Learning Constant-Depth Circuits in Malicious Noise Models

Adam Klivans KLIVANS@UTEXAS.EDU
University of Texas at Austin

Konstantinos Stavropoulos KSTAVROP@UTEXAS.EDU
University of Texas at Austin

Arsen Vasilyan ARSENVASILYAN@GMAIL.COM
University of Texas at Austin

Editors: Nika Haghtalab and Ankur Moitra

Abstract

The seminal work of Linial, Mansour, and Nisan gave a quasipolynomial-time algorithm for learning constant-depth circuits (AC^0) with respect to the uniform distribution on the hypercube. Extending their algorithm to the setting of malicious noise, where both covariates and labels can be adversarially corrupted, has remained open. Here we achieve such a result, inspired by recent work on learning with distribution shift. Our running time essentially matches their algorithm, which is known to be optimal assuming various cryptographic primitives.

Our proof uses a simple outlier-removal method combined with Braverman's theorem for fooling constant-depth circuits. We attain the best possible dependence on the noise rate and succeed in the harshest possible noise model (i.e., contamination or so-called "nasty noise").

Keywords: PAC learning, malicious noise, constant-depth circuits, contamination

1. Introduction

In their famous paper, Linial, Mansour, and Nisan (Linial et al., 1993) introduced the "low-degree" algorithm for learning Boolean functions with respect to the uniform distribution on $\{\pm 1\}^d$. The running time and sample complexity of their algorithm scales in terms of the Fourier concentration of the underlying concept class, and, using this framework, they obtained a quasipolynomial-time algorithm for learning constant-depth, polynomial-size circuits (AC^0).

Prior work (Kalai et al., 2008) had extended their result to the agnostic setting, where the *labels* can be adversarially corrupted, but the marginal distribution on inputs must still be uniform over $\{\pm 1\}^d$. Remarkably, there had been no progress on this problem in the last three decades for *malicious* noise models where *both* covariates and labels can be adversarially corrupted (Valiant, 1985; Kearns and Li, 1993). In fact, in the malicious noise setting, nothing was known even for the special case of arbitrary polynomial-size DNF formulas (i.e., disjunctions of conjuntions).

In this paper, we completely resolve this problem and obtain a quasipolynomial-time algorithm for learning AC^0 in the harshest possible noise model, the so-called "nasty noise" model of (Bshouty et al., 2002). We define this model below and refer to it simply as learning with contamination, in line with recent work in robust statistics (see, e.g., (Diakonikolas and Kane, 2023)).

Definition 1 (Learning from Contaminated Samples) *A set of N labeled examples $\bar{S}_{\mathrm{inp}}$ is an η-contaminated (uniform) sample with respect to some class $\mathcal{C} \subseteq \{\{\pm 1\}^d \to \{\pm 1\}\}$, where $N \in \mathbb{N}$ and $\eta \in (0, 1)$, if it is formed by an adversary as follows.*

1. *The adversary receives a set of N clean i.i.d. labeled examples $\bar{S}_{\mathrm{cln}}$, drawn from the uniform distribution over $\{\pm 1\}^d$ and labeled by some unknown concept f^* in $\mathcal{C}$.*

2. *The adversary removes an arbitrary set $\bar{S}_{\mathrm{rem}}$ of $\lfloor \eta N \rfloor$ labeled examples from $\bar{S}_{\mathrm{cln}}$ and substitutes it with an adversarial set of $\lfloor \eta N \rfloor$ labeled examples $\bar{S}_{\mathrm{adv}}$.*

Namely, $\bar{S}_{\mathrm{inp}} = (\bar{S}_{\mathrm{cln}} \setminus \bar{S}_{\mathrm{rem}}) \cup \bar{S}_{\mathrm{adv}}$. For the corresponding unlabeled set S_{inp}, we say that it is an η-contaminated (uniform) sample.

In this model, the goal of the learner is to output (with probability $1-\delta$) a hypothesis $h : \{\pm 1\}^d \to \{\pm 1\}$ such that $\mathbb{P}_{\mathbf{x}\sim \mathrm{Unif}(\{\pm 1\}^d)}[h(\mathbf{x}) \neq f^*(\mathbf{x})] \leq 2\eta + \epsilon$. The factor 2 is known to be the best possible constant achievable by any algorithm (Bshouty et al., 2002).

Although there is now a long line of research giving computationally efficient algorithms for learning Boolean function classes in malicious noise models, these algorithms primarily apply to geometric concept classes and continuous marginal distributions, such as halfspaces or intersections of halfspaces with respect to Gaussian or log-concave densities and provide suboptimal error bounds (Kalai et al., 2008; Klivans et al., 2009; Awasthi et al., 2017; Diakonikolas et al., 2018; Shen and Zhang, 2021). In particular, nothing was known for the case of AC^0.

