(x)}{2}, \tag{19}$$

$$\forall M\subseteq X : \sum_{x\in M} t(x) \leq 16\varepsilon \;\Rightarrow\; \sum_{x\in M}\widehat{t}(x) \geq \sum_{x\in M} t(x) - \frac{\Delta}{8}. \tag{20}$$

To prove (18) we apply (25) with $p' = \eta$ and $\lambda = \Delta/(2\eta)$. To prove (19) we apply (23) with $p = \Delta/(24d)$ and $\lambda = 1/2$. Finally, to prove (20) we use (23) to find that

$$\Pr\left\{S_m \leq \left(1-\frac{\lambda}{p}\right)mp\right\} \leq \exp\left(-\frac{\lambda^2 m}{2p}\right) \leq \exp\left(-\frac{\lambda^2 m}{2p'}\right)$$

where the last inequality holds for all $p' \geq p$ by monotonicity. Setting $\lambda = \Delta/8$ and $p' = 16\varepsilon$ concludes the proof of (20). These three conditions are assumed to hold in the sequel. An instance x is called *light* if $t(x) < \Delta/(24d)$, and *heavy* otherwise. Note that $\eta < \eta_{\text{rand}} \leq 2/3$ (recall that $\eta_{\text{rand}} = 2\varepsilon/(1+2\varepsilon)$.) Thus, $D(x) < \Delta/(8d)$ for all light points. The total contribution of light points to the error probability of the coin rule F is therefore less than $\Delta/8$. The following analysis focuses on heavy points; note that for these points the implication in (19) is valid. We will show that the total error probability on heavy points is bounded by $\varepsilon - \Delta/8$.

It will be instructive to consider the error probability on x of our coin rule F as the return of the adversary at x (denoted by RETURN(x) henceforth) and the quantity $\widehat{f}(x)$, defined above, as its investment at x. Our goal is to show that the total return of the adversary is smaller than

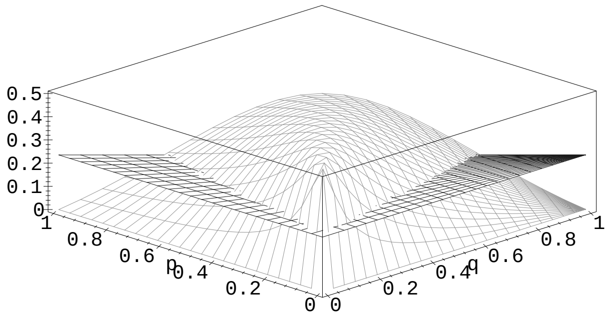

Figure 5: Plot of the return curve $R(p,q) = (pq)/(p^2+q^2)$ and of the constant function $c(p,q) = 4/17$, for $p,q \in [0,1]$, $p+q=1$.

$\varepsilon - \Delta/8$, given that its total investment is $\widehat{\eta}$. The function $R(p,q) = pq/(p^2+q^2)$ plays a central role in the analysis of the relation between return and investment. (A plot of this function is shown in Figure 4.) Function R attains its maximal value $1/2$ for $p = q$. For $q \le p/4$ or $p \le q/4$, the maximal value is $4/17$, see also Figure 5. Before bounding the total return, we will analyze the term $\textsc{return}(x)$.

If $C(x) = 0$, then

$$\widehat{f}(x) = \widehat{p}_1(x), \quad \widehat{t}(x) = \widehat{p}_0(x), \quad \textsc{return}(x) = F(x)\cdot D(x) = \frac{\widehat{p}_1(x)^2\cdot D(x)}{\widehat{p}_0(x)^2+\widehat{p}_1(x)^2}.$$

If $C(x) = 1$, then

$$\widehat{f}(x) = \widehat{p}_0(x), \quad \widehat{t}(x) = \widehat{p}_1(x), \quad \textsc{return}(x) = (1-F(x))\cdot D(x) = \frac{\widehat{p}_0(x)^2\cdot D(x)}{\widehat{p}_0(x)^2+\widehat{p}_1(x)^2}.$$

