arXiv:2505.20177v1 [cs.LG] 26 May 2025

The Power of Iterative Filtering for Supervised Learning with (Heavy) Contamination

Adam R. Klivans*
klivans@cs.utexas.edu
UT Austin

Konstantinos Stavropoulos†
kstavrop@cs.utexas.edu
UT Austin

Kevin Tian
kjtian@cs.utexas.edu
UT Austin

Arsen Vasilyan‡
arsenvasilyan@gmail.com
UT Austin

Abstract

Inspired by recent work on learning with distribution shift, we give a general outlier removal algorithm called *iterative polynomial filtering* and show a number of striking applications for supervised learning with contamination:

1. We show that any function class that can be approximated by low-degree polynomials with respect to a hypercontractive distribution can be efficiently learned under bounded contamination (also known as *nasty noise*). This is a surprising resolution to a longstanding gap between the complexity of agnostic learning and learning with contamination, as it was widely believed that low-degree approximators only implied tolerance to label noise.
2. For any function class that admits the (stronger) notion of sandwiching approximators, we obtain near-optimal learning guarantees even with respect to heavy additive contamination, where far more than 1/2 of the training set may be added adversarially. Prior related work held only for regression and in a list-decodable setting.
3. We obtain the first efficient algorithms for tolerant testable learning of functions of halfspaces with respect to any fixed log-concave distribution. Even the non-tolerant case for a single halfspace in this setting had remained open.

These results significantly advance our understanding of efficient supervised learning under contamination, a setting that has been much less studied than its unsupervised counterpart.

*Supported by NSF award AF-1909204 and the NSF AI Institute for Foundations of Machine Learning (IFML).

†Supported by the NSF AI Institute for Foundations of Machine Learning (IFML) and by scholarships from Bodossaki Foundation and Leventis Foundation.

‡Supported by the NSF AI Institute for Foundations of Machine Learning (IFML).

1 Introduction

Dataset curation is a fundamental part of the training pipeline of modern machine learning models and often appears to be the bottleneck in obtaining models with improved performance [SKL17, LRB+21, BGMMS21, GIF+23]. One way to theoretically model this problem is to assume that the learner has access to a—potentially heavily—contaminated dataset and the goal is to learn a model that performs well on some clean underlying target distribution. While there has been tremendous recent progress for unsupervised learning with contamination [HM13, CSV17, KS17a, KS17b, BDLS17, DKS18b, HL18, CDG19, RY20, BDH+20, CMY20, BK21, IK22, ZS22, DKK+22a, DHPT24], relatively little is known for *supervised* learning with contamination, especially for binary classification.

Many efficient algorithms with strong error guarantees have been developed for *agnostic* learning, a special case of contamination where only the labels are adversarially corrupted [KKMS08, KOS08, BOW10, ABL17, DKS18a, DKTZ20b, DKK+22b]. Most of these guarantees, however, had seemed difficult to extend to the more challenging setting of contamination, where both labels and covariates can be adversarially corrupted. In this paper—building on recent work on robust learning and learning with distribution shift [DKS18a, GSSV24, KSV24c]—we give a general iterative polynomial filtering procedure that greatly expands the set of known positive results for learning binary classifiers from contaminated datasets (see Tables 1, 2, and 3). In particular, we obtain the surprising conclusion that all known near-optimal error guarantees for agnostic learning (that can be achieved efficiently) can indeed be extended to the setting of contamination.

Learning with Bounded Contamination. The earliest works that explored learnability beyond label noise date back more than 30 years ago [Val85, KL93]. Since then, the problem has been studied in the context of learning with *malicious* [Val85, KL93, KLS09, She25] and *nasty* [BEK02, ABL17, DKS18a, GSSV24, KSV24c] noise. Here we focus on the harshest among these noise models, nasty noise, which we call bounded contamination (defined below), in line with recent work in robust learning (see [DK23, KSV24c]). In this model, the adversary is allowed to replace an arbitrarily chosen but bounded fraction of a clean dataset with arbitrary labeled datapoints.

Definition 1.1 (Bounded Contamination (BC))**.** *Let $\mathcal{D}$ be some distribution over $\mathcal{X}$, $\eta \in (0,1)$ and $f : \mathcal{X} \to \{\pm 1\}$. We say that a set of samples $\bar{S}_{\mathrm{inp}}$ is generated by $(\mathcal{D}, f)$ with bounded contamination of rate η if it is generated as follows for some $M \geq 1$.*[1]

1. *First, a set $\bar{S}_{\mathrm{cln}}$ of M i.i.d. examples of the form $(\mathbf{x}, f(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{D}$, is drawn.*
2. *Then, an adversary receives $\bar{S}_{\mathrm{cln}}$, chooses at most ηM labeled examples in $\bar{S}_{\mathrm{cln}}$ and substitutes them with an equal number of arbitrary labeled examples $\bar{S}_{\mathrm{adv}}$ to form $\bar{S}_{\mathrm{inp}}$.*

The learner receives a dataset with bounded contamination, and the goal is to output a classifier that enjoys approximately optimal error guarantees on the clean underlying target distribution. It is known that achieving an error better than twice the contamination rate is, in general, impossible [BEK02].

Definition 1.2 (BC-Learning)**.** *An algorithm $\mathcal{A}$ is a BC-learner for $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ if on input $(\epsilon, \delta, \bar{S}_{\mathrm{inp}})$, where $\epsilon, \delta \in (0,1)$, and $\bar{S}_{\mathrm{inp}}$ is generated by $(\mathcal{D}, f)$ with bounded contamination η for some distribution $\mathcal{D}$ over $\mathcal{X}$, some $f \in \mathcal{C}$ and $\eta \in [0,1)$, the algorithm $\mathcal{A}$ outputs some hypothesis*

[1] We use the notation $\bar{S}$ to denote a labeled dataset and distinguish it from its unlabeled counterpart S.

$h : \mathcal{X} \to \{\pm 1\}$ *such that with probability at least* $1 - \delta$ *over the clean examples in* $\bar{S}_{\text{inp}}$*, and the randomness of* $\mathcal{A}$*:*

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{D}}[f(\mathbf{x}) \neq h(\mathbf{x})] \leq 2\eta + \epsilon$$

The sample complexity of $\mathcal{A}$ *is the minimum number of examples required to achieve the above guarantee. Moreover, a distribution-specific BC-learner with respect to some distribution* $\mathcal{D}^*$ *over* $\mathcal{X}$ *is a BC-learner that is guaranteed to work only when* $\mathcal{D} = \mathcal{D}^*$*.*

Most of the computationally efficient algorithms for learning with bounded contamination provide suboptimal error guarantees and apply only to special concept classes [KLS09, ABL17, DKS18a]. Nevertheless, a recent line of works inspired by advances in learning with distribution shift [GKKM20, KSV24b, GSSV24] has given efficient algorithms with near-optimal guarantees for concept classes that admit sandwiching polynomial approximators [GSSV24, KSV24c]. In contrast, for agnostic learning (i.e., adversarial label noise), it is well known that the weaker notion of (non-sandwiching) approximating polynomials is sufficient [KKMS08], and there is strong evidence of its necessity [DSFT+14, DKPZ21].

Therefore, the following question naturally arises: *does the existence of low-degree approximating polynomials imply efficient learnability even with respect to bounded contamination?*

In Theorem 4.2, we give a positive answer to this question, thereby resolving a longstanding gap between the complexity of learning with bounded contamination and with adversarial label noise. This is particularly important as it implies exponential improvements for BC-learning of fundamental concept classes like intersections of halfspaces, monotone functions, and convex sets (see Table 2).

Learning with Heavy Contamination. Perhaps surprisingly, to our knowledge, binary classification beyond bounded contamination is completely unexplored. In contrast, there is a substantial body of work for learning from datasets where only a small proportion comes from the clean distribution in unsupervised settings [CSV17, DKS18b, RY20, BK21, IK22, ZS22, DHPT24] and linear regression [RY20, KKK19, DHPT24]. These works typically provide list-decodable guarantees (i.e., provide multiple candidate hypotheses only one of which has low error) or require access to a small trusted clean sample. Here, we define a new model for learning binary classifiers with heavy additive contamination that outputs a single hypothesis with a strong error guarantee under the clean distribution. Our model is inspired by recent work on regression (a basic supervised learning task) in additive semi-random models [JLM+23, KLL+23].

Definition 1.3 (Heavily Contaminated (HC) Datasets)**.** *Let* $\bar{\mathcal{D}}$ *be some distribution over* $\mathcal{X} \times \{\pm 1\}$ *(we think of* $\bar{\mathcal{D}}$ *as the clean or uncorrupted distribution). We say that a set of samples* $\bar{S}_{\text{inp}}$ *is generated by* $\bar{\mathcal{D}}$ *with* Q*-heavy contamination if it is generated as follows for some* $m \leq M$ *with* $M/m \leq Q$*.*

1. *First, a set* $\bar{S}_{\text{cln}}$ *of* m *i.i.d. labeled examples from* $\bar{\mathcal{D}}$ *is drawn.*
2. *Then, an adversary receives* $\bar{S}_{\text{cln}}$ *and adds* $M - m$ *arbitrary labeled examples to form* $\bar{S}_{\text{inp}}$*.*

The heavy contamination model only allows the adversary to add points, since removing an arbitrary fraction of the clean samples would correspond to learning with truncation [DGTZ18, DGTZ19, KZZ24], which is beyond the scope of this work (see also Remark 4.6). Another difference between Definition 1.1 and Definition 1.3 is that in the HC model, clean labels need not be realized by some function in the given concept class. We instead consider the following quantity

$$\mathsf{opt}_{\text{total}} = \min_{f \in \mathcal{C}} \frac{1}{|\bar{S}_{\text{inp}}|} \sum_{(\mathbf{x}, y) \in \bar{S}_{\text{inp}}} \mathbb{1}\{y \neq f(\mathbf{x})\}\,, \tag{1.1}$$

which is the minimum error achievable by the concept class $\mathcal{C}$ on the whole (contaminated) input dataset $\bar{S}_{\text{inp}}$ (including the misclassification errors on the clean samples). The error benchmark we consider is a rescaling of $\mathsf{opt}_{\text{total}}$, proportional to the heavy contamination ratio Q.

Definition 1.4 (HC-Learning)**.** *An algorithm $\mathcal{A}$ is an HC-learner for $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ if on input $(\epsilon, \delta, Q, \bar{S}_{\text{inp}})$, where $\epsilon, \delta \in (0,1)$, $Q \geq 1$ and $\bar{S}_{\text{inp}}$ is a Q-heavily contaminated set of labeled examples generated by distribution $\bar{\mathcal{D}}$ (as described in Definition 1.3), the algorithm $\mathcal{A}$ outputs some hypothesis $h : \mathcal{X} \to \{\pm 1\}$ such that with probability at least $1 - \delta$ over the clean examples in $\bar{S}_{\text{inp}}$, and the randomness of $\mathcal{A}$:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq h(\mathbf{x})] \leq Q \cdot \mathsf{opt}_{\text{total}} + \epsilon\,, \text{ where } \mathsf{opt}_{\text{total}} \text{ is given by Eq. (1.1)}$$

The (clean) sample complexity of $\mathcal{A}$ is the minimum number of clean examples $\bar{S}_{\text{inp}}$ needs to contain in order to achieve the above guarantee. Moreover, a distribution-specific HC-learner with respect to $\mathcal{D}^*$ is an HC-learner that is guaranteed to work only when the marginal of $\bar{\mathcal{D}}$ on $\mathcal{X}$ is $\mathcal{D} = \mathcal{D}^*$.