Our main theorem is as follows:

Theorem 2 *For any $s, \ell, d \in \mathbb{N}$, and $\epsilon, \delta \in (0,1)$, there is an algorithm that learns the class of AC^0 circuits of size s and depth ℓ and achieves error $2\eta + \epsilon$, with running time and sample complexity $d^{O(k)} \log(1/\delta)$, where $k = (\log(s))^{O(\ell)} \log(1/\epsilon)$, from contaminated samples of any noise rate η.*

Our running time essentially matches the Linial, Mansour, and Nisan result, which is known to be optimal assuming various cryptographic primitives (Kharitonov, 1993).

More generally, we prove that any concept class $\mathcal{C}$ that admits ℓ_1-*sandwiching polynomials* of degree k can be learned in time $d^{O(k)}$ from contaminated samples. Recent work due to (Goel et al., 2024) had obtained a similar result achieving the weaker bound of $O(\eta) + \epsilon$ for learning functions with ℓ_2-sandwiching polynomials. Crucially, it remains unclear how to obtain such ℓ_2 sandwiching approximators for constant depth circuits [1], and so their result does not apply here.

In 2005, Kalai et al. (Kalai et al., 2008) showed that ℓ_1-approximation suffices for agnostic learning. Here we complete the analogy for malicious learning, showing that ℓ_1-sandwiching implies learnability with respect to contamination.

Proof Overview. The input set $\bar{S}_{\mathrm{inp}}$ is η-contaminated. This might make it hard to find a hypothesis with near-optimal error on $\bar{S}_{\mathrm{inp}}$. However, we are only interested in finding a hypothesis with error $2\eta + \epsilon$ on the clean distribution, which is structured (in particular, the marginal distribution on the features is uniform over $\{\pm 1\}^d$). In order to take advantage of the structure of the clean distribution despite only having access to the contaminated sample, we make use of the notion of sandwiching polynomials:

Definition 3 (Sandwiching polynomials) *Let $f : \{\pm 1\}^d \to \{\pm 1\}$. We say that the ($\ell_1$) ϵ-sandwiching degree of f with respect to the uniform distribution over the hypercube $\{\pm 1\}^d$ is k if there are polynomials $p_{\mathrm{up}}, p_{\mathrm{down}} : \{\pm 1\}^d \to \mathbb{R}$ of degree at most k such that (1) $p_{\mathrm{down}}(\mathbf{x}) \leq f(\mathbf{x}) \leq p_{\mathrm{up}}(\mathbf{x})$ for all $\mathbf{x} \in \{\pm 1\}^d$ and (2) $\mathbb{E}_{\mathbf{x}\sim \mathrm{Unif}(\{\pm 1\}^d)}[p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})] \leq \epsilon$.*

1. Braverman's celebrated result on AC^0 (Braverman, 2008) obtains only ℓ_1-sandwiching.

The sandwiching degree of size-s depth-ℓ AC^0 circuits is bounded by $k = (\log(s))^{O(\ell)} \log(1/\epsilon)$, due to the result of Braverman on fooling constant-depth circuits (see Theorem 6 from (Braverman, 2008; Tal, 2017; Harsha and Srinivasan, 2019)). Suppose that $\bar{S}$ is a subset of $\bar{S}_{\mathrm{inp}}$ that preserves the expectations of low-degree and non-negative polynomials (e.g., $p_{\mathrm{up}} - p_{\mathrm{down}}$) compared to the uniform distribution. Under this condition, low-degree polynomial regression gives a hypothesis with near-optimal error on $\bar{S}$ (see Section 4).

We show in Lemma 4 that a simple procedure that iteratively removes samples from $\bar{S}_{\mathrm{inp}}$ can be used to form such a set $\bar{S}$ (that preserves the expectations of non-negative, degree-k and low-expectation polynomials) and, moreover, this procedure removes more contaminated points than clean points. The last property is important, because it implies that $\bar{S}$ is representative for the ground truth distribution, i.e., any near-optimal hypothesis for $\bar{S}$ will also have error $2\eta + \epsilon$ on the ground truth.

This is possible because the only way the adversary can significantly increase the expectation of a non-negative polynomial p is by inserting examples $\mathbf{x}$ where $p(\mathbf{x})$ is unreasonably large compared to the typical values of p over the uniform distribution. Our algorithm iteratively finds the non-negative polynomial q with the largest expectation over a given set through a simple linear program and then removes the points $\mathbf{x}$ for which $q(x)$ is large.

Our iterative outlier removal procedure is inspired by prior work on TDS learning (Testable Learning with Distribution Shift) and PQ learning (Klivans et al., 2024; Goel et al., 2024) as well as the work of (Diakonikolas et al., 2018) on learning geometric concepts from contaminated examples. Both of these works use outlier removal procedures that give bounds on the variance of polynomials rather than the expectation of non-negative polynomials and, instead of linear programming, they use spectral algorithms.

2. Notation

Throughout this work, when we refer to a set S of examples from the hypercube $\{\pm 1\}^d$, we consider every example in S to be a unique and separate instance of the corresponding element in $\{\pm 1\}^d$. Moreover, we denote with $\bar{S}$ the corresponding labeled set of examples in $\{\pm 1\}^d \times \{\pm 1\}$. We denote with $\mathrm{Unif}(\{\pm 1\}^d)$ or simply Unif_d the uniform distribution over the hypercube.