Note that in both cases $\widehat{f}(x)\cdot\widehat{t}(x) = \widehat{p}_0(x)\cdot\widehat{p}_1(x)$ and $\textsc{return}(x) = \widehat{f}(x)^2\cdot D(x)/(\widehat{p}_0(x)^2+\widehat{p}_1(x)^2)$. Setting $\widehat{\alpha}(x) = t(x) - \widehat{t}(x)$, we obtain $D(x) = t(x)/(1-\eta) = (\widehat{t}(x)+\widehat{\alpha}(x))/(1-\eta)$. For the sake of simplicity, we will use the abbreviations

$$\widehat{p}_0 = \widehat{p}_0(x), \quad \widehat{p}_1 = \widehat{p}_1(x), \quad \widehat{f} = \widehat{f}(x), \quad \widehat{t} = \widehat{t}(x), \quad \widehat{\alpha} = \widehat{\alpha}(x).$$

With these abbreviations, $\widehat{t}\cdot\widehat{f} = \widehat{p}_0\cdot\widehat{p}_1$ is valid and the term $\textsc{return}(x)$ can be written as follows:

$$\begin{aligned}\textsc{return}(x) = \frac{\widehat{f}^2\cdot(\widehat{t}+\widehat{\alpha})}{(\widehat{p}_0^2+\widehat{p}_1^2)\cdot(1-\eta)} &= \frac{1}{1-\eta}\left(\frac{\widehat{t}\cdot\widehat{f}}{\widehat{p}_0^2+\widehat{p}_1^2}\widehat{f} + \frac{\widehat{\alpha}\cdot\widehat{f}^2}{\widehat{p}_0^2+\widehat{p}_1^2}\right)\\ &= \frac{1}{1-\eta}\left(R(\widehat{p}_0,\widehat{p}_1)\cdot\widehat{f} + \frac{\widehat{\alpha}\cdot\widehat{f}^2}{\widehat{p}_0^2+\widehat{p}_1^2}\right) \qquad (21)\end{aligned}$$

We now bound separately each one of the last two terms in (21). If $\widehat{f} \ge \widehat{t}/4$, then

$$\frac{R(\widehat{p}_0,\widehat{p}_1)\cdot\widehat{f}}{(1-\eta)} \le \frac{\widehat{f}}{2(1-\eta)}.$$

Furthermore, as either $\widehat{f} = \widehat{p}_0$ or $\widehat{f} = \widehat{p}_1$,

$$\frac{\widehat{\alpha}\cdot\widehat{f}^2}{(\widehat{p}_0^2+\widehat{p}_1^2)(1-\eta)} \le \frac{\widehat{\alpha}}{1-\eta}.$$

If $\widehat{f} < \widehat{t}/4$, we bound the last term in (21) using (19) and get

$$\frac{\widehat{\alpha} \cdot \widehat{f}^2}{(\widehat{p}_0^2 + \widehat{p}_1^2)} = \frac{(t - \widehat{t}) \cdot \widehat{f}^2}{(\widehat{p}_0^2 + \widehat{p}_1^2)} \leq \frac{\widehat{t} \cdot \widehat{f}^2}{(\widehat{p}_0^2 + \widehat{p}_1^2)} = \frac{(\widehat{p}_0 \cdot \widehat{p}_1) \cdot \widehat{f}}{(\widehat{p}_0^2 + \widehat{p}_1^2)} = R(\widehat{p}_0, \widehat{p}_1) \cdot \widehat{f}.$$

Hence, using (21) and $R(\widehat{p}_0, \widehat{p}_1) \leq 4/17 < 1/4$ for $\widehat{f} < \widehat{t}/4$, we finally get

$$\text{RETURN}(x) \leq \frac{2R(\widehat{p}_0, \widehat{p}_1) \cdot \widehat{f}}{1 - \eta} < \frac{\widehat{f}}{2(1-\eta)}.$$