We show that the dependence on $\Omega(Q \cdot \mathsf{opt}_{\text{total}})$ is, in fact, necessary, even if the clean labels are realized by the learned class $\mathcal{C}$. The quantity $Q \cdot \mathsf{opt}_{\text{total}}$ equals the number of errors $|\bar{S}_{\text{inp}}| \cdot \mathsf{opt}_{\text{total}}$ of the optimal classifier on the input set $\bar{S}_{\text{inp}}$ divided by the size of the clean dataset. Our lower bound essentially shows the existence of a contamination strategy that forces any HC learner to pay for all the mistakes of the optimal classifier on $\bar{S}_{\text{inp}}$, even if these are not made on the clean dataset. This is possible as the learner does not know which subset of $\bar{S}_{\text{inp}}$ is clean. In the following, we let $\mathsf{opt}_{\text{clean}} := \min_{f\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})]$ be the optimum error under the clean distribution.

Proposition 1.5 (Informal, see Propositions B.2 and B.3)**.** *Let $\mathcal{C}$ be any non-trivial class. Then, no HC-learner for $\mathcal{C}$ can guarantee error better than $\frac{1}{2} \cdot Q \cdot \mathsf{opt}_{\text{total}}$, even when the clean distribution is realizable, i.e., $\mathsf{opt}_{\text{clean}} = 0$. Moreover, if $|\mathcal{X}| < \infty$, then no HC-learner for $\mathcal{C}_{\text{all}} = \{\pm 1\}^{\mathcal{X}}$ can guarantee error better than $Q \cdot \mathsf{opt}_{\text{total}}$, even when $\mathsf{opt}_{\text{clean}} = 0$.*

Naturally, one might wonder whether the existence of approximating polynomials is sufficient for HC-learning. In Theorem 5.2 we give a negative answer to this question, by providing a lower bound on the sample complexity of HC-learning of monotone functions, which admit low-degree approximators. A recent line of works has used the stronger notion of sandwiching approximators to provide efficient algorithms for various challenging learning tasks [GKK23, KSV24b, GSSV24, CKK+24b, KSV24c]. Here, we expand on this paradigm and show that the existence of low-degree sandwiching polynomials implies efficient HC-learning as well (Theorem 4.4).

1.1 Our Results

Learning from Contaminated Datasets. In Table 1, we present an overview of the upper bounds we obtain as applications of our main theorems on learning with contamination (Theorems 4.2 and 4.4). For comparison, we also provide the corresponding results on learning with label noise from prior work. For the case of bounded contamination, and for constant error and confidence parameters (ϵ, δ), our results match the best known bounds for learning with label noise, since in both cases, the existence of low-degree polynomial approximators is sufficient. See Tables 2 and 3 for further details.

All of our results work in the distribution-specific setting. Distributional assumptions are unavoidable, as there is strong evidence that distribution-free learning of even the simplest classes (e.g., linear classifiers) in the presence of noise is computationally hard [FGKP06, Dan16]. Although we present our results for specific standard target marginal distributions, our results hold for

Concept Class	Target Marginal	Adversarial Label Noise *(Prior work)*	Bounded Contamin. *(This work)*	2-Heavy Contamin. *(This work)*
Intersections of k Halfspaces	$\mathcal{N}(0, \mathbf{I}_d)$	$d^{O(\log k)}$	$d^{O(\log k)}$	$d^{O(k^6)}$
Depth-t, Size-d Boolean Circuits	$\mathrm{Unif}\{\pm 1\}^d$	$d^{O(\log d)^{t-1}}$	$d^{O(\log d)^{t-1}}$	$d^{O(\log d)^{O(t)}}$
Degree-k PTFs	$\mathcal{N}(0, \mathbf{I}_d)$	$d^{O(k^2)}$	$d^{O(k^2)}$	$d^{O_k(1)}$
Monotone Functions	$\mathrm{Unif}\{\pm 1\}^d$	$2^{\tilde{O}(\sqrt{d})}$	$2^{\tilde{O}(\sqrt{d})}$	$2^{\Omega(d)}$

Table 1: Upper and lower bounds on the time complexity of learning in different noise models up to excess error $\epsilon = 0.1$ and failure probability $\delta = 0.01$. See Appendix D for more details on the complexity of learning with adversarial label noise.

any hypercontractive (Definition 3.1) marginal $\mathcal{D}^*$ that can be sampled efficiently, as long as the degree of approximation (resp. sandwiching for HC-learning) of the learned class is low under $\mathcal{D}^*$. Hypercontractive distributions are an extremely wide class of probability distributions, which includes Gaussians, all log-concave distributions over $\mathbb{R}^d$, as well as product distributions over $\{\pm 1\}^d$. See Appendix C for an overview of relevant results in approximation theory.

On the lower bound side, we show that learning monotone functions with heavy contamination requires exponentially many samples, thereby separating bounded and heavy contamination. For a more thorough discussion on lower bounds for learning with contamination, see Section 5.[2]

Testable Learning. A recent line of works in learning theory has focused on providing learning algorithms that can verify their distributional assumptions [RV23, GKK23, GKSV23, DKK+23, GKSV24, STW24, GSSV24]. These algorithms are allowed to either accept and output a classifier with certified optimal performance, or detect a violation of their target distributional assumptions and reject. Here, we provide improved results for a tolerant version of this problem where the algorithm has to accept even if the input distribution is close to the target distribution. Tolerant testable learning was first studied in [GSSV24] (see Definition 4.7).

In Theorem 4.8, we show that the existence of low-degree sandwiching polynomials implies efficient tolerant testable learning. Previously, [GSSV24] required the existence of low-degree $\mathcal{L}_2$-sandwiching polynomials, which is a stronger assumption. Moreover, all of the previous results in testable learning—even the non-tolerant variants—required that the sandwiching polynomials have bounded coefficients. Here, we do not impose such a requirement and obtain the first testable learning results for functions of halfspaces with respect to any fixed log-concave distribution. All prior work either gave worse error guarantees [GKSV24, GKSV23], or required target marginals with strictly sub-exponential tails [GKK23]. See Section 4.3 for more details.

[2]We note that although learning with bounded contamination is clearly more challenging than learning with adversarial label noise, many of the known lower bounds for agnostic learning do not transfer directly to BC-learning as formalized in Definition 1.1 for technical reasons. See Section 5 for a way to circumvent this.

1.2 Our Techniques

Iterative Polynomial Filtering. All of our results use the same iterative polynomial filtering algorithm of Theorem 3.2 with appropriate hyperparameter choices. The algorithm receives a set S_{inp} of data points and filters it, outputting a subset S_{filt} of S_{inp} that satisfies the following two conditions. First, any low-degree polynomial p whose absolute expectation $\mathbb{E}[|p(\mathbf{x})|]$ over the pre-specified target distribution $\mathcal{D}^*$ is small, will also have bounded average over S_{filt}, i.e. $\frac{1}{|S_{\mathrm{filt}}|}\sum_{\mathbf{x}\in S_{\mathrm{filt}}} p(\mathbf{x})$ is small. Second, if S_{inp} contains a set of points S that were initially generated independently by $\mathcal{D}^*$, then only a limited number of points from S can be removed by the filtering (the allowed number of removed points is controlled by an appropriate hyperparameter).

Our algorithm is a refined version of bounded-degree outlier removal procedures from robust learning and learning with distribution shift [DKS18a, DKK+19a, GSSV24, KSV24c]. The general principle behind these algorithms is that one can iteratively find polynomials that violate the desired condition over the input set and use them to filter the input points. In particular, the algorithm removes the points that give such polynomials values larger than a threshold. By choosing this threshold appropriately, one can control the proportion of removed points that lie in the clean set S in each step.

Prior work only gave guarantees for squared polynomials [DKS18a, GSSV24] and for non-negative polynomials [KSV24c]. Here, we give a filtering algorithm that preserves the expectation of any polynomial, as long as its absolute expectation is small with respect to the target distribution. This is crucial for our application in BC-learning. Moreover, our filtering procedure works for any hypercontractive target distribution $\mathcal{D}^*$, which is also true for [DKS18a, GSSV24] but not for [KSV24c], which only works for the uniform distribution over the hypercube.

Learning with Bounded Contamination. Our algorithm follows a two-phase approach: (1) Run our outlier-removal algorithm, and obtain a filtered subset S_{filt}. (2) Following an approach similar to [KKMS08], construct a predictor based on the polynomial $\widehat{p}$ with smallest $\mathcal{L}_1$ error on S_{filt}. We now give a sketch of the analysis of this algorithm and explain how an error of $O(\eta+\epsilon)$ can be guaranteed. (The optimal dependence of $2\eta+\epsilon$ is obtained by carefully refining the analysis below.)

What would happen if we ran the phase (2) without filtering the dataset beforehand? As shown in [KKMS08], this approach works in the agnostic setting, i.e. when an adversary can corrupt only the labels but not the examples. A key observation in [KKMS08] is that if p^* is an ϵ-approximating polynomial for the ground truth f^*, then after η fraction of data labels are corrupted, the polynomial p^* will have an $\mathcal{L}_1$-error of only at most $\eta+\epsilon$. However, in the more challenging setting of learning with contamination over $\mathbb{R}^d$, even a single corrupted data-point can cause the $\mathcal{L}_1$-error of p^* to be arbitrarily large. The reason is that any non-zero polynomial over $\mathbb{R}^d$ will be arbitrarily large in absolute value when evaluated at inputs $\mathbf{x}$ far enough from the origin.

Hence, the first phase of our algorithm aims to filter out such bad input datapoints. The following basic observation is key to our approach: if p^* is an ϵ^4-approximator in $\mathcal{L}_2$ norm for a $\{\pm 1\}$-valued function f^*, then the average $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[|(p^*(\mathbf{x}))^2-1|]$ is at most $O(\epsilon^2)$. This observation, together with our filtering guarantee, ensures that the average $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{filt}}}[(p^*(\mathbf{x}))^2-1]$ is likewise bounded by $O(\epsilon)$ while removing almost exclusively outliers.

Yet, the set S_{filt} might still contain many outliers. We show that the condition $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{filt}}}[(p^*(\mathbf{x}))^2-1]\le O(\epsilon)$ implies that the remaining outliers are not dangerous when it comes to phase (2) of our algorithm. Indeed, this condition tells us that the number of remaining outliers $\mathbf{x}$ in S_{filt} with $|p^*(\mathbf{x})|>\tau$ is at most $O(|S_{\mathrm{filt}}|\epsilon/\tau^2)$. Taking $\tau=2$, we see that the total contribution to the $\mathcal{L}_1$ error of p^* on S_{filt} of outliers $\mathbf{x}$ with $|p^*(\mathbf{x})|>2$ is $O(\epsilon)$. The remaining outliers contribute at most

$O(\eta)$ to this error, since there are at most $O(\eta)$ of such outliers and each satisfies $|p^*(\mathbf{x})| \leq 2$.

Overall, we see that the $\mathcal{L}_1$ error of p^* on S_{filt} is at most $O(\eta + \epsilon)$, and therefore the polynomial $\widehat{p}$ found in phase 2 of our algorithm will also have an $\mathcal{L}_1$ error of at most $O(\eta + \epsilon)$ on S_{filt}. Since phase (1) only removed at most $O(\epsilon)$ clean datapoints, we conclude that the $\mathcal{L}_1$ error of $\widehat{p}$ on the clean dataset S_{cln} is likewise $O(\eta + \epsilon)$, which we use to bound the out-of-distribution error of $\widehat{p}$.

Learning with Heavy Contamination. Our heavy contamination algorithm follows the same structure as the one for bounded contamination: we first run the iterative filtering algorithm and then run $\mathcal{L}_1$ polynomial regression on the filtered dataset. This time, however, we choose the hyperparameters of the iterative filtering algorithm so that the proportion of removed points that are clean is inversely proportional to the heavy contamination ratio Q. In this way, we make sure to remove only a small fraction of the clean points.