Recall that polynomials over $\{\pm 1\}^d$ are functions of the form $p(\mathbf{x}) = \sum_{\mathcal{I} \subseteq [d]} c_p(\mathcal{I}) \prod_{i \in \mathcal{I}} x_i$, where $\mathbf{x} = (x_i)_{i \in [d]}$ and $c_p(\mathcal{I}) \in \mathbb{R}$. We denote with $\mathbf{x}^{\mathcal{I}}$ the quantity $\prod_{i \in \mathcal{I}} x_i$. We say that the degree of p is at most k if for any $\mathcal{I} \subseteq [d]$ with $|\mathcal{I}| > k$, we have $c_p(\mathcal{I}) = 0$. For a polynomial p, we denote with $\|p\|_{\mathrm{coef}}$ the ℓ_1 norm of its coefficients, i.e., $\|p\|_{\mathrm{coef}} = \sum_{\mathcal{I} \subseteq [d]} |c_p(\mathcal{I})|$.

3. Removing the Outliers

The input set $\bar{S}_{\mathrm{inp}}$ includes an η fraction of contaminated examples. It is, of course, impossible to identify the exact subset of $\bar{S}_{\mathrm{inp}}$ that is contaminated. However, we show how to remove contaminated examples that lead to inflation of the expectations of low-degree non-negative polynomials, which we call "outliers." We remove only a relatively small number of clean examples from $\bar{S}_{\mathrm{inp}}$, as we show in the following lemma.

Lemma 4 (Outlier removal) *Let S_{inp} be an η-contaminated uniform sample (see Definition 1) with size N. For any choice of the parameters $\epsilon, \delta \in (0, 1)$, and $k \in \mathbb{N}$, the output S_{filt} of Algo-*

rithm 1 satisfies the following, whenever $N \geq C\frac{(3d)^{2k}}{\epsilon^2}\log(1/\delta)$*, for some sufficiently large constant* $C \geq 1$.

1. *With probability at least* $1-\delta$*, the number of clean examples in* S_{inp} *that are removed from* S_{filt} *is at most equal to the number of adversarial examples that are removed from* S_{filt} *(see Figure 1). Namely,* $|(S_{\text{inp}} \cap S_{\text{cln}}) \setminus S_{\text{filt}}| \leq |S_{\text{adv}} \setminus S_{\text{filt}}|$.

2. *For any non-negative polynomial* p *over* $\{\pm 1\}^d$ *with degree at most* k *and* $\mathbb{E}_{\mathbf{x}\sim \text{Unif}_d}[p(\mathbf{x})] \leq \frac{\epsilon}{8}$*, we have* $\sum_{\mathbf{x}\in S_{\text{filt}}} p(\mathbf{x}) \leq \epsilon N$ *with probability at least* $1-\delta$.

Algorithm 1: Outlier removal through Linear Programming

Input: Set $S_{\text{inp}} \subseteq \{\pm 1\}^d$ of size N and parameters $\epsilon \in (0,1)$, $B > 0$ and $k \in \mathbb{N}$
Output: Filtered set $S_{\text{filt}} \subseteq S_{\text{inp}}$.
Let $B = 3^k d^{k/2}$, $\Delta = \frac{\epsilon}{2B}$.
$S^{(0)} \leftarrow S_{\text{inp}}$ and let S_{ref} consist of N i.i.d. examples from $\text{Unif}(\{\pm 1\}^d)$
for $i = 0, 1, 2, \ldots, N$ **do**
Let p^* be the solution of the following linear program (P) and $\lambda^* = \frac{1}{N}\sum_{\mathbf{x}\in S^{(i)}} p^*(\mathbf{x})$.

$$\left\{\begin{array}{ll} \max_p & \sum_{\mathbf{x}\in S^{(i)}} p(\mathbf{x}) \\ \text{s.t.:} & p \text{ polynomial, } \deg(p) \leq k \text{ and } \|p\|_{\text{coef}} \leq B \\ & p(\mathbf{x}) \geq 0, \text{ for all } \mathbf{x} \in S_{\text{ref}} \cup S_{\text{inp}} \\ & \frac{1}{N}\sum_{\mathbf{x}\in S_{\text{ref}}} p(\mathbf{x}) \leq \epsilon/4 \end{array}\right\} \tag{P}$$

if $\lambda^* \leq \epsilon$ **then** output $S_{\text{filt}} \leftarrow S^{(i)}$ and terminate;
else
let $\tau^* \geq 0$ be the smallest value such that
$\frac{|S^{(i)}|}{N}\mathbb{P}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau^*] \geq 2\,\mathbb{P}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x}) > \tau^*] + \Delta$
$S^{(i+1)} \leftarrow S^{(i)} \setminus \{\mathbf{x} \in S^{(i)} : p^*(\mathbf{x}) > \tau^*\}$
end
end

Our Algorithm 1 is similar in spirit to outlier removal procedures that have been used previously in the context of learning with contaminated samples (Diakonikolas et al., 2018) and tolerant learning with distribution shift (Goel et al., 2024): we iteratively find the non-negative polynomial with largest expectation and remove the examples that give this polynomial unusually large values. Here we focus on the expectations of non-negative polynomials, while in all previous works, the guarantees after outlier removal concerned the variance of arbitrary polynomials. In this sense, our guarantees are stronger, but only hold for non-negative polynomials. Our algorithm solves, in every iteration, one linear program (P) in place of the usual spectral techniques from prior work.