Piecing the above together we obtain

$$\text{RETURN}(x) \leq \frac{\widehat{f}}{2(1-\eta)} + \begin{cases} \frac{\widehat{\alpha}}{(1-\eta)} & \text{if } \widehat{f} \geq \widehat{t}/4, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

We are now in the position to bound the total return on all heavy instances x. For the first term in the left-hand-side of (22) we obtain the bound

$$\frac{1}{2(1-\eta)} \sum \widehat{f}(x) \leq \frac{\widehat{\eta}}{2(1-\eta)}$$

where the sum is over all heavy x. The treatment of the second term in the left-hand-side of (22) is more subtle. Let M denote the set of heavy instances x where $\widehat{f}(x) \geq \widehat{t}(x)/4$. $D(M)$ is therefore bounded as follows:

$$\sum_{x \in M} t(x) \leq 2 \sum_{x \in M} \widehat{t}(x) \leq 8 \sum_{x \in M} \widehat{f}(x) \leq 8\widehat{\eta} < 16\varepsilon.$$

From (20), we conclude that:

$$\frac{1}{1-\eta} \sum_{x \in M} \widehat{\alpha}(x) \leq \frac{1}{1-\eta} \sum_{x \in M} (t(x) - \widehat{t}(x)) \leq \frac{\Delta}{8(1-\eta)}.$$

A straightforward computation shows that

$$\frac{\widehat{\eta}}{2(1-\eta)} + \frac{\Delta}{8(1-\eta)} \leq \varepsilon - \Delta/8.$$

As the probability of all light points is at most $\Delta/8$, the expected error of SQ-RULE is at most ε. This completes the proof of Theorem 4.4 □

The upper bound of Theorem 4.4 has a matching lower bound (up to logarithmic factors). The proof, which is a somewhat involved modification of the proof of Theorem 3.9, is only sketched.

Theorem 4.5 *For every target class $\mathcal{C}$ with VC dimension $d \geq 3$, for every $0 < \varepsilon \leq 1/38$, $0 < \delta \leq 1/74$, and for every $0 < \Delta = o(\varepsilon)$, the sample size needed by any strategy (even using randomized hypotheses) for learning $\mathcal{C}$ with accuracy ε, confidence δ, and tolerating malicious noise rate $\eta = 2\varepsilon/(1+2\varepsilon) - \Delta$, is $\Omega\left(d\varepsilon/\Delta^2\right)$.*

Proof. One uses d shattered points with a suitable distribution D and shows that, with constant probability, there exists a constant fraction of these points which occur with a frequency much lower than expected. The measure according to D of these points is 2ε, but the adversary can balance the occurrences of both labels on these points while using noise rate less than η_{rand}. Hence, even a randomized hypothesis cannot achieve an error smaller than ε. □

Remark 4.6 *Algorithm* SQ-RULE *can be modified to learn the class I_k of unions of at most k intervals. Similarly to Corollary 3.15, one first computes a suitable partition of the domain into bins and applies the algorithm for the powerset afterwards as a subroutine.*

Noise model	classification		malicious with deterministic hyp.		malicious with randomized hyp.	
Theoretical limit on noise rate	$\eta_0 := 1/2$		$\eta_0 := \varepsilon/(1+\varepsilon)$		$\eta_0 := 2\varepsilon/(1+2\varepsilon)$	
Type of bound	upper	lower	upper	lower	upper	lower
$\eta = \eta_0 - \Delta$, $\Delta \to 0$	$\frac{d}{\varepsilon\Delta^2}$	$\frac{d}{\varepsilon\Delta^2}$	$\frac{d}{\varepsilon} + \frac{\varepsilon}{\Delta^2}$	$\frac{d}{\varepsilon} + \frac{\varepsilon}{\Delta^2}$	$\frac{d\varepsilon}{\Delta^2}$	$\frac{d\varepsilon}{\Delta^2}$
$\eta = c \cdot \eta_0$	$\frac{d}{\varepsilon}$	$\frac{d}{\varepsilon}$	$\frac{d}{\varepsilon}$	$\frac{d}{\varepsilon}$	—	—
$\eta < \frac{(7/6)\varepsilon}{1+(7/6)\varepsilon}$	—	—	—	—	$\frac{d}{\varepsilon}$	$\frac{d}{\varepsilon}$
Min. Dis. $\eta = c \cdot \eta_0$	$\frac{d}{\varepsilon}$	$\frac{d}{\varepsilon}$	$\frac{d}{\varepsilon}$	$\frac{d}{\varepsilon}$	—	—
Min. Dis. $\eta = \eta_0 - \Delta$, $\Delta \to 0$	$\frac{d}{\varepsilon\Delta^2}$	$\frac{d}{\varepsilon\Delta^2}$	$\frac{d\varepsilon}{\Delta^2}$	$\frac{d\varepsilon}{\Delta^2}$	—	—