To conclude the proof, we use the notion of sandwiching polynomials. In particular, if $f^* \in \mathcal{C}$ is the optimum classifier on the input dataset, and $p_{\mathrm{up}}, p_{\mathrm{down}}$ are two low-degree polynomials such that (1) $p_{\mathrm{up}}(\mathbf{x}) \geq f^*(\mathbf{x}) \geq p_{\mathrm{down}}(\mathbf{x})$ for all $\mathbf{x}$ and (2) $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}^*}[p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})] \leq O(\epsilon^2/Q)$, then the filtering process guarantees that $\mathbb{E}_{\mathbf{x} \sim S_{\mathrm{filt}}}[p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})]$ scales proportionally to ϵ. Overall, this implies that p_{down} has low $\mathcal{L}_1$-error under S_{filt} and that $\mathcal{L}_1$-polynomial regression achieves near-optimal error guarantees.

1.3 Related Work

Supervised Learning with Noise. The majority of existing works on robust supervised learning focus on label noise. In order to obtain efficient algorithms, it is standard and often necessary to make assumptions on the marginal distribution [Dan16, DK22, DKMR22], although there have been recent attempts to relax those assumptions [CKK+24a]. Even under common distributional assumptions, the best possible error guarantees often require exponential dependence on the excess error parameter ϵ [DSFT+14, DKPZ21]. A line of works was focused on providing faster learning algorithms for some classes at the expense of relaxed error guarantees [ABL17, DKS18a, DKTZ20b] or under restricted noise models [DKTZ20a, DKK+22b]. We consider the more challenging scenario of learning with contamination, where there is noise on both the input examples and their labels. Before this work, efficient algorithms for learning with bounded contamination up to optimal error were known only for classes with low sandwiching degree [GSSV24, KSV24c], and nothing was known about classification under heavy contamination.

Semi-Random Models. Our Definition 1.3 is inspired by semi-random models, which lie between the average and worst case settings [BS95, FK01]. In particular, it resembles an instantiation of the semi-random model framework known as a *monotone adversary*, which breaks a statistical assumption used by a learning algorithm (e.g., i.i.d. draws from a known distribution) by providing additional data. Our approach is inspired by recent algorithms for supervised regression problems, e.g., solving linear systems, sparse recovery, or matrix completion [CG18, KLL+23, JLM+23, KLL+24], that are tolerant to monotone adversaries. These algorithms also use reweightings that come with certificates of success. However, a major qualitative difference between the aforementioned works and ours is that our Definition 1.3 does not require that the adversary uses labels consistent with a "clean hypothesis." Instead, our algorithms can tolerate label noise (alongside covariate noise) and achieve the information-theoretically optimal clean error under such a model (Proposition 1.5).

Testable Learning. In recent years, a number of works has focused on verifying the assumptions of learning algorithms. Testable learning was introduced by [RV23] in the context of verifying

the distributional assumptions of agnostic learners and there are several subsequent works on this setting [GKK23, GKSV24, DKK+23, GKSV23, DKLZ24, STW24]. This paradigm has since been expanded to testing for distribution shifts that may harm the performance of supervised learning algorithms [KSV24b, KSV24a, CKK+24b, CKLS25], or even testing noise assumptions [GKSV25]. Here, we study a tolerant version of testable agnostic learning that was first studied by [GSSV24], and provide the first guarantees for halfspaces with respect to any fixed log-concave measure.

Learning with heavy contamination can be thought of as a search version of testable learning. More specifically, in testable learning the goal is to decide whether the input dataset is structured enough so that a near-optimal hypothesis can be found efficiently, while in HC-learning the goal is to *find* a subset of the input that is structured enough. An analogous connection was observed in [GSSV24] between TDS learning [KSV24b] and PQ learning [GKKM20] in the context of learning under distribution shift. There, the goal was to either decide whether the (unlabeled) test examples come from a distribution that is similar to the one the learner has trained on (TDS), or find a subset of the test examples where the learner is confident in its predictions (PQ). See Remark 4.5.

2 Notation

We consider a d-dimensional feature space $\mathcal{X}$ which will either be $\mathbb{R}^d$ or the hypercube $\{\pm 1\}^d$. A polynomial p over $\mathbb{R}^d$ is of the form $p(\mathbf{x}) = \sum_{\alpha \subseteq \mathbb{N}^d} c_p(\alpha)\, \mathbf{x}^\alpha$, where $\mathbf{x}^\alpha = \prod_{i \in [d]} x_i^{\alpha_i}$ is a monomial of degree $\|\alpha\|_1$. The degree of p is equal to the maximum $\|\alpha\|_1$ such that $c_p(\alpha) \neq 0$. Over $\{\pm 1\}^d$ we use the multilinear expansion $p(\mathbf{x}) = \sum_{\mathcal{I} \subseteq [d]} c_p(\mathcal{I}) \mathbf{x}^{\mathcal{I}}$, where $\mathbf{x}^{\mathcal{I}} = \prod_{i \in \mathcal{I}} x_i$. In both cases, we call $\mathbf{c}_p$ the vector of coefficients. We consider sets of points S to contain examples that are separate instances of the correponding elements of $\mathcal{X}$. We denote with $\bar{S}$ the corresponding labeled set of examples. We use $\mathbf{x} \sim S$ to say that $\mathbf{x}$ is drawn uniformly from S. For a labeled distribution $\bar{\mathcal{D}}$ over $\mathcal{X} \times \{\pm 1\}$, $\mathcal{D}$ is the marginal on $\mathcal{X}$. In the following the distribution-specific algorithms have sample access to the unlabeled target distribution $\mathcal{D}^*$. We use $\mathcal{N}_d = \mathcal{N}(0, \mathbf{I}_d)$ to denote the standard d-dimensional Gaussian and $\mathrm{Unif}_d = \mathrm{Unif}\{\pm 1\}^d$ for the uniform over the d-dimensional hypercube.

3 Iterative Polynomial Filtering

We first define hypercontractivity, which is a crucial assumption for our filtering procedure.

Definition 3.1 (Hypercontractivity)**.** We say that a distribution $\mathcal{D}$ over $\mathcal{X}$ is A-hypercontractive with respect to polynomials for some $A \geq 1$ if for any polynomial p over $\mathcal{X}$ and any $t \geq 2$ we have

1. $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[|p(\mathbf{x})|^t] \leq (At)^{\ell t} \left(\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[|p(\mathbf{x})|] \right)^t$, where $\ell = \deg(p)$
2. The absolute value of any degree-1 monomial has finite expectation under $\mathcal{D}$.

Our main algorithmic tool is the following theorem, which gives a way to filter an arbitrary set of examples in order to preserve the expectations of polynomials whose expected absolute value under some hypercontractive target distribution is small.

Theorem 3.2 (Iterative Filtering)**.** *Let $\mathcal{D}^*$ be a A-hypercontractive distribution over a d-dimensional space $\mathcal{X}$. Consider parameters $\epsilon, \delta \in (0,1)$ and $R, \ell, m \geq 1$, and let S_{inp} be an arbitrary set of examples in $\mathcal{X}$ and S_{ref} a set of m_{ref} i.i.d. examples from $\mathcal{D}^*$. For a sufficiently large universal constant $C \geq 1$, if $m_{\mathrm{ref}} \geq \frac{R^2 (CAd)^{2\ell}}{\epsilon^3} (\log \frac{1}{\delta})^{4\ell+1}$, then Algorithm 1 on input $(S_{\mathrm{inp}}, S_{\mathrm{ref}}, m, \ell, R, \epsilon)$ runs in time $\mathrm{poly}(|S_{\mathrm{inp}}|, m_{\mathrm{ref}}, (d+1)^\ell)$ and outputs $S_{\mathrm{filt}} \subseteq S_{\mathrm{inp}}$ such that the following hold:*

1. *Let S_{cln} be any set of m i.i.d. examples from $\mathcal{D}^*$ where $m \geq CR^2 \frac{(2A(d+1))^{2\ell}}{\epsilon^3} \log \frac{1}{\delta}$. Suppose that S_{inp} is formed by first removing an arbitrary fraction of points in S_{cln} and then adding any number of arbitrary points. Then, the algorithm removes a relatively small number of the examples in S_{cln} that appear in S_{inp}:*

$$|(S_{\mathrm{cln}} \cap S_{\mathrm{inp}}) \setminus S_{\mathrm{filt}}| \leq \frac{1}{R} \cdot |S_{\mathrm{inp}} \setminus S_{\mathrm{filt}}| + \frac{\epsilon m}{2}\,, \text{ with probability at least } 1-\delta \text{ over } S_{\mathrm{cln}}, S_{\mathrm{ref}}$$

2. *For any polynomial p of degree at most ℓ, and $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[|p(\mathbf{x})|] \leq \frac{\epsilon}{4R}$ we have:*

$$\sum_{\mathbf{x}\in S_{\mathrm{filt}}} p(\mathbf{x}) \leq \epsilon m\,, \text{ with probability at least } 1-\delta \text{ over } S_{\mathrm{ref}}$$

Algorithm 1: Iterative Polynomial Filtering

Input: S_{inp} set of M points, S_{ref} set of m_{ref} points, $m, \ell \in \mathbb{N}$, $R \geq 1$, $\epsilon \in (0,1)$
Output: Set $S_{\mathrm{filt}} \subseteq S_{\mathrm{inp}}$.

Let $\beta \leftarrow 2(2A)^{2\ell}$; $\gamma \leftarrow \frac{\epsilon}{2R}$; $B \leftarrow 4(d+1)^{\frac{\ell}{2}}(\frac{\beta}{\epsilon})^{\frac{1}{2}}$; $\Delta \leftarrow \frac{\epsilon}{2B}$;
Let $\mathcal{P}$ denote the family of polynomials p of degree at most ℓ for which we have: $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{ref}}}[|p(\mathbf{x})|] \leq \gamma$ and $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{ref}}}[(p(\mathbf{x}))^2] \leq \beta$;
$S \leftarrow \{\mathbf{x} \in S_{\mathrm{inp}} : |p(\mathbf{x})| \leq B \text{ for all } p \in \mathcal{P}\}$;
for $i = 0, 1, 2, \ldots, M$ **do**
  Compute p^* and λ^* as follows.

$$p^* = \arg\max_{p\in\mathcal{P}} \sum_{\mathbf{x}\in S} p(\mathbf{x}) \quad \text{and} \quad \lambda^* = \frac{1}{m}\sum_{\mathbf{x}\in S} p^*(\mathbf{x})$$

  if $\lambda^* \leq \epsilon$ **then** return $S_{\mathrm{filt}} \leftarrow S$;
  else
    Let $\tau^* \geq 0$ be the smallest value such that $\frac{|S|}{m}\mathbb{P}_{\mathbf{x}\sim S}[|p^*(\mathbf{x})| > \tau^*] \geq R \cdot \mathbb{P}_{\mathbf{x}\sim S_{\mathrm{ref}}}[|p^*(\mathbf{x})| > \tau^*] + \Delta$;
    $S \leftarrow S \setminus \{\mathbf{x} \in S : |p^*(\mathbf{x})| > \tau^*\}$;

The algorithm iteratively removes the points that give large values to polynomials that do not satisfy the stopping criterion of line 6. The hyperparameter R determines how selective the filtering is: larger values of R imply that less points will be removed in each iteration, and the proportion of removed points that are clean is smaller. The price one has to pay for larger choices of R is that the guarantee of part 2 holds for polynomials with smaller absolute expectation under the target distribution.

Our algorithm requires access to a set of reference samples from the target distribution $\mathcal{D}^*$ and uses them to restrict its attention to polynomials with the desired properties under $\mathcal{D}^*$. The stopping criterion ensures that upon completion all of these polynomials will have bounded expectations under the empirical distribution over the filtered set. The bound depends on the hyperparameters ϵ (target error) and m (effective size), but not on the degree ℓ of the polynomials considered.

We now give the proof of our main technical tool, Theorem 3.2. The proof follows the approach of [KSV24c], but has a number of technical differences. In particular, the algorithm of [KSV24c] only gives guarantees for non-negative polynomials and works with respect to the uniform distribution

on the hypercube. Here, we preserve the expectation of any polynomial, with respect to any target hypercontractive distribution that can be sampled efficiently.