The proof idea is that whenever there is a non-negative polynomial p^* with unreasonably large expectation, there have to be many outliers that give unusually large values to p^*. By removing all the points where p^* is large, we can, therefore, be confident that we remove more outliers than clean examples (part 1 of Lemma 4). When the algorithm terminates, all non-negative polynomials with

low expectation under the uniform distribution, will also have low expectation under the remaining set of examples (part 2 of Lemma 4).

For part 1, we analyze the non-terminating iterations and we show that for each clean point that is filtered out by the procedure, at least one adversarial point is filtered out as well. We first show that in such an iteration, there is a $\tau \in [0, B]$ such that $\frac{|S^{(i)}|}{N} \mathbb{P}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau] \geq 2\,\mathbb{P}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x}) > \tau] + \Delta > 0$. This implies that in every non-terminating iteration at least one point is removed and, therefore, some iteration $i \leq N$ will satisfy the stopping criterion and terminate (there are only N points in total).

Claim *In any non-terminating iteration (i.e. an iteration where $\lambda^* > \epsilon$), there is $\tau^* \in [0, B]$ such that*

$$\frac{|S^{(i)}|}{N} \mathop{\mathbb{P}}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau^*] \geq 2 \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x}) > \tau^*] + \Delta.$$

Proof Suppose, for contradiction, that for all $\tau \in [0, B]$ we have

$$\frac{|S^{(i)}|}{N} \mathop{\mathbb{P}}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau] < 2 \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x}) > \tau] + \Delta$$

We may integrate over $\tau \in [0, B]$ both sides of the above inequality, since the corresponding functions of τ have finite number of discontinuities (at most equal to $|S^{(i)}| + |S_{\text{ref}}|$).

$$\frac{|S^{(i)}|}{N} \int_{\tau=0}^{B} \mathop{\mathbb{P}}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau]\, d\tau < 2 \int_{\tau=0}^{B} \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x}) > \tau]\, d\tau + \Delta B \tag{1}$$

We will now substitute the integrals above with expectations, i.e., $\int_{\tau=0}^{B} \mathbb{P}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau]\, d\tau = \mathbb{E}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x})]$ and $\int_{\tau=0}^{B} \mathbb{P}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x}) > \tau]\, d\tau = \mathbb{E}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x})]$. We use the simple fact that for any non-negative random variable X with values in $[0, B]$, we have $\mathbb{E}[X] = \int_{\tau=0}^{B} \mathbb{P}[X > \tau]\, d\tau$.

We first set $X = p^*(\mathbf{x})$, where $\mathbf{x} \sim S^{(i)}$ and observe that (1) $p^*(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in S^{(i)}$, and also that (2) $p^*(\mathbf{x}) \leq \|p\|_{\text{coef}} \leq B$ for all $\mathbf{x} \in \{\pm 1\}^d \supseteq S^{(i)}$, since p^* satisfies $\|p\|_{\text{coef}} \leq B$ according to the constraints of (P) and $p^*(\mathbf{x}) = \sum_{\mathcal{I}\subseteq[d]} c_{p^*}(\mathcal{I})\mathbf{x}^{\mathcal{I}} \leq \sum_{\mathcal{I}\subseteq[d]} |c_{p^*}(\mathcal{I})| \cdot |\mathbf{x}^{\mathcal{I}}| = \sum_{\mathcal{I}\subseteq[d]} |c_{p^*}(\mathcal{I})| = \|p^*\|_{\text{coef}}$, since $\mathbf{x} \in \{\pm 1\}^d$ and therefore $|\mathbf{x}^{\mathcal{I}}| = 1$. This shows that $X \in [0, B]$ almost surely over $\mathbf{x} \sim S^{(i)}$. Using an analogous argument for $\mathbf{x} \sim S_{\text{ref}}$, we overall obtain the following.

$$\int_{\tau=0}^{B} \mathop{\mathbb{P}}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau]\, d\tau = \mathop{\mathbb{E}}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x})] \quad \text{and} \quad \int_{\tau=0}^{B} \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x}) > \tau]\, d\tau = \mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x})] \tag{2}$$

We may now substitute (2) in the inequality (1), and use the fact that $\mathbb{E}_{\mathbf{x}\sim S_{\text{ref}}}[p^*(\mathbf{x})] \leq \epsilon/4$ (by the constraints of (P)) to conclude that

$$\lambda^* = \frac{1}{N} \sum_{\mathbf{x}\sim S^{(i)}} p^*(\mathbf{x}) = \frac{|S^{(i)}|}{N} \mathop{\mathbb{E}}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x})] \leq 2\epsilon/4 + \epsilon/2 = \epsilon$$