Table 1: Survey on sample size results of learning in the presence of noise, ignoring logarithmic terms. Empty entries are unknown or are combinations which make no sense. The lower part contains the results for the minimum disagreement strategy. For the sake of comparison we also add the corresponding results for classification noise.

5 Summary

Table 1 shows the known results on learning in the malicious and classification noise models. The latter is a noise model where independently for every example the label is inverted with probability $\eta < 1/2$.[5] There are still a few problems open. One is the question whether the strong adversary in the lower bound proofs of Theorems 3.9 and 4.5 can be replaced by the weaker KL-adversary. Also it would be interesting to see whether the constant 7/6 in Theorem 4.2 can improved to arbitrary constants $0 \le c < 2$. It seem that both questions are not easy to answer.

Acknowledgements

Preliminary versions of this paper were presented at the 28th Annual Symposium on Theory of Computing, 1996, and at the 3rd European Conference on Computational Learning Theory, 1997. This research was supported by a Grant from the G.I.F., the German-Israeli Foundation for Scientific Research and Development. The authors gratefully acknowledge the support of ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II). The first author gratefully acknowledges the support of Progetto Vigoni. The third and fifth author gratefully acknowledge the support of Bundesministerium für Forschung und Technologie grant 01IN102C/2, the support of: Deutsche Forschungsgemeinschaft grant Si 498/3-1, Deutscher Akademischer Austauschdienst grant 322-vigoni-dr, Deutsche Forschungsgemeinschaft research grant We 1066/6-2.

[5]See e.g. [11] for a survey on this noise model and for the upper bound on the sample size in the case of finite concept classes. See [15] for the lower bound on the sample size, and [16] for a generalization of Laird's upper bound to arbitrary concept classes.

The authors take the responsibility for the contents. Thanks to an anonymous reviewer for pointing out [12].

References

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

[2] R.R. Bahadur and R. Ranga-Rao. On deviations of the sample mean. *Annals of Mathematical Statistics*, 31:1015–1027, 1960.

[3] 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.

[4] Yuan Shih Chow and Henry Teicher. *Probability Theory*. Springer-Verlag, 1988.

[5] 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.

[6] Claudio Gentile and David Helmbold. Improved lower bounds for learning from noisy examples: an information-theoretic approach. Submitted for journal publication. A preliminary version appeared in the Proceedings of the 11th Annual Conference on Computational Learning Theory, ACM Press, 1998, 1998.

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

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

[9] Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and Systems Sciences*, 48(3):464–497, 1994. An extended abstract appeared in the *Proceedings of the 30th Annual Symposium on the Foundations of Computer Science*.

[10] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2/3):115–142, 1994.

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

[12] J.E. Littlewood. On the probability in the tail of a binomial distribution. *Advances in Applied Probability*, 1:43–72, 1969.

[13] N. Sauer. On the Density of Families of Sets. *J. Combin. Th. A*, 13:145–147, 1972.

[14] S. Shelah. A Combinatorial Problem; Stability and Order for Models and Theories in Infinitary Languages. *Pacific J. of Math.*, 41:247–261, 1972.

[15] Hans Ulrich Simon. General bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 52(2):239–255, 1996.