Proof of Theorem 3.2. We first observe that the family $\mathcal{P}$ of polynomials p of degree at most ℓ for which $\mathbb{E}_{\mathbf{x}\sim S_{\text{ref}}}[|p(\mathbf{x})|] \le \gamma$ and $\mathbb{E}_{\mathbf{x}\sim S_{\text{ref}}}[(p(\mathbf{x}))^2] \le \beta$ can be described by $O(m_{\text{ref}})$ linear constraints plus one convex quadratic constraint over the coefficient vectors whose dimension is $(d+1)^\ell$. Therefore, lines 3 and 5 can be implemented as convex programs in time $\text{poly}(|S_{\text{inp}}|, m_{\text{ref}}, (d+1)^\ell)$. Lines 8 and 9 can be implemented with a single pass of the input points.

We first prove the following claim, which ensures that lines 8 and 9 are well defined.

Claim. *When $\lambda^* > \epsilon$, there exists $\tau^* \ge 0$ that satisfies the guarantees of line 8.*

Proof. Suppose, for contradiction, that for any $\tau \ge 0$ we have

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

We may now integrate both sides of the inequality over $\tau \in [0, B]$, since it holds for all $\tau \ge 0$. Note that $|p^*(\mathbf{x})| \ge 0$ for all $\mathbf{x}$ and, therefore, the following are true

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim S}[|p^*(\mathbf{x})|] = \int_{\tau=0}^{\infty} \mathop{\mathbb{P}}_{\mathbf{x}\sim S}[|p^*(\mathbf{x})| > \tau]\, d\tau = \int_{\tau=0}^{B} \mathop{\mathbb{P}}_{\mathbf{x}\sim S}[|p^*(\mathbf{x})| > \tau]\, d\tau \tag{3.1}$$

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\text{ref}}}[|p^*(\mathbf{x})|] = \int_{\tau=0}^{\infty} \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\text{ref}}}[|p^*(\mathbf{x})| > \tau]\, d\tau \ge \int_{\tau=0}^{B} \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\text{ref}}}[|p^*(\mathbf{x})| > \tau]\, d\tau \tag{3.2}$$

The second equality in Eq. (3.1) follows from the fact that for any $p^* \in \mathcal{P}$, S contains only points such that $p^*(\mathbf{x}) \in [-B, B]$, due to line 3. The inequality in Eq. (3.2) follows from the fact that the integrated function is non-negative. Overall, we obtain the following inequality

$$\lambda^* = \frac{|S|}{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim S}[p^*(\mathbf{x})] \le \frac{|S|}{m} \mathop{\mathbb{E}}_{\mathbf{x}\sim S}[|p^*(\mathbf{x})|] < R \cdot \mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\text{ref}}}[|p^*(\mathbf{x})|] + \Delta B \le \epsilon\,,$$

where the last inequality follows from the fact that $p^* \in \mathcal{P}$ and, hence, $\mathbb{E}_{\mathbf{x}\sim S_{\text{ref}}}[|p^*(\mathbf{x})|] \le \frac{\epsilon}{2R}$ and $\Delta = \frac{\epsilon}{2B}$. We have reached contradiction, because we showed that $\lambda^* \le \epsilon$. □

Note that due to the claim above, we have $\{\mathbf{x} \in S : |p^*(\mathbf{x})| > \tau^*\} \ne \emptyset$. This means that at each iteration we remove at least one point from the input dataset and, therefore, we do not need more than $M = |S_{\text{inp}}|$ iterations. The following claim ensures that we only remove a small fraction of clean points from S_{inp}.

Claim. *With probability at least $1 - \delta$ over S_{cln} and S_{ref}, we have*

$$|(S_{\text{cln}} \cap S_{\text{inp}}) \setminus S_{\text{filt}}| \le \frac{\epsilon m}{2} + \frac{1}{R} \cdot |S_{\text{inp}} \setminus S_{\text{filt}}|$$

Proof. We will show that with high probability over the clean and reference datasets, the number of removed points from S_{cln} is small. First, we account for the initial filtering of line 3. In this step, we remove points $\mathbf{x}$ from S_{inp} that give large absolute values to polynomials in $\mathcal{P}$, i.e., polynomials who, in particular, have bounded second norms over S_{ref}. Due to Lemma A.2 and since $m_{\text{ref}} \ge (CAd)^{2\ell}(\log \frac{1}{\delta})^{4\ell+1}$, the probability that some $\mathbf{x}$ drawn from $\mathcal{D}$ gives $|p(\mathbf{x})| > B$ for some $p \in \mathcal{P}$ is at most $\epsilon/4$. Therefore, the total number of points removed from S_{cln} in this step follows

the binomial distribution with m number of trials and probability of success at most $\epsilon/4$. By a standard Chernoff bound, and since $m \geq \frac{C}{\epsilon}\log(1/\delta)$, with probability at least $1-\delta/2$ we have

$$|\{\mathbf{x} \in S_{\mathrm{cln}} : |p(\mathbf{x})| > B \text{ for some } p \in \mathcal{P}\}| \leq \epsilon m/2 \tag{3.3}$$

It remains to account for the points removed in line 9. The removed points are always of the form $\{\mathbf{x} \in S : |p^*(\mathbf{x})| > \tau^*\}$ for some $p^* \in \mathcal{P}$. Moreover, according to line 8, we have

$$\frac{1}{m}\sum_{\mathbf{x}\in S} \mathbb{1}\{|p^*(\mathbf{x})| > \tau^*\} \geq R \cdot \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\mathrm{ref}}}[|p^*(\mathbf{x})| > \tau^*] + \Delta$$

Since $m, m_{\mathrm{ref}} \geq C' \frac{(d+1)^\ell + \log(1/\delta)}{(\Delta/R)^2}$ for some sufficiently large universal constant $C' \geq 1$ and the VC dimension of the class of polynomial threshold functions of degree ℓ is at most $(d+1)^\ell$, we have that with probability at least $1-\delta/2$ over S_{cln}, S_{ref} the following holds for all polynomials p of degree at most ℓ and all $\tau \geq 0$

$$R \cdot \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\mathrm{cln}}}[|p(\mathbf{x})| > \tau] \leq R \cdot \mathop{\mathbb{P}}_{\mathbf{x}\sim S_{\mathrm{ref}}}[|p(\mathbf{x})| > \tau] + \Delta\,,$$

since S_{cln} and S_{ref} are both i.i.d. samples from the distribution $\mathcal{D}^*$ and $\mathbb{1}\{|p(\mathbf{x})| > \tau\}$ is equal to the sum of two degree-ℓ polynomial threshold functions, $\mathbb{1}\{p(\mathbf{x}) > \tau\} + \mathbb{1}\{p(\mathbf{x}) < -\tau\}$. Therefore:

$$|\{\mathbf{x} \in S : |p^*(\mathbf{x})| > \tau^*\}| \geq R \cdot |\{\mathbf{x} \in S_{\mathrm{cln}} : |p^*(\mathbf{x})| > \tau^*|\,,$$

for any update we make. Summing over all the updates, we obtain that the total number of points removed from S_{cln} by the updates of line 9 is at most

$$(1/R) \cdot |S_{\mathrm{inp}} \setminus S_{\mathrm{filt}}| \tag{3.4}$$

By combining Eq. (3.3) and (3.4), we obtain the desired result, where the probability of failure is at most δ due to a union bound. □

So far, we have proven the first part of Theorem 3.2. We will now prove the second part, by showing the following claim.

Claim. *Any polynomial p of degree at most ℓ and $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|p(\mathbf{x})|] \leq \epsilon/(4R)$, lies within $\mathcal{P}$ with probability at least $1-\delta$ over S_{ref}.*

Proof. Let $\mathcal{D} = \mathcal{D}^*$. Fix any polynomial p with $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|p(\mathbf{x})|] \leq \epsilon/(4R)$. Due to the hypercontractivity of $\mathcal{D}$ (Definition 3.1), we have $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(p(\mathbf{x}))^2] \leq (2A)^{2\ell} = \beta/2$, since $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|p(\mathbf{x})|] \leq 1$. By Lemma A.1, since m_{ref} is large enough, w.p. at least $1-\frac{\delta}{2}$ over S_{ref}, we have $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{ref}}}[(p(\mathbf{x}))^2] \leq \beta$.

It remains to show that with probability at least $1-\delta/2$, we have $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{ref}}}[|p(\mathbf{x})|] \leq \gamma = \epsilon/(2R)$. The following bound can be obtained for any $t \geq 1$ by applying the Marcinkiewicz–Zygmund inequality (see [Fer14]), due to the fact that S_{ref} consists of m_{ref} i.i.d. examples.

$$\mathop{\mathbb{P}}_{S_{\mathrm{ref}}\sim\mathcal{D}^{m_{\mathrm{ref}}}}\Big[\Big|\mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\mathrm{ref}}}[|p(\mathbf{x})|] - \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[|p(\mathbf{x}')|]\Big| > \frac{\epsilon}{4R}\Big] \leq 2\Big(\frac{32R^2t}{\epsilon^2 m_{\mathrm{ref}}}\Big)^t \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[\Big||p(\mathbf{x})| - \mathop{\mathbb{E}}_{\mathbf{x}'\sim\mathcal{D}}[|p(\mathbf{x}')|]\Big|^{2t}\Big]$$

To conclude the proof of the claim, we use the simple inequality $(a+b)^{2t} \leq 4^t \max\{a^{2t}, b^{2t}\}$ for all $a, b, t \geq 0$ as well as hypercontractivity to bound the expectation on the right-hand side by $(8At)^{2\ell t}\,\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|p(\mathbf{x})|]^{2t}$. Due to the choice of m_{ref}, if we choose $t = \log(1/\delta)$, we obtain a bound of $\delta/2$, as desired. □

This concludes the proof of Theorem 3.2. □

4 Applications

4.1 Learning with Bounded Contamination

We give results for BC-learning of any concept class that can be approximated by low-degree polynomials in $\mathcal{L}_2$ distance.

Definition 4.1 (Polynomial Approximators)**.** For $\epsilon \in (0,1)$, we say that a class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ has ϵ-approximate degree $\ell = \ell(\epsilon)$ with respect to some distribution $\mathcal{D}^*$ over $\mathcal{X}$ if for any $f \in \mathcal{C}$ there is a polynomial p of degree at most $\ell(\epsilon)$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[(f(\mathbf{x}) - p(\mathbf{x}))^2] \leq \epsilon$.

Our main result additionally requires that the target marginal distribution is hypercontractive. This assumption is inherited by Theorem 3.2.

Theorem 4.2 (Polynomial Approximation implies BC-Learning)**.** *Let $\epsilon, \delta \in (0,1)$ and $A \geq 1$. Let $\mathcal{D}^*$ be some A-hypercontractive distribution over a d-dimensional space and let $\mathcal{C}$ be a concept class whose $\frac{\epsilon^4}{C}$-approximate degree w.r.t. $\mathcal{D}^*$ is ℓ for some large enough universal constant $C \geq 1$. Then, there is an algorithm that (ϵ,δ)-learns $\mathcal{C}$ under bounded contamination with respect to $\mathcal{D}^*$ in time* $\mathrm{poly}(A^\ell, (\log\frac{1}{\delta})^\ell, (d+1)^\ell, \frac{1}{\epsilon})$*, and has (clean) sample complexity at most* $\frac{1}{\epsilon^6}O(Ad\log(1/\delta))^{4\ell+1}$.

In Table 2, we summarize some of the new results we obtain as corollaries of Theorem 4.2, combined with appropriate known bounds on the approximate degree (see Appendix C.1). We also present the previous state-of-the-art results for comparison. For monotone functions and convex sets, no non-trivial results were known before this work. For halfspace intersections, we obtain exponential improvements over prior work, and for polynomial threshold functions, we obtain the first near-optimal error bounds. For small-depth circuits, we obtain an improved dependence on the depth.