We reached a contradiction, since $\lambda^* > \epsilon$, and, therefore, τ^* exists. ■

We still need to show that whenever the procedure filters out clean examples, it also filters out an equal number of adversarial examples. Let $S_r^{(i)} = \{\mathbf{x} \in S^{(i)} : p^*(\mathbf{x}) > \tau^*\}$ be the set of points

that are filtered out during iteration i. We can write $S_r^{(i)}$ as a disjoint union $S_{r,\mathrm{cln}}^{(i)} \cup S_{r,\mathrm{adv}}^{(i)}$, where $S_{r,\mathrm{cln}}^{(i)} = S_r^{(i)} \cap S_{\mathrm{cln}}$ are the clean examples that are removed and $S_{r,\mathrm{adv}}^{(i)} = S_r^{(i)} \cap S_{\mathrm{adv}}$ are the adversarial examples that are removed.

Claim *With probability at least $1-\delta$, we have that for all non-terminating iterations, $|S_{r,\mathrm{cln}}^{(i)}| \leq |S_{r,\mathrm{adv}}^{(i)}|$.*

Proof By the previous claim, we know that τ^* (which defines the set $S_r^{(i)}$) exists and has the property that $\frac{|S^{(i)}|}{N}\,\mathbb{P}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau^*] \geq 2\,\mathbb{P}_{\mathbf{x}\sim S_{\mathrm{ref}}}[p^*(\mathbf{x}) > \tau^*] + \Delta$.

We first focus on the quantity $\mathbb{P}_{\mathbf{x}\sim S_{\mathrm{ref}}}[p^*(\mathbf{x}) > \tau^*]$, which is proportional to the number of reference examples that would be removed by the thresholding operation $p^*(\mathbf{x}) > \tau^*$. However, we are interested in the number of actual clean examples that would be removed. The reference examples can be shown to provide an estimate of the number of removed clean examples, through uniform convergence. In particular, the thresholding operation corresponds to a polynomial threshold function of degree at most d^k and, therefore, by standard VC dimension arguments (and uniformly for all iterations) we have that $\mathbb{P}_{\mathbf{x}\sim S_{\mathrm{ref}}}[p^*(\mathbf{x}) > \tau^*] \geq \mathbb{P}_{\mathbf{x}\sim S_{\mathrm{cln}}}[p^*(\mathbf{x}) > \tau^*] - \Delta/2$, except with probability δ , as long as the sample size is $N \geq C' \frac{d^k + \log(1/\delta)}{\Delta^2}$. This is because both S_{ref} and S_{cln} consist of N i.i.d. samples from the uniform distribution.

Overall, we have that $\frac{|S^{(i)}|}{N}\,\mathbb{P}_{\mathbf{x}\sim S^{(i)}}[p^*(\mathbf{x}) > \tau^*] \geq 2\,\mathbb{P}_{\mathbf{x}\sim S_{\mathrm{cln}}}[p^*(\mathbf{x}) > \tau^*]$. We can write the empirical probabilities in terms of the sizes of the removed sets to obtain the following, where we also use the fact that $|S_r^{(i)}| = |S_{r,\mathrm{cln}}^{(i)}| + |S_{r,\mathrm{adv}}^{(i)}|$ and that $|S_{r,\mathrm{cln}}^{(i)}|$ is at most equal to the number of clean examples that would be removed by the i-th filtering operation (some clean examples could already have been removed either by the adversary or by some previous iteration and these will not be contained in $S_{r,\mathrm{cln}}^{(i)}$).

$$\frac{|S^{(i)}|}{N} \cdot \frac{|S_r^{(i)}|}{|S^{(i)}|} \geq 2\frac{|S_{r,\mathrm{cln}}^{(i)}|}{N} \quad \text{or} \quad |S_{r,\mathrm{cln}}^{(i)}| + |S_{r,\mathrm{adv}}^{(i)}| \geq 2|S_{r,\mathrm{cln}}^{(i)}| \quad \text{or} \quad |S_{r,\mathrm{adv}}^{(i)}| \geq |S_{r,\mathrm{cln}}^{(i)}|$$

This concludes the proof of the claim. ■

Overall, if we sum over $i \in [N]$, we obtain that the number of clean examples that are removed by the procedure is at most equal to the number of adversarial examples that are removed by the procedure.

For part 2 of Lemma 4, we observe that $\sum_{\mathbf{x}\in S_{\mathrm{filt}}} p(\mathbf{x}) \leq \lambda^* N \leq \epsilon N$, as long as p satisfies all the constraints of the program (P). It suffices to prove the following claim.