[16] Hans Ulrich Simon. A general upper bound on the number of examples sufficient to learn in the presence of classification noise, 1996. Unpublished Manuscript.

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

[18] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, New York, 1982.

A Some Statistical and Combinatorial Relations

Let $S_{m,p}$ and $S'_{m,p'}$ be the sums of successes in a sequence of m Bernouilli trials each succeeding with probability respectively at least p and at most p'.

Lemma A.1 *For all* $0 < \lambda < 1$*,*

$$\begin{aligned} \Pr\{S_{m,p} \leq (1-\lambda)mp\} &\leq e^{-\lambda^2 mp/2} && (23) \\ \Pr\{S_{m,p} \leq m(p-\lambda)\} &\leq e^{-2\lambda^2 m}, && (24) \\ \Pr\{S'_{m,p'} \geq (1+\lambda)mp'\} &\leq e^{-\lambda^2 mp'/3}. && (25) \end{aligned}$$

Let $\mathcal{C}$ be a target class of VC dimension d over some domain X. Let D be a distribution over X. Let S be an unlabeled sample of size m drawn from X under D. For $C \in \mathcal{C}$ let $D_S(C) = |\{x \in S \,:\, x \in C\}|/m$, the empirical probability of C.

Lemma A.2 ([18, 3]) *For every* $0 < \varepsilon, \gamma \leq 1$ *and every* $0 < \delta < 1$*, the probability that there exists a* $C \in \mathcal{C}$ *such that* $D(C) > \varepsilon$ *and* $D_S(C) \leq (1-\gamma)D(C)$ *is at most* $8(2m)^d e^{-\gamma^2 \varepsilon m/4}$*, which in turn is at most* δ *if*

$$m \geq \max\left\{\frac{8}{\gamma^2\varepsilon}\ln\left(\frac{8}{\delta}\right), \frac{16d}{\gamma^2\varepsilon}\ln\left(\frac{16}{\gamma^2\varepsilon}\right)\right\} .$$

Lemma A.3 ([13, 14]) *Let* $\mathcal{C}$ *be a target class over* X *of VC dimension* d*. For all* $(x_1, \ldots, x_m) \in X^m$

$$|\{(C(x_1), \ldots, C(x_m)) \,:\, C \in \mathcal{C}\}| \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Proof of Fact 3.2. 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

$$\sqrt{2\pi N}\left(\frac{N}{e}\right)^N < N! < \sqrt{2\pi N}\left(\frac{N}{e}\right)^N e^{\frac{1}{12N}} . \tag{26}$$

Using (26) one can lower bound the binomial coefficient $\binom{N}{Np}$ as follows (assuming that N is a multiple of $1/p$, which will be justified later in the proof)

$$\binom{N}{Np} = \frac{N!}{(Np)!(Nq)!}$$

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

This leads to

$$
\sqrt{Npq} > \frac{1}{\sqrt{2\pi}p^{Np}\ q^{Nq}}\binom{N}{Np}^{-1} e^{-\frac{1}{12Npq}}. \tag{27}
$$

[1] proved the following lower bound on the tail of the binomial distribution, where $0 \leq k \leq N$,

$$
\Pr\{S_{N,p} \geq k\} \geq \binom{N}{k} p^k q^{(N-k)} \cdot \frac{q(k+1)}{k+1-p(N+1)} \cdot \left(1 + \frac{Npq}{(k-Np)^2}\right)^{-1}. \tag{28}
$$

In order to be able to apply (27) we first remove the "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.

$$
\begin{aligned}
pq = (p'+\gamma)(q'-\gamma) &= p'q' + \gamma(q'-p') - \gamma^2 \\
&= p'q' + \gamma(q'-p'-\gamma) \\
&= p'q' + \gamma(q-p') \\
&< p'q' + \gamma < p'q' + \frac{1}{N}.
\end{aligned} \tag{29}
$$