Concept Class	Target Marginal	Runtime	Error	Reference
Intersections of k Halfspaces	$\mathcal{N}_d$ or Unif_d	$d^{\tilde{O}(\log(k)/\epsilon^8)}$	$2\eta+\epsilon$	This work
	$\mathcal{N}_d$ or Unif_d	$d^{\tilde{O}(k^6/\epsilon^4)}$	$4\eta+\epsilon$	[GSSV24]
	$\mathcal{N}_d$	$(\frac{dk}{\epsilon})^{O(1)} + (\frac{k}{\epsilon})^{O(k^2)}$	$\tilde{O}(k^{\frac{4}{11}}\eta^{\frac{1}{11}})+\epsilon$	[DKS18a]
Degree-k PTFs	$\mathcal{N}_d$	$d^{\tilde{O}(k^2/\epsilon^8)}$	$2\eta+\epsilon$	This work
	$\mathcal{N}_d$	$\mathrm{poly}(d^k, 1/\epsilon)$	$\tilde{O}(k^2\eta^{\frac{1}{1+k}})+\epsilon$	[DKS18a]
	Unif_d	$d^{(\log(1/\epsilon))^{\tilde{O}(k^2)}/\epsilon^8}$	$2\eta+\epsilon$	This work
Monotone Functions	Unif_d	$d^{\tilde{O}(\sqrt{d}/\epsilon^8)}$	$2\eta+\epsilon$	This work
Convex Sets	$\mathcal{N}_d$	$d^{\tilde{O}(\sqrt{d}/\epsilon^8)}$	$2\eta+\epsilon$	This work
Depth-t, Size-s Circuits	Unif_d	$d^{O(\log s)^{t-1}\log 1/\epsilon}$	$2\eta+\epsilon$	This work
	Unif_d	$d^{O(\log s)^{O(t)}\log 1/\epsilon}$	$2\eta+\epsilon$	[KSV24c]

Table 2: Bounds on the time complexity of learning with bounded contamination of (unknown) rate $\eta \in (0,1)$ up to failure probability $\delta = 0.01$.

Our results nearly match the best known upper bounds for agnostic learning, albeit with a worse dependence on the excess error parameter ϵ. In particular, to achieve excess error ϵ, we require $O(\epsilon^4)$-approximating polynomials, while for agnostic learning an $O(\epsilon^2)$-approximation suffices. Therefore,

our results imply, for example, a runtime of $d^{\tilde{O}(\log(k)/\epsilon^8}$ for BC-learning of k-halfspace intersections with respect to $\mathcal{N}_d$, but agnostic learning can be done in $d^{\tilde{O}(\log(k)/\epsilon^4}$ [KOS08].

The proof of Theorem 4.2 is based on a delicate analysis of the error on the filtered dataset obtained by applying Theorem 3.2. The filtering algorithm is used to preserve the following important property of any polynomial p^* that approximates a boolean function:

$$\mathbb{E}[(p^*(\mathbf{x}))^2] \leq 1 + O(\epsilon) \tag{4.1}$$

We show that obtaining a filtered set that preserves this property is sufficient for learning with bounded contamination. To prove this, we crucially use the fact that the noise is bounded and that the right hand side of Eq. (4.1) is approximately equal to 1 (rather than some larger constant). On the other hand, by applying part 2 of Theorem 3.2 on the polynomial $q(\mathbf{x}) = (p^*(\mathbf{x}))^2 - 1$, we are able to ensure the desired property for p^* on the filtered set. We give the full proof of our main result on learning with bounded contamination (Theorem 4.2) below.

Proof of Theorem 4.2. The algorithm receives a dataset $\bar{S}_{\text{inp}}$ generated by $(\mathcal{D}^*, f^*)$ with bounded contamination of rate $\eta \in (0,1)$, where $f^* \in \mathcal{C}$ and η, f^* are unknown, draws a set S_{ref} of $m_{\text{ref}} = \frac{(C'Ad)^{4\ell}}{\epsilon^6}(\log\frac{1}{\delta})^{8\ell+1}$ i.i.d. unlabeled examples from $\mathcal{D}^*$, where $C' \geq 1$ is a sufficiently large universal constant and does the following.

1. First, the algorithm runs the filtering procedure of Theorem 3.2 (that is, Algorithm 1) on input $(S_{\text{inp}}, S_{\text{ref}}, m = |S_{\text{inp}}|, 2\ell, R = 2, 24\epsilon^2/\sqrt{C})$ to form the filtered dataset $\bar{S}_{\text{filt}}$, where the labels are consistent with $\bar{S}_{\text{inp}}$.

2. Then, the algorithm finds a polynomial $\widehat{p}$ of degree at most ℓ that minimizes the following convex objective.

$$\widehat{p} = \arg\min_{p} \mathbb{E}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[|y - p(\mathbf{x})|]$$
$$\text{s.t. } p \text{ has degree at most } \ell$$

3. The algorithm outputs $h(\mathbf{x}) = \text{sign}(\widehat{p}(\mathbf{x}) + \widehat{\tau})$, where $\widehat{\tau} \in \mathbb{R}$ minimizes the one-dimensional objective $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq \text{sign}(\widehat{p}(\mathbf{x}) + \tau)]$ over $\tau \in \mathbb{R}$.

We will now bound the error of h on the clean dataset $\bar{S}_{\text{cln}}$ according to which $\bar{S}_{\text{inp}}$ was formed (see Definition 1.1). This suffices because, due to standard VC theory, and since h is a polynomial threshold function of degree at most ℓ, as long as $m = |\bar{S}_{\text{cln}}| \geq C'\frac{(d+1)^\ell + \log\frac{1}{\delta}}{\epsilon^2}$ for some sufficiently large universal constant $C' \geq 1$, we have that $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[f^*(\mathbf{x}) \neq h(\mathbf{x})]$ is approximately equal to its empirical counterpart $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq h(\mathbf{x})]$.

Consider p^* to be a polynomial of degree at most ℓ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(f^*(\mathbf{x}) - p^*(\mathbf{x}))^2] \leq \epsilon'$, where $\epsilon' = \epsilon^4/C$. Since p^* approximates a Boolean-valued function f^*, the values of p^* should, in expectation, be close to either 1 or -1. In particular, we obtain the following bound using the Cauchy-Schwarz inequality:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[|(p^*(\mathbf{x}))^2 - 1|] &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|(p^*(\mathbf{x}))^2 - (f^*(\mathbf{x}))^2|] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[|p^*(\mathbf{x}) - f^*(\mathbf{x})| \cdot |p^*(\mathbf{x}) + f^*(\mathbf{x})|] \\
&\leq \sqrt{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[(p^*(\mathbf{x}) - f^*(\mathbf{x}))^2]} \cdot \sqrt{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[(p^*(\mathbf{x}) + f^*(\mathbf{x}))^2]} \\
&\leq \epsilon' + \sqrt{2\epsilon'} \leq 3\epsilon^2/\sqrt{C}
\end{aligned} \tag{4.2}$$

Note that $q(\mathbf{x}) = (p^*(\mathbf{x}))^2 - 1$ is a polynomial of degree at most 2ℓ. Therefore, Algorithm 1 can be used to filter out the points $\mathbf{x}$ such that $q(\mathbf{x})$ is too large and, in particular, due to Theorem 3.2, the following holds with probability at least $1 - \delta/4$ over the random choice of S_{ref} and $\bar{S}_{\mathrm{cln}}$ as long as $m \geq C' \frac{(2A(d+1))^{4\ell}}{\epsilon^6} \log \frac{1}{\delta}$ and $m_{\mathrm{ref}} \geq \frac{(C'Ad)^{4\ell}}{\epsilon^6} (\log \frac{1}{\delta})^{8\ell+1}$:

$$\sum_{\mathbf{x} \in S_{\mathrm{filt}}} q(\mathbf{x}) \leq \frac{\epsilon^2 m}{C''}\,, \tag{4.3}$$

where $C'' = \sqrt{C}/24$. We will now show that this bound is sufficient for our purposes.

Consider the quantity $P = \frac{|\bar{S}_{\mathrm{filt}}|}{m} \mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{filt}}}[y \neq h(\mathbf{x})]$. We first bound the quantity P by $P \leq \frac{|\bar{S}_{\mathrm{filt}}|}{2m} \mathbb{E}_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{filt}}}[|y - \widehat{p}(\mathbf{x})|]$, where the factor $1/2$ appears due to the fact that the choice of $\widehat{\tau}$ is optimal (and a random choice would yield the factor $1/2$ with positive probability, see [KKMS08]). Furthermore, we have the following, due to the fact that $\widehat{p}$ is optimal among low-degree polynomials for $\bar{S}_{\mathrm{filt}}$ with respect to the absolute error.

$$\begin{aligned}
P &\leq \frac{1}{2m} \sum_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}} |y - \widehat{p}(\mathbf{x})| \leq \frac{1}{2m} \sum_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}} |y - p^*(\mathbf{x})| \\
&\leq \frac{1}{2m} \sum_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}\cap\bar{S}_{\mathrm{cln}}} |f^*(\mathbf{x}) - p^*(\mathbf{x})| + \frac{1}{2m} \sum_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}\setminus\bar{S}_{\mathrm{cln}}} |y - p^*(\mathbf{x})| \\
&\leq \frac{1}{2m} \sum_{(\mathbf{x},y)\in\bar{S}_{\mathrm{cln}}} |f^*(\mathbf{x}) - p^*(\mathbf{x})| + \frac{|\bar{S}_{\mathrm{filt}} \setminus \bar{S}_{\mathrm{cln}}|}{2m} + \frac{1}{2m} \sum_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}\setminus\bar{S}_{\mathrm{cln}}} |p^*(\mathbf{x})|
\end{aligned} \tag{4.4}$$

For the first term, we choose $m = |\bar{S}_{\mathrm{inp}}| = |\bar{S}_{\mathrm{cln}}|$ and we have $\frac{1}{2m} \sum_{(\mathbf{x},y)\in\bar{S}_{\mathrm{cln}}} |f^*(\mathbf{x}) - p^*(\mathbf{x})| = \frac{1}{2} \mathbb{E}_{\mathbf{x}\sim S_{\mathrm{cln}}}[|f^*(\mathbf{x}) - p^*(\mathbf{x})|]$. In order to bound this quantity, we use the fact that S_{cln} is a set of m i.i.d. samples from a hypercontractive distribution, combining the Marcinkiewicz-Zygmund inequality (which is a generalization of Chebyshev's inequality to higher moments), hypercontractivity and boundedness of f^* to obtain that, as long as $m \geq \frac{1}{\epsilon^2}(C'A)^{2\ell}(\log(1/\delta))^{2\ell+1}$, with probability at least $1 - \delta/4$ over the choice of S_{cln} we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\mathrm{cln}}}[|f^*(\mathbf{x}) - p^*(\mathbf{x})|] &\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}^*}[|f^*(\mathbf{x}) - p^*(\mathbf{x})|] + \epsilon/6 \\
&\leq \sqrt{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}^*}[(f^*(\mathbf{x}) - p^*(\mathbf{x}))^2]} + \epsilon/6 \\
&\leq \epsilon^2/\sqrt{C} + \epsilon/6 \leq \epsilon/3
\end{aligned} \tag{4.5}$$

For the last term in Eq. (4.4), we first use the basic fact that the variance of any random variable is non-negative and hence $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{filt}}\setminus S_{\mathrm{cln}}}[|p^*(\mathbf{x})|] \leq (\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{filt}}\setminus S_{\mathrm{cln}}}[(p^*(\mathbf{x}))^2])^{1/2}$. Then, we observe that $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{filt}}\setminus S_{\mathrm{cln}}}[(p^*(\mathbf{x}))^2] = \mathbb{E}_{\mathbf{x}\sim S_{\mathrm{filt}}\setminus S_{\mathrm{cln}}}[q(\mathbf{x})] + 1$. Overall, we have:

$$\frac{1}{2m} \sum_{\mathbf{x}\in S_{\mathrm{filt}}\setminus S_{\mathrm{cln}}} |p^*(\mathbf{x})| \leq \frac{|S_{\mathrm{filt}} \setminus S_{\mathrm{cln}}|}{2m} \Big(1 + \mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\mathrm{filt}}\setminus S_{\mathrm{cln}}}[q(\mathbf{x})]\Big)^{1/2} \tag{4.6}$$