Claim *Any non-negative polynomial p with degree at most k and $\mathbb{E}_{\mathbf{x}\sim \mathrm{Unif}_d}[p(\mathbf{x})] \leq \epsilon/8$ satisfies all the constraints of the program (P) with probability at least $1-\delta$.*

Proof The degree bound and non-negativity are satisfied directly by the definition of p. We now need to show that $\|p\|_{\mathrm{coef}} \leq 3^k d^{k/2}$. Recall that $p(\mathbf{x}) = \sum_{\mathcal{I}\subseteq[d]} c_p(\mathcal{I})\mathbf{x}^S$, where $c_p(\mathcal{I}) = 0$ for any $|\mathcal{I}| > k$ and $\|p\|_{\mathrm{coef}} = \sum_{\mathcal{I}\subseteq[d]} |c_p(\mathcal{I})|$. By viewing c_p as a vector with $\sum_{j=0}^{k}\binom{d}{j} \leq d^k$ dimensions (assuming $2 \leq k \leq d$), we have that $\|p\|_{\mathrm{coef}} = \|c_p\|_1 \leq d^{k/2}\|c_p\|_2 = d^{k/2}(\mathbb{E}_{\mathbf{x}\sim \mathrm{Unif}_d}[(p(\mathbf{x}))^2])^{1/2}$.

The uniform distribution is $(2,1)$-hypercontractive (see Theorem 9.22 in (O'Donnell, 2014)), and we, therefore, have

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathrm{Unif}_d}[(p(\mathbf{x}))^2])^{1/2} \leq e^k \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathrm{Unif}_d}[|p(\mathbf{x})|] \tag{3}$$

Recall that the polynomial p is non-negative. This implies that $|p(\mathbf{x})| = p(\mathbf{x})$ for all $\mathbf{x} \in \{\pm 1\}^d$ and therefore $\mathbb{E}_{\mathbf{x}\sim\mathrm{Unif}_d}[|p(\mathbf{x})|] = \mathbb{E}_{\mathbf{x}\sim\mathrm{Unif}_d}[p(\mathbf{x})] \leq \epsilon/8$. Overall, we have

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathrm{Unif}_d}[(p(\mathbf{x}))^2])^{1/2} \leq e^k\epsilon/8 \leq 3^k \tag{4}$$

Recall that $\|p\|_{\mathrm{coef}} \leq d^{k/2}(\mathbb{E}_{\mathbf{x}\sim\mathrm{Unif}_d}[(p(\mathbf{x}))^2])^{1/2}$. We obtain the desired bound $\|p\|_{\mathrm{coef}} \leq 3^k d^{k/2}$.

It remains to show that with probability at least $1-\delta$, we have $\frac{1}{N}\sum_{\mathbf{x}\in S_{\mathrm{ref}}} p(\mathbf{x}) \leq \epsilon/4$. Consider the random variable $X = \frac{1}{N}\sum_{\mathbf{x}\in S_{\mathrm{ref}}} p(\mathbf{x})$, where S_{ref} is drawn i.i.d. form Unif_d. We have that $\mathbb{E}[X] = \mathbb{E}_{\mathbf{x}\sim\mathrm{Unif}_d}[p(\mathbf{x})] \leq \epsilon/8$. Moreover, $p(\mathbf{x}) \leq \|p\|_{\mathrm{coef}} \leq 3^k d^{k/2}$, for all $\mathbf{x} \in \{\pm 1\}^d$ and, from a standard Hoeffding bound on the random variable X, we obtain that $\frac{1}{N}\sum_{\mathbf{x}\in S_{\mathrm{ref}}} p(\mathbf{x}) \leq \epsilon/4$ with probability at least $1 - \exp(-\frac{\epsilon^2}{64Nd^k9^k})$. Due to the choice of $N \geq C'\frac{d^k9^k}{\epsilon^2}\log(1/\delta)$, we have that the probability of failure is bounded by δ, as desired. ■

4. Finding a Low-Error Hypothesis

The outlier removal process of Lemma 4 enables us to find a subset S_{filt} of the input set such that all non-negative and low-degree polynomials with small expectation under the uniform distribution also have small empirical expectation under S_{filt}. Moreover, the number of clean examples removed to form S_{filt} is smaller than the number of removed outliers (see Figure 1). We show that these two properties are all we need in order to learn constant-depth circuits with contamination (Definition 1).

In order to take advantage of Lemma 4, we will use two main tools. The first one is the following theorem originally proposed by (Kalai et al., 2008) to show that $\mathcal{L}_1$ polynomial regression implies agnostic learning for classes that can be approximated by low-degree polynomials.