Then (29) and $N \geq 37/(pq)$ imply that

$$
Np'q' > Npq - 1 \geq 36. \tag{30}
$$

Hence (1) can be lower bounded as follows

$$
\begin{aligned}
&\Pr\left\{S_{N,p} \geq \lfloor Np \rfloor + \left\lfloor\sqrt{Npq-1}\right\rfloor\right\} \\
&\quad \geq \Pr\left\{S_{N,p'} \geq Np' + \left\lfloor\sqrt{Npq-1}\right\rfloor\right\} \\
&\quad \geq \Pr\left\{S_{N,p'} \geq Np' + \left\lfloor\sqrt{N(p'q'+\frac{1}{N})-1}\right\rfloor\right\} \\
&\quad \geq \Pr\left\{S_{N,p'} \geq Np' + \left\lfloor\sqrt{Np'q'}\right\rfloor\right\}.
\end{aligned} \tag{31}
$$

In order to bound (31) we apply inequality (28) 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 (28), denoted by F_1, F_2 and F_3, are separately bounded as follows.

$$
F_1 = \binom{N}{Np' + \lfloor\sqrt{Np'q'}\rfloor} p'^{Np'+\lfloor\sqrt{Np'q'}\rfloor} q'^{N-Np'-\lfloor\sqrt{Np'q'}\rfloor} \tag{32}
$$

$$
= \binom{N}{Np' + \lfloor\sqrt{Np'q'}\rfloor} p'^{Np'} q'^{Nq'} \left(\frac{p'}{q'}\right)^{\lfloor\sqrt{Np'q'}\rfloor} \tag{33}
$$

$$
F_2 = \frac{q'(Np' + \lfloor\sqrt{Np'q'}\rfloor + 1)}{Np' + \lfloor\sqrt{Np'q'}\rfloor + 1 - Np' - p'} \tag{34}
$$

$$
\begin{aligned}
&= \frac{Np'q' + q'(\lfloor\sqrt{Np'q'}\rfloor + 1)}{\lfloor\sqrt{Np'q'}\rfloor + q'} > \sqrt{Np'q'} \\
&> \frac{1}{\sqrt{2\pi}\,{p'}^{Np'}\,{q'}^{Nq'}}\binom{N}{Np'}^{-1} e^{-\frac{1}{12Np'q'}}. \qquad (35)
\end{aligned}
$$

(The inequality (35) follows from (27).)

$$
\begin{aligned}
F_3 &= \left(1 + \frac{Np'q'}{(Np' + \lfloor\sqrt{Np'q'}\rfloor - Np')^2}\right)^{-1} \\
&= \left(1 + \frac{Np'q'}{\lfloor\sqrt{Np'q'}\rfloor^2}\right)^{-1} \\
&= \frac{\lfloor\sqrt{Np'q'}\rfloor^2}{\lfloor\sqrt{Np'q'}\rfloor^2 + Np'q'} \\
&> \frac{(\sqrt{Np'q'} - 1)^2}{(\sqrt{Np'q'} + 1)^2 + Np'q'} \\
&= \frac{Np'q' - 2\sqrt{Np'q'} + 1}{2Np'q' - 2\sqrt{Np'q'} + 1} \\
&> \frac{Np'q' - 2\sqrt{Np'q'}}{2Np'q' - 2\sqrt{Np'q'}} \\
&= \frac{1}{2}\left(1 - \frac{\sqrt{Np'q'}}{Np'q' - \sqrt{Np'q'}}\right) \\
&= \frac{1}{2}\left(1 - \frac{1}{\sqrt{Np'q'} - 1}\right). \qquad (36)
\end{aligned}
$$

The following calculation shows how the product of (33) and (35) can be lower bounded. For notational convenience let $T = \left(e^{\frac{1}{12Np'q'}}\sqrt{2\pi}\right)^{-1}$ and let $K = \lfloor\sqrt{Np'q'}\rfloor$.