We will now bound the quantity $P' = \mathbb{E}_{\mathbf{x}\sim S_{\text{filt}}\setminus S_{\text{cln}}}[q(\mathbf{x})] = \frac{1}{|S_{\text{filt}}\setminus S_{\text{cln}}|}\sum_{\mathbf{x}\in S_{\text{filt}}\setminus S_{\text{cln}}} q(\mathbf{x})$.

$$\begin{aligned} P' &= \frac{1}{|S_{\text{filt}}\setminus S_{\text{cln}}|}\sum_{\mathbf{x}\in S_{\text{filt}}} q(\mathbf{x}) - \frac{1}{|S_{\text{filt}}\setminus S_{\text{cln}}|}\sum_{\mathbf{x}\in S_{\text{filt}}\cap S_{\text{cln}}} q(\mathbf{x}) \\ &\overset{(4.3)}{\leq} \frac{\epsilon^2 m}{C''|S_{\text{filt}}\setminus S_{\text{cln}}|} + \frac{1}{|S_{\text{filt}}\setminus S_{\text{cln}}|}\sum_{\mathbf{x}\in S_{\text{filt}}\cap S_{\text{cln}}} |q(\mathbf{x})| \\ &\leq \frac{\epsilon^2 m}{C''|S_{\text{filt}}\setminus S_{\text{cln}}|} + \frac{m}{|S_{\text{filt}}\setminus S_{\text{cln}}|}\mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\text{cln}}}[|q(\mathbf{x})|] \end{aligned} \tag{4.7}$$

The first inequality in the expression above follows from the guarantee of Eq. (4.3), which is provided by the filtering algorithm. We may now use the Marcinkiewicz-Zygmund inequality once more to obtain that, as long as $m \geq \frac{1}{\epsilon^4}(C'A)^{4\ell}(\log(1/\delta))^{4\ell+1}$, with probability at least $1-\delta/4$, we have:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim S_{\text{cln}}}[|q(\mathbf{x})|] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim \mathcal{D}^*}[|q(\mathbf{x})|] + \epsilon^2/\sqrt{C} \leq \epsilon^2/C'' \,,$$

where the last inequality follows from Eq. (4.2), the fact that $q(\mathbf{x}) = (p^*(\mathbf{x}))^2 - 1$, and recall that $C'' = \sqrt{C}$. Overall, we obtain the bound $P' \leq \frac{2\epsilon^2 m}{C''|S_{\text{filt}}\setminus S_{\text{cln}}|}$. Note that since for any $a \geq 0$ we have $\sqrt{1+a} \leq 1 + \sqrt{a}$, by substituting the bound for P' in Eq. (4.6), we obtain

$$\frac{1}{2m}\sum_{\mathbf{x}\in S_{\text{filt}}\setminus S_{\text{cln}}} |p^*(\mathbf{x})| \leq \frac{|S_{\text{filt}}\setminus S_{\text{cln}}|}{2m} + \frac{\epsilon}{\sqrt{C''/2}} \,,$$

where we used the fact that $|S_{\text{filt}}\setminus S_{\text{cln}}| \leq m$ to bound the factor $\sqrt{|S_{\text{filt}}\setminus S_{\text{cln}}|/m}$ in the second term of the above bound by 1. In total, by choosing C appropriately large, we have the following bound, where we combine Eq. (4.4), (4.5), and (4.6).

$$P = \frac{1}{m}\sum_{(\mathbf{x},y)\in \bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} \leq \frac{|\bar{S}_{\text{filt}}\setminus \bar{S}_{\text{cln}}|}{m} + \frac{2\epsilon}{3} \,,$$

with probability at least $1 - 3\delta/4$ over the random choice of $\bar{S}_{\text{inp}}$ and S_{ref}.

We will now bound $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq h(\mathbf{x})]$ by $2\eta + 5\epsilon/6$. We have the following:

$$\begin{aligned} \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\bar{S}_{\text{cln}}}[y \neq h(\mathbf{x})] &= \frac{1}{m}\sum_{(\mathbf{x},y)\in\bar{S}_{\text{cln}}} \mathbb{1}\{y \neq h(\mathbf{x})\} \\ &= \frac{1}{m}\sum_{(\mathbf{x},y)\in\bar{S}_{\text{cln}}\cap\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} + \frac{1}{m}\sum_{(\mathbf{x},y)\in\bar{S}_{\text{cln}}\setminus\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} \\ &\leq \sum_{(\mathbf{x},y)\in\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} + \frac{|\bar{S}_{\text{cln}}\setminus\bar{S}_{\text{filt}}|}{m} \\ &\leq \frac{|\bar{S}_{\text{filt}}\setminus\bar{S}_{\text{cln}}| + |(\bar{S}_{\text{cln}}\cap\bar{S}_{\text{inp}})\setminus\bar{S}_{\text{filt}}| + |\bar{S}_{\text{cln}}\setminus\bar{S}_{\text{inp}}|}{m} + 2\epsilon/3 \end{aligned}$$

In order to complete the proof, we use the guarantee of Algorithm 1 that with probability at least $1-\delta/4$, we have

$$\begin{aligned} |(\bar{S}_{\text{cln}}\cap\bar{S}_{\text{inp}})\setminus\bar{S}_{\text{filt}}| &\leq \frac{1}{2}|\bar{S}_{\text{inp}}\setminus\bar{S}_{\text{filt}}| + \frac{\epsilon m}{12} \\ &= \frac{1}{2}|(\bar{S}_{\text{cln}}\cap\bar{S}_{\text{inp}})\setminus\bar{S}_{\text{filt}}| + \frac{1}{2}|(\bar{S}_{\text{inp}}\setminus\bar{S}_{\text{cln}})\setminus\bar{S}_{\text{filt}}| + \epsilon m/12 \,. \end{aligned}$$

Therefore, we have $|(\bar{S}_{\text{cln}} \cap \bar{S}_{\text{inp}}) \setminus \bar{S}_{\text{filt}}| \leq |(\bar{S}_{\text{inp}} \setminus \bar{S}_{\text{cln}}) \setminus \bar{S}_{\text{filt}}| + \epsilon m/6$. Note that the union of the sets $(\bar{S}_{\text{inp}} \setminus \bar{S}_{\text{cln}}) \setminus \bar{S}_{\text{filt}}$ and $\bar{S}_{\text{filt}} \setminus \bar{S}_{\text{cln}}$ is disjoint and equals to the set of adversarial examples $\bar{S}_{\text{adv}} = \bar{S}_{\text{inp}} \setminus \bar{S}_{\text{cln}}$. By the definition of the noise model (Definition 1.1), we have that $|\bar{S}_{\text{adv}}| \leq \eta m$ and $|\bar{S}_{\text{cln}} \setminus \bar{S}_{\text{inp}}| = |\bar{S}_{\text{adv}}| \leq \eta m$. Therefore, we overall obtain the desired bound of

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{cln}}}[y \neq h(\mathbf{x})] \leq 2\eta + \frac{5\epsilon}{6}\,.$$

This concludes the proof of Theorem 4.2. □

4.2 Learning with Heavy Contamination

For heavy contamination, our results are based on the stronger notion of sandwiching approximators from pseudorandomness [Baz09].

Definition 4.3 (Sandwiching Approximators)**.** For $\epsilon \in (0,1)$, we say that a class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ has ϵ-sandwiching degree $\ell = \ell(\epsilon)$ with respect to some distribution $\mathcal{D}^*$ over $\mathcal{X}$ if for any $f \in \mathcal{C}$ there are two polynomials $p_{\text{up}}, p_{\text{down}}$ of degree at most $\ell(\epsilon)$ such that:

1. $p_{\text{down}}(\mathbf{x}) \leq f(\mathbf{x}) \leq p_{\text{up}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and
2. $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[p_{\text{up}}(\mathbf{x}) - p_{\text{down}}(\mathbf{x})] \leq \epsilon$

Once more, our main result requires hypercontractivity of the marginal distribution.

Theorem 4.4 (Sandwiching implies HC-Learning)**.** *Let $\epsilon, \delta \in (0,1)$ and $A, Q \geq 1$. Let $\mathcal{D}^*$ be some A-hypercontractive distribution over a d-dimensional space and let $\mathcal{C}$ be a concept class whose $\frac{\epsilon^2}{CQ}$-sandwiching degree w.r.t. $\mathcal{D}^*$ is ℓ for some large enough constant $C \geq 1$. Then, there is an algorithm that (ϵ, δ, Q)-HC learns $\mathcal{C}$ with respect to $\mathcal{D}^*$ in time $\text{poly}(A^\ell, (\log(1/\delta))^\ell, (d+1)^\ell, Q/\epsilon)$, and has (clean) sample complexity at most $\frac{Q^2}{\epsilon^5} \cdot O(Ad)^{2\ell} \cdot \log\frac{1}{\delta}$.*

In Table 3, we provide a number of end-to-end results for learning with heavy contamination, which are obtained by combining Theorem 4.4 with appropriate known bounds on the sandwiching degree (see Appendix C.2). The first row of Table 3 gives the bound $d^{\tilde{O}(Q^2k^6/\epsilon^4)}$ for learning intersections of k halfspaces with Q-heavy contamination by choosing $s = t = k$, as well as the bound $d^{\tilde{O}(Q^2k^44^k/\epsilon^4)}$ for arbitrary functions of k halfspaces, with the choice $s = 2^k$, $t = k$.

Concept Class	Target Marginal	Runtime
Depth-t, Size-s Decision Trees of Halfspaces	$\mathcal{N}_d$ or Unif_d	$d^{\tilde{O}(Q^2t^4s^2/\epsilon^4)}$
Depth-t, Size-s Boolean Circuits	Unif_d	$d^{O(\log s)^{O(t)}\log(Q/\epsilon)}$
Degree-k PTFs	$\mathcal{N}_d$	$d^{\tilde{O}_k((Q/\epsilon^2)^{4k\cdot 7^k})}$
Degree-2 PTFs	Unif_d	$d^{\tilde{O}(Q^9/\epsilon^{18})}$

Table 3: Bounds on the time complexity of learning with Q-heavy contamination up to error $Q \cdot \mathsf{opt}_{\text{total}} + \epsilon$ and failure probability $\delta = 0.01$.

The proof of Theorem 4.4 uses once more Algorithm 1 to filter the input set, but this time we set the hyperparameter R to a value that scales with the heavy contamination ratio Q. This ensures that we keep most of the clean points, because the filtering process is highly selective. Due to the choice of R, the required sandwiching degree scales with Q as well. As a consequence, for many classes, our bounds are exponential in Q. After filtering, the approximation property of the sandwiching polynomials is preserved under the empirical distribution over the filtered set, and this is sufficient for efficient learnability of the optimum hypothesis on the filtered set, due to Theorem A.3 by [KKMS08]. We give the full proof of our main result on learning with heavy contamination (Theorem 4.4) below.

Proof of Theorem 4.4. The algorithm receives a Q-heavily contaminated dataset $\bar{S}_{\mathrm{inp}}$, draws a set S_{ref} of $m_{\mathrm{ref}} = \frac{Q^2(C'Ad)^{2\ell}}{\epsilon^5}(\log\frac{1}{\delta})^{4\ell+1}$ i.i.d. unlabeled examples from $\mathcal{D} = \mathcal{D}^*$, where $C' \geq 1$ is a sufficiently large universal constant and does the following.