Theorem 5 (Learning through $\mathcal{L}_1$ polynomial regression (Kalai et al., 2008)) *Let $\mathcal{D}$ be any distribution over $\{\pm 1\}^d \times \{\pm 1\}$ and $\mathcal{C}$ some class of concepts from $\{\pm 1\}^d$ to $\{\pm 1\}$. If for each $f \in \mathcal{C}$ there is some polynomial p over $\{\pm 1\}^d$ of degree at most k such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|f(\mathbf{x}) - p(\mathbf{x})|] \leq \epsilon$, then there is an algorithm (based on degree-k $\mathcal{L}_1$ polynomial regression) which outputs a degree-k polynomial threshold function h such that $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq h(\mathbf{x})] \leq \min_{f\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq f(\mathbf{x})] + 2\epsilon$, in time $O(\frac{1}{\epsilon^2})d^{O(k)}\log(1/\delta)$.*

Our overall learning algorithm will first filter the input set of examples $\bar{S}_{\mathrm{inp}}$ using Algorithm 1 and then run the algorithm of Theorem 5 on the uniform distribution over the filtered set $\bar{S}_{\mathrm{filt}}$. All we need to show is that there is a low-degree polynomial p with $\mathbb{E}_{\mathbf{x}\sim\bar{S}_{\mathrm{filt}}}[|f(\mathbf{x}) - p(\mathbf{x})|] \leq \epsilon$. This is ensured by combining part 2 of Lemma 4 with the sandwiching approximators for constant-depth circuits originally proposed by (Braverman, 2008) in the context of pseudorandomness.

Theorem 6 (Sandwiching polynomials for AC^0) (Braverman, 2008; Tal, 2017; Harsha and Srinivasan, 2019) *Let $f : \{\pm 1\}^d \to \{\pm 1\}$ be any AC^0 circuit of size s and depth ℓ and let $\epsilon \in (0,1)$. Then, there are polynomials $p_{\mathrm{up}}, p_{\mathrm{down}}$ over $\{\pm 1\}^d$, each of degree at most $k = (\log(s))^{O(\ell)} \cdot \log(1/\epsilon)$ such that (1) $p_{\mathrm{up}}(\mathbf{x}) \geq f(\mathbf{x}) \geq p_{\mathrm{down}}(\mathbf{x})$ for all $\mathbf{x} \in \{\pm 1\}^d$ and (2) $\mathbb{E}_{\mathbf{x}\sim\mathrm{Unif}_d}[p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})] \leq \epsilon$.*

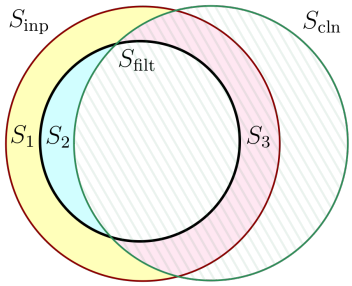

Figure 1: The diagram shows the input set of points S_{inp} (red circle), the clean points S_{cln} (green circle), the output S_{filt} (black circle) of Algorithm 1 and the sets S_1 (yellow region), S_2 (blue region), S_3 (pink region). The set S_{inp} consists of clean points, except from an η fraction of adversarial points. S_1 contains the adversarial points that are filtered out by the outlier removal process and S_2 contains the adversarial points that were not removed and are kept in S_{filt}. S_3 contains the clean points that were filtered out during outlier removal. Lemma 4 states that $|S_3| \leq |S_1|$ w.h.p.

Proof of Theorem 2. Consider the polynomial $p = p_{\text{up}} - p_{\text{down}}$, where $p_{\text{up}}, p_{\text{down}}$ are the $(\epsilon/8)$-sandwiching polynomials of some circuit f of size s and depth ℓ. Observe that p is non-negative and $\mathbb{E}_{\mathbf{x}\sim \text{Unif}_d}[p(\mathbf{x})] \leq \epsilon/8$. Therefore, according to part 2 of Lemma 4, we have $\sum_{\mathbf{x}\in S_{\text{filt}}} p(\mathbf{x}) \leq \epsilon N$. Since $p_{\text{up}} \geq f \geq p_{\text{down}}$ we also have $\mathbb{E}_{\mathbf{x}\sim S_{\text{filt}}}[|f(\mathbf{x}) - p_{\text{down}}(\mathbf{x})|] \leq \epsilon N/|\bar{S}_{\text{filt}}|$. By Theorem 5, we find $h : \{\pm 1\}^d \to \{\pm 1\}$ with

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq h(\mathbf{x})] \leq \min_{f\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq f(\mathbf{x})] + 2\epsilon N/|\bar{S}_{\text{filt}}| \text{ , or equivalently}$$

$$\sum_{(\mathbf{x},y)\in\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} \leq \min_{f\in\mathcal{C}} \sum_{(\mathbf{x},y)\in\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq f(\mathbf{x})\} + 2\epsilon N \tag{5}$$

The error of the hypothesis h on the set $\bar{S}_{\text{cln}}$ gives a bound on its error under the uniform distribution with high probability, due to classical VC theory, as long as $N \geq C' \frac{d^k + \log(1/\delta)}{\epsilon^2}$, because h is a PTF of degree at most k. We can provide an upper bound for $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{cln}}}[y \neq h(\mathbf{x})]$ in terms of the sizes of the sets depicted in Figure 1. In particular, we give a high-probability upper bound on the number of mistakes that h makes on $\bar{S}_{\text{cln}}$.