$$
\begin{aligned}
F_1 \cdot F_2 &> \frac{\binom{N}{Np'+K}\left(\frac{p'}{q'}\right)^K}{\sqrt{2\pi}\binom{N}{Np'} e^{\frac{1}{12Np'q'}}} \\
&= T\left(\frac{p'}{q'}\right)^K \cdot \frac{N!(Np')!(N - Np')!}{N!(Np' + K)!(N - Np' - K)!} \\
&= T\left(\frac{p'}{q'}\right)^K \frac{(Nq' - K + 1)\cdots(Nq')}{(Np' + 1)\cdots(Np + K)} \\
&= T\left(\frac{p'}{q'}\right)^K \left(\prod_{i=1}^{K} \frac{Nq' - K + i}{Np' + i}\right) \qquad (37) \\
&> T\left(\frac{p'}{q'}\right)^K \left(\frac{Nq'}{Np' + K}\right)^K \qquad (38) \\
&= T\left(\frac{Np'}{Np' + K}\right)^K \\
&= T\left(\frac{Np' + K}{Np'}\right)^{-K}
\end{aligned}
$$

$$
\begin{aligned}
&= T\left(1+\frac{K}{Np'}\right)^{-K} \\
&\geq T\left(1+\frac{\sqrt{Np'q'}}{Np'}\right)^{-\sqrt{Np'q'}} \\
&= T\left(1+\frac{q'}{\sqrt{Np'q'}}\right)^{-\sqrt{Np'q'}} && (39)\\
&\geq Te^{-q'} && (40)\\
&= \frac{1}{\sqrt{2\pi}\ e^{\frac{1}{12Np'q'}}}e^{-q'} && (41)\\
&\geq \frac{1}{\sqrt{2\pi}\ e^{\frac{1}{12\cdot 36}}}e^{-1} \geq 0.14642\ldots && (42)
\end{aligned}
$$

In (41) and (42) we used that $Np'q' > 36$, by (30). For the step from (37) to (38) we assume that $Nq' - k \geq Np'$. If $Nq' - k < Np'$ the steps from (38) in the above calculation are replaced by the following:

$$
\begin{aligned}
& T\ \left(\frac{p'}{q'}\right)^{K}\ \left(\prod_{i=1}^{K}\frac{Nq'-K+i}{Np'+i}\right) && (43)\\
&> T\ \left(\frac{p'}{q'}\right)^{K}\ \left(\frac{Nq'-K}{Np'}\right)^{K} && (44)\\
&= T\left(\frac{Nq'-K}{Nq'}\right)^{K} \\
&= T\left(\frac{Nq'-\lfloor\sqrt{Np'q'}\rfloor}{Nq'}\right)^{\lfloor\sqrt{Np'q'}\rfloor} \\
&\geq T\left(\frac{Nq'-\sqrt{Np'q'}}{Nq'}\right)^{\sqrt{Np'q'}} \\
&= T\left(1-\frac{p'}{\sqrt{Np'q'}}\right)^{\sqrt{Np'q'}} && (45)\\
&\geq \frac{1}{\sqrt{2\pi}\ e^{\frac{1}{12Np'q'}}}\ \frac{10}{11}e^{-p'} && (46)\\
&\geq \frac{1}{\sqrt{2\pi}\ e^{\frac{1}{12\cdot 36}}}\ \frac{10}{11}e^{-1} \geq 0.133112\ldots && (47)
\end{aligned}
$$

The step from (45) to (46) follows from an elementary analysis of the function $(1-\frac{a}{b})^b - \frac{10}{11}e^{-a}$. Using (47) (which is less than the bound in (42)) and (36) we can lower bound the product $F_1F_2F_3$ as follows:

$$
\begin{aligned}
F_1 \cdot F_2 \cdot F_3 &> 0.133112 \cdot \frac{1}{2} \cdot \left(1-\frac{1}{\lfloor\sqrt{Np'q'}\rfloor - 1}\right) \\
&\geq 0.066556 \cdot \left(1-\frac{1}{\left\lfloor\sqrt{36}\right\rfloor - 1}\right) && (48)\\
&> 0.05324\ldots > \frac{1}{19}\ . && (49)
\end{aligned}
$$

For the step from (48) to (49) we again used inequality (30). $\square$