1. First, the algorithm runs the filtering procedure of Theorem 3.2 (that is, Algorithm 1) on input $(S_{\mathrm{inp}}, S_{\mathrm{ref}}, m = \frac{Q^2(C'Ad)^{2\ell}\log\frac{1}{\delta}}{\epsilon^5}, \ell, R = \frac{2Q}{\epsilon}, \epsilon/3)$ to form the filtered dataset $\bar{S}_{\mathrm{filt}}$.
2. Then, the algorithm finds a polynomial $\widehat{p}$ of degree at most ℓ that minimizes the following convex objective.
$$\widehat{p} = \arg\min_{p} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{filt}}}[|y - p(\mathbf{x})|]$$
$$\text{s.t. } p \text{ has degree at most } \ell$$
3. The algorithm outputs $h(\mathbf{x}) = \mathrm{sign}(\widehat{p}(\mathbf{x}) + \widehat{\tau})$, where $\widehat{\tau} \in \mathbb{R}$ minimizes the one-dimensional objective $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{filt}}}[y \neq \mathrm{sign}(\widehat{p}(\mathbf{x}) + \tau)]$ over $\tau \in \mathbb{R}$.

We will now bound the error of h on the clean distribution $\bar{\mathcal{D}}$. To this end, we first note that the empirical error of h on the clean samples $\bar{S}_{\mathrm{cln}}$ is, with high probability, close to the error on $\bar{\mathcal{D}}$, since $|\bar{S}_{\mathrm{cln}}|$ is large enough and h is a polynomial threshold function of degree at most ℓ. It therefore, suffices to bound the error of h on $\bar{S}_{\mathrm{cln}}$. We have the following.

$$\begin{aligned}\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{cln}}}[y \neq h(\mathbf{x})] &= \frac{1}{m}\sum_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{cln}}\cap\bar{S}_{\mathrm{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} + \frac{1}{m}\sum_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{cln}}\setminus\bar{S}_{\mathrm{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} \\ &\leq \frac{1}{m}\sum_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} + \frac{1}{m}\cdot|\bar{S}_{\mathrm{cln}} \setminus \bar{S}_{\mathrm{filt}}| \qquad (4.8)\end{aligned}$$

The second term in the bound above can be bounded by $\epsilon/3$, due to part 1 of Theorem 3.2. It remains to bound the first term.

Consider $f^* \in \mathcal{C}$ to be the concept in class $\mathcal{C}$ with minimum error on $\bar{S}_{\mathrm{filt}}$, i.e., we have that $f^* = \arg\min_{f\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}}[y \neq f(\mathbf{x})]$. Note that there is some $f' \in \mathcal{C}$ that makes at most $M \cdot \mathsf{opt}_{\mathrm{total}}$ errors on the whole input dataset $\bar{S}_{\mathrm{inp}}$, due to the definition of $\mathsf{opt}_{\mathrm{total}}$ (see Definition 1.4). We have that $\mathbb{P}_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}}[y \neq f^*(\mathbf{x})] \leq \mathbb{P}_{(\mathbf{x},y)\in\bar{S}_{\mathrm{filt}}}[y \neq f'(\mathbf{x})] \leq M \cdot \mathsf{opt}_{\mathrm{total}}/|\bar{S}_{\mathrm{filt}}|$.

Moreover, let $p_{\mathrm{up}}, p_{\mathrm{down}}$ be the $\frac{\epsilon^2}{72Q}$-sandwiching polynomials for f^* and $p = p_{\mathrm{up}} - p_{\mathrm{down}}$, which is a non-negative polynomial with $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|p(\mathbf{x})|] \leq \frac{\epsilon^2}{72Q}$. Due to part 2 of Theorem 3.2, the expectation of p under S_{filt} is at most $\frac{\epsilon m}{3|\bar{S}_{\mathrm{filt}}|}$, i.e., $\mathbb{E}_{\mathbf{x}\sim S_{\mathrm{filt}}}[p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})] \leq \frac{\epsilon m}{3|\bar{S}_{\mathrm{filt}}|}$. Since $p_{\mathrm{up}} \geq f^* \geq p_{\mathrm{down}}$, we also have that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\bar{S}_{\mathrm{filt}}}[|f^*(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x})|] \leq \frac{\epsilon m}{3|\bar{S}_{\mathrm{filt}}|}$$

We now observe that h is the output of the $\mathcal{L}_1$ polynomial regression algorithm on $\bar{S}_{\text{filt}}$ and, therefore, according to Theorem A.3 by [KKMS08], we overall have the following bound

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq h(\mathbf{x})] \leq \frac{M}{|\bar{S}_{\text{filt}}|} \cdot \mathsf{opt}_{\text{total}} + \frac{\epsilon m}{3|\bar{S}_{\text{filt}}|},$$

which implies the desired bound:

$$\frac{1}{m} \sum_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} \leq \frac{M}{m} \cdot \mathsf{opt}_{\text{total}} + \frac{\epsilon m}{2} \leq Q \cdot \mathsf{opt}_{\text{total}} + \epsilon/3 \tag{4.9}$$

The result follows by combining Eq. (4.8) and (4.9). □

Remark 4.5. Our result on HC-learning highlights the following analogy:

- Learning with heavy contamination can be thought of as a version of testable agnostic learning (Definition 4.7), where the algorithm is asked to find a subset of the input that is structured enough, instead of merely verifying whether the whole input is structured.
- Similarly, in the context of learning with distribution shift, PQ learning [GKKM20] requires finding a subset of an unlabeled test dataset where the learner is confident in its own predictions, while TDS learning [KSV24b] aims to verify whether the whole unlabeled test dataset is drawn from some distribution that is similar to the one the learner has trained on.
- In [GSSV24], it is shown that by using a spectral iterative filtering algorithm, TDS learning results can be extended to PQ learning. Here, we complete the analogy by showing that our iterative filtering algorithm can extend known results from testable learning (i.e., that sandwiching is sufficient, see [GKK23]) to HC-learning.

Remark 4.6. One could define a hybrid model where a bounded fraction of clean points are replaced adversarially and a proportionally large number of adversarial points are then added. Our results should apply to this setting as well, but we focus on BC and HC learning separately for simplicity of presentation.

4.3 Tolerant Testable Learning

We give new results for testable learning [RV23] and tolerant testable learning [GSSV24].

Definition 4.7 (Tolerant Testable Learning [GSSV24])**.** *An algorithm $\mathcal{A}$ is a tolerant tester-learner for $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ with respect to some target distribution $\mathcal{D}^*$ over $\mathcal{X}$ if on input $(\epsilon, \delta, \tau, \bar{S}_{\text{inp}})$, where $\epsilon, \tau, \delta \in (0,1)$ and $\bar{S}_{\text{inp}}$ is a set of i.i.d. examples from some arbitrary distribution $\bar{\mathcal{D}}$, the algorithm $\mathcal{A}$ outputs either outputs* Reject *or outputs* (Accept, h)*, where $h : \mathcal{X} \to \{\pm 1\}$ such that with probability at least $1 - \delta$ over $\bar{S}_{\text{inp}}$, and the randomness of $\mathcal{A}$, the following conditions hold.*

1. *(Soundness) Upon acceptance, the error of h is bounded as follows:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq h(\mathbf{x})] \leq \min_{f\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})] + \tau + \epsilon$$

2. *(Completeness) If $\mathrm{d}_{\text{TV}}(\mathcal{D}, \mathcal{D}^*) \leq \tau$, where $\mathcal{D}$ is the marginal of $\bar{\mathcal{D}}$ on $\mathcal{X}$, then $\mathcal{A}$ accepts.*

The sample complexity of $\mathcal{A}$ is the minimum number of examples required to achieve the above guarantee.

Previous work by [GSSV24] showed that the existence of $\mathcal{L}_2$ sandwiching polynomials with bounded coefficients, i.e., polynomials $p_{\mathrm{up}}, p_{\mathrm{down}}$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[(p_{\mathrm{up}}(\mathbf{x}) - p_{\mathrm{down}}(\mathbf{x}))^2] \leq \epsilon$ implies tolerant testable learning. Here, we relax this condition to $\mathcal{L}_1$ sandwiching polynomials (Definition 4.3), without requiring any bound on the coefficients. Note that even for non-tolerant testable learning, where $\mathcal{L}_1$ sandwiching is known to suffice [GKK23], all prior work required bounds on the coefficients.

Theorem 4.8 (Sandwiching implies Tolerant Testable Learning)**.** *Let $\epsilon, \delta, \tau \in (0,1)$ and $Q \geq 1$. Let $\mathcal{D}^*$ be some A-hypercontractive distribution over a d-dimensional space and let $\mathcal{C}$ be a concept class whose $\frac{\epsilon}{C(1+\tau/\epsilon)}$-sandwiching degree w.r.t. $\mathcal{D}^*$ is ℓ for some large enough constant $C \geq 1$. Then, there is an (ϵ, δ, τ)-tester-learner for $\mathcal{C}$ w.r.t. $\mathcal{D}^*$ with runtime $\mathrm{poly}(A^\ell, (\log(1/\delta))^\ell, (d+1)^\ell, 1/\epsilon)$, and sample complexity at most $\frac{1}{\epsilon^5} \cdot O(Ad)^{2\ell} \cdot \log\frac{1}{\delta}$.*

The following corollary follows from Theorem 4.8, combined with the fact that log-concave distributions are hypercontractive (see [Bob01]), as well as the results of [KM13, KKM13] on the sandwiching degree of functions of halfspaces (see Appendix C.2). This is the first result for testably learning even a single halfspace with respect to log-concave distributions up to optimal error.

Corollary 4.9. *Let $\mathcal{C}$ be the class of arbitrary functions of k halfspaces over $\mathbb{R}^d$ and let $\mathcal{D}^*$ be any log-concave distribution. There is an (ϵ, δ, τ)-tester-learner for $\mathcal{C}$ with respect to $\mathcal{D}^*$ that runs in time $(d\log(1/\delta))^{\tilde{O}(\ell)}$, where $\ell = \exp((\log(\log(k)/\epsilon))^{O(k)}/\epsilon^8)$.*

We will now prove Theorem 4.8, using once more the iterative polynomial filtering algorithm of Theorem 3.2.

Proof of Theorem 4.8. The algorithm receives a dataset $\bar{S}_{\mathrm{inp}}$, draws a reference set S_{ref} of $m_{\mathrm{ref}} = \frac{(C'Ad)^{2\ell}}{\epsilon^5}(\log\frac{1}{\delta})^{4\ell+1}$ i.i.d. unlabeled examples from $\mathcal{D}$, where $C' \geq 1$ is a sufficiently large universal constant and does the following.

1. First, the algorithm runs the filtering procedure of Theorem 3.2 (that is, Algorithm 1) on input $(S_{\mathrm{inp}}, S_{\mathrm{ref}}, m = |S_{\mathrm{inp}}|, \ell, R = \frac{4\tau}{\epsilon} + 2, \epsilon/8)$ to form the filtered dataset $\bar{S}_{\mathrm{filt}}$.
2. Then, the algorithm checks if $m - |\bar{S}_{\mathrm{filt}}| \leq (\tau + \epsilon/2)m$ and rejects if the inequality does not hold.
3. Otherwise, the algorithm finds a polynomial $\widehat{p}$ of degree at most ℓ that minimizes the following convex objective.
$$\widehat{p} = \arg\min_{p} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{filt}}} [|y - p(\mathbf{x})|]$$
$$\text{s.t. } p \text{ has degree at most } \ell$$
4. The algorithm outputs $h(\mathbf{x}) = \mathrm{sign}(\widehat{p}(\mathbf{x}) + \widehat{\tau})$, where $\widehat{\tau} \in \mathbb{R}$ minimizes the one-dimensional objective $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\mathrm{filt}}}[y \neq \mathrm{sign}(\widehat{p}(\mathbf{x}) + \tau)]$ over $\tau \in \mathbb{R}$.