1. The points in $\bar{S}_{\text{cln}}$ that are removed by the adversary are not taken into account while forming h, so, in the worst case, h classifies them incorrectly. This gives at most $|S_{\text{cln}} \setminus S_{\text{inp}}|$ mistakes.

2. Similarly, h makes at most $|S_3|$ mistakes corresponding to the clean points that are removed during the outlier removal process.

3. Finally, h will make at most $|S_2| + 2\epsilon N$ mistakes on S_{filt}, according to the inequality (5), corresponding to the adversarially corrupted points that were not removed during the outlier removal process. In the worst case, all of these mistakes are made in the part of S_{filt} that intersects S_{cln}.

The overall error is $\frac{1}{N}(|S_{\text{cln}} \setminus S_{\text{inp}}| + |S_3| + |S_2|) + O(\epsilon)$. According to part 1 of Lemma 4, we have $|S_3| \leq |S_1|$. Moreover, by Definition 1, we have that $|S_{\text{cln}} \setminus S_{\text{inp}}| = |S_{\text{inp}} \setminus S_{\text{cln}}| = |S_1| + |S_2| = \eta N$. The error bound we obtain overall is $2\eta + O(\epsilon)$, as desired. ■

Remark 7 *For the proof of Theorem 2, the only property we used for the class of* AC^0 *circuits is that it admits low-degree sandwiching approximators (as per Theorem 6). Therefore, our results imply algorithms with runtime* $d^{O(k)}\,\mathrm{poly}(1/\epsilon)$ *for any class that admits degree-*k *sandwiching approximators with respect to the uniform distribution over the hypercube.*

Remark 8 *Several fundamental concept classes are known to have bounded sandwiching degree (see, for example, (Gopalan et al., 2010)), but the sandwiching degree can be significantly higher than the approximation degree in general. For example, the class of monotone functions in* d *dimensions is known to have approximation degree* $O(\sqrt{d})$ *(Bshouty and Tamon, 1996), but the sandwiching degree is* $\Omega(d)$ *(Rubinfeld and Vasilyan, 2023; Gollakota et al., 2023).*

Acknowledgments

We thank Mark Braverman and Sasha Razborov for useful conversations.

Adam Klivans was supported by NSF award AF-1909204 and the NSF AI Institute for Foundations of Machine Learning (IFML). Konstantinos Stavropoulos was supported by the NSF AI Institute for Foundations of Machine Learning (IFML) and by scholarships from Bodossaki Foundation and Leventis Foundation. Arsen Vasilyan was supported in part by NSF awards CCF-2006664, DMS-2022448, CCF-1565235, CCF-1955217, CCF-2310818, Big George Fellowship and Fintech@CSAIL. Part of this work was conducted while the author was visiting the Simons Institute for the Theory of Computing.

References

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):1–27, 2017.

Mark Braverman. Polylogarithmic independence fools AC 0 circuits. *Journal of the ACM (JACM)*, 57(5):1–10, 2008.

Nader H Bshouty and Christino Tamon. On the fourier spectrum of monotone functions. *Journal of the ACM (JACM)*, 43(4):747–770, 1996.

Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. Pac learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002. ISSN 0304-3975. doi: https://doi.org/10.1016/S0304-3975(01)00403-0. URL https://www.sciencedirect.com/science/article/pii/S0304397501004030. Algorithmic Learning Theory.

Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073, 2018.

Surbhi Goel, Abhishek Shetty, Konstantinos Stavropoulos, and Arsen Vasilyan. Tolerant algorithms for learning with arbitrary covariate shift. *Advances in Neural Information Processing Systems*, 37, 2024.

Aravind Gollakota, Adam R Klivans, and Pravesh K Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023.

Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *2010 IEEE 25th Annual Conference on Computational Complexity*, pages 223–234. IEEE, 2010.

Prahladh Harsha and Srikanth Srinivasan. On polynomial approximations to AC^0. *Random Structures & Algorithms*, 54(2):289–303, 2019.

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22 (4):807–837, 1993. doi: 10.1137/0222052. URL https://doi.org/10.1137/0222052.

Michael Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '93, page 372–381, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897915917. doi: 10.1145/167088.167197. URL https://doi.org/10.1145/167088.167197.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2887–2943. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/klivans24a.html.

Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023.

Jie Shen and Chicheng Zhang. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 1072–1113. PMLR, 2021. URL http://proceedings.mlr.press/v132/shen21a.html.

Avishay Tal. Tight Bounds on the Fourier Spectrum of AC0. In Ryan O'Donnell, editor, *32nd Computational Complexity Conference (CCC 2017)*, volume 79 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 15:1–15:31, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-040-8. doi: 10.4230/LIPIcs.CCC.2017.15. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.CCC.2017.15.

Leslie G. Valiant. Learning disjunction of conjunctions. In Aravind K. Joshi, editor, *Proceedings of the 9th International Joint Conference on Artificial Intelligence. Los Angeles, CA, USA, August 1985*, pages 560–566. Morgan Kaufmann, 1985. URL http://ijcai.org/Proceedings/85-1/Papers/107.pdf.