Soundness. Suppose, first, that the algorithm accepts. Observe that $|\bar{S}_{\mathrm{inp}} \setminus \bar{S}_{\mathrm{filt}}| = |\bar{S}_{\mathrm{inp}}| - |\bar{S}_{\mathrm{filt}}|$, since $\bar{S}_{\mathrm{filt}} \subseteq \bar{S}_{\mathrm{inp}}$. Since the algorithm has accepted, we know the following.

$$|\bar{S}_{\mathrm{inp}}| - |\bar{S}_{\mathrm{filt}}| = m - |\bar{S}_{\mathrm{filt}}| \leq m(\tau + \epsilon/2)$$

Recall that $\bar{S}_{\text{inp}}$ is a set of m i.i.d. examples from the input distribution $\bar{\mathcal{D}}$. Since m is large enough, due to standard VC dimension arguments (and the fact that h is a polynomial threshold function of degree at most ℓ), we have that with high probability:

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq h(\mathbf{x})] \leq \mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq h(\mathbf{x})] + \epsilon/4 \tag{4.10}$$

Therefore, it suffices to show a bound on the empirical error of h on $\bar{S}_{\text{inp}}$. We have the following:

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq h(\mathbf{x})] &= \frac{1}{m}\sum_{(\mathbf{x},y)\in\bar{S}_{\text{inp}}\setminus\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} + \frac{1}{m}\sum_{(\mathbf{x},y)\in\bar{S}_{\text{filt}}} \mathbb{1}\{y \neq h(\mathbf{x})\} \\
&\leq \frac{1}{m}|\bar{S}_{\text{inp}} \setminus \bar{S}_{\text{filt}}| + \frac{|\bar{S}_{\text{filt}}|}{m}\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq h(\mathbf{x})] \\
&\leq \tau + \frac{\epsilon}{2} + \frac{|\bar{S}_{\text{filt}}|}{m}\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq h(\mathbf{x})]
\end{aligned}$$

It suffices to show that $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq h(\mathbf{x})] \leq \frac{m}{|\bar{S}_{\text{filt}}|}(\min_{f\in\mathcal{C}}\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq h(\mathbf{x})] + \epsilon/4)$. Let f^* be the function that minimizes the error on $\bar{\mathcal{D}}$ and let $p_{\text{up}}, p_{\text{down}}$ be its $\frac{\epsilon^2}{128\tau+64\epsilon}$-sandwiching polynomials. Then according to part 2 of Theorem 3.2, and since for $p = p_{\text{up}} - p_{\text{down}}$ we have $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[p(\mathbf{x})] \leq \frac{\epsilon^2}{128\tau+64\epsilon}$, we obtain:

$$|S_{\text{filt}}|\mathop{\mathbb{E}}_{\mathbf{x}\sim\bar{S}_{\text{filt}}}[p_{\text{up}}(\mathbf{x}) - p_{\text{down}}(\mathbf{x})] = \sum_{\mathbf{x}\in S_{\text{filt}}}(p_{\text{up}}(\mathbf{x}) - p_{\text{down}}(\mathbf{x})) \leq \frac{\epsilon m}{8}$$

Therefore, we also have $\mathbb{E}_{\mathbf{x}\sim S_{\text{filt}}}[|f^*(\mathbf{x}) - p_{\text{down}}(\mathbf{x})|] \leq \mathbb{E}_{\mathbf{x}\sim S_{\text{filt}}}[p_{\text{up}}(\mathbf{x}) - p_{\text{down}}(\mathbf{x})] \leq \frac{\epsilon m}{8|S_{\text{filt}}|}$. Observe, now that h is the output of the low-degree polynomial regression algorithm and, hence, according to Theorem A.3 by [KKMS08], we have the following.

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq h(\mathbf{x})] \leq \mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq f^*(\mathbf{x})] + \frac{\epsilon m}{8|\bar{S}_{\text{filt}}|}$$

To conclude the argument, observe $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{filt}}}[y \neq f^*(\mathbf{x})] \leq \frac{m}{|\bar{S}_{\text{filt}}|}\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq f^*(\mathbf{x})]$, since $\bar{S}_{\text{filt}} \subseteq \bar{S}_{\text{inp}}$ and, due to a Hoeffding bound, with high probability we have $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq f^*(\mathbf{x})] \leq \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f^*(\mathbf{x})] + \epsilon/8$. Note that due to the choice of f^*, we have $\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f^*(\mathbf{x})] = \min_{f\in\mathcal{C}}\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})] =: \mathsf{opt}$. Overall, we have the following bound:

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq h(\mathbf{x})] \leq \mathsf{opt} + \tau + 3\epsilon/4\,,$$

which, together with the generalization bound of Eq. (4.10) implies the soundness of the algorithm.

Completeness. We will now show that the algorithm accepts with high probability whenever $\mathrm{d_{TV}}(\mathcal{D}, \mathcal{D}^*) \leq \tau$. It suffices to show that $|\bar{S}_{\text{inp}} \setminus \bar{S}_{\text{filt}}| \leq (\tau + \epsilon/2)m$ with high probability, since $|\bar{S}_{\text{inp}} \setminus \bar{S}_{\text{filt}}| = m - |\bar{S}_{\text{filt}}|$, which is exactly the quantity based upon which the algorithm decides whether to reject or not.

Let $S_{\text{cln}} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\}$ be a set of m i.i.d. examples drawn from $\mathcal{D}$. Let S_{inp} be drawn as follows:

- For each $i \in [m]$, draw an independent Bernoulli random variable ξ_i with success rate $1 - \tau$.

- If $\xi_i = 1$, then let $\mathbf{z}^{(i)} = \mathbf{x}^{(i)}$.
- If $\xi_i = 0$, then draw $\mathbf{z}^{(i)}$ independently from the residual distribution $\mathcal{D}'$ in the maximal coupling between $\mathcal{D}$ and $\mathcal{D}^*$ (i.e., the joint distribution $(\mathbf{x}, \mathbf{x}^*)$ that has marginals $\mathcal{D}$ and $\mathcal{D}^*$ respectively and maximizes the likelihood that $\mathbf{x} = \mathbf{x}^*$).
- Let $S_{\text{cln}} = \{\mathbf{z}^{(i)}\}_{i \in [m]}$.

Note that the distribution of each $\mathbf{z}^{(i)}$ is $\mathcal{D}^*$ and they are drawn independently. Due to a Chernoff bound, since m is large enough, we have that, with high probability, $|S_{\text{inp}} \setminus S_{\text{cln}}| \leq m(\tau + \epsilon/8)$. According to Theorem 3.2, we have the following, with high probability:

$$|(S_{\text{inp}} \cap S_{\text{cln}}) \setminus S_{\text{filt}}| \leq \frac{1}{R} \cdot |S_{\text{inp}} \setminus S_{\text{filt}}| + \frac{\epsilon m}{8}$$

Observe, now, that $|S_{\text{inp}} \setminus S_{\text{filt}}| = |(S_{\text{inp}} \cap S_{\text{cln}}) \setminus S_{\text{filt}}| + |S_{\text{inp}} \setminus S_{\text{cln}}|$. Therefore, if we solve for $|S_{\text{inp}} \setminus S_{\text{filt}}|$, we have the following

$$\left(1 - \frac{1}{R}\right) \cdot |S_{\text{inp}} \setminus S_{\text{filt}}| \leq |S_{\text{inp}} \setminus S_{\text{cln}}| + \frac{\epsilon m}{8} \leq m(\tau + \epsilon/4)$$

Due to the choice of R, we have the desired bound $|\bar{S}_{\text{inp}} \setminus \bar{S}_{\text{filt}}| \leq m(\tau + \epsilon/2)$. □

5 Lower Bounds for Learning with Contamination

5.1 Bounded Contamination

Although the problem of learning from contaminated datasets is, in principle, more challenging than learning with label noise, it is not clear whether the common formulations of the former are formally stronger than those of the latter. In particular, the problem of learning with label noise is usually formalized in terms of agnostic learning, where the goal is to achieve error $\mathsf{opt} + \epsilon$ on the *input* distribution $\bar{\mathcal{D}}$ for $\mathsf{opt} = \min_{f \in \mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})]$, where $\mathcal{C}$ is the learned class. In contrast, learning with contamination is usually formulated as learning with nasty noise, where the goal is to achieve error $2\eta + \epsilon$ on the *target* distribution $(\mathcal{D}^*, f^*)$, where $f^* \in \mathcal{C}$ and η is the noise rate. Therefore, given an algorithm for learning with contamination, one may obtain an agnostic learner with error guarantee $3\mathsf{opt} + \epsilon$, but it is not clear whether the guarantee of $\mathsf{opt} + \epsilon$ can be achieved in a black-box way. This is because in one formulation the error is measured with respect to the input distribution and in the other it is measured with respect to the clean target distribution.

Furthermore, many of the lower bounds for agnostic learning seem to exclusively rule out algorithms with error guarantees of $\mathsf{opt} + \epsilon$ [DKPZ21]. Therefore, these lower bounds do not transfer directly to learning with bounded contamination as defined in Definition 1.1. One natural way to go around this, is to observe that our result in Theorem 4.2 would also hold for the following stronger version of Definition 1.1.

Definition 5.1 (Agnostic BC-Learning)**.** *An algorithm $\mathcal{A}$ is an agnostic BC-learner for $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ if on input $(\epsilon, \delta, \bar{S}_{\text{inp}})$, where $\epsilon, \delta \in (0,1)$, and $\bar{S}_{\text{inp}}$ is generated by $\bar{\mathcal{D}}$ with bounded contamination η for some labeled distribution $\bar{\mathcal{D}}$, and $\eta \in [0,1)$, the algorithm $\mathcal{A}$ outputs some hypothesis $h : \mathcal{X} \to \{\pm 1\}$ such that with probability at least $1 - \delta$ over the clean examples in $\bar{S}_{\text{inp}}$, and the randomness of $\mathcal{A}$:*

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq h(\mathbf{x})] \leq 2\eta + \mathsf{opt}_{\text{clean}} + \epsilon\,, \text{ where } \mathsf{opt}_{\text{clean}} = \min_{f \in \mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})]$$

The sample complexity of $\mathcal{A}$ is the minimum number of examples required to achieve the above guarantee. Moreover, a distribution-specific BC-learner with respect to some distribution $\mathcal{D}^*$ over $\mathcal{X}$ is a BC-learner that is guaranteed to work only when $\mathcal{D} = \mathcal{D}^*$.

Note that Definition 5.1 is a generalization of agnostic learning, since one can obtain agnostic learning by setting $\eta = 0$. Our Theorem 4.2 holds even under this definition, and all the lower bounds for agnostic learning (Table 4) are also inherited. For example, our results indicate that a runtime of $d^{\mathrm{poly}(\frac{1}{\epsilon}, \log k)}$ is both sufficient and necessary for learning intersections of k halfspaces in the agnostic BC setting. However, agnostic learning is known to be characterized by the $\mathcal{L}_1$-approximate degree [KKMS08, DSFT+14, DKPZ21], whereas we require the stronger assumption of low $\mathcal{L}_2$-approximate degree. Therefore, while for most interesting classes our results are essentially tight, a complete characterization of efficient learnability in the setting of Definition 5.1 remains open.

5.2 Heavy Contamination

We provide an information-theoretic lower bound for learning monotone functions with heavy contamination. Our lower bound highlights a separation between bounded and heavy contamination and justifies the need for a stronger notion of approximation than standard polynomial approximators. In particular, monotone functions over $\{\pm 1\}^d$ are known to admit ϵ-approximate polynomials of degree $\ell(\epsilon) = \sqrt{d}/\epsilon$, and hence admit a BC-learner that runs in time $2^{O(\sqrt{d})}$ for $\epsilon = 1/4$, but we show that every HC learner requires $2^{\Omega(d)}$ samples (and runtime).

Theorem 5.2 (Lower Bound for HC-Learning of Monotone Functions)**.** *Let $\mathcal{C}$ be the class of monotone functions over $\{\pm 1\}^d$. Any HC-learner for $\mathcal{C}$ with respect to the uniform distribution over the hypercube* $\mathrm{Unif}(\{\pm 1\}^d)$ *requires sample complexity $2^{\Omega(d)}$, even when $\epsilon = 1/4$, $Q = 2$ and $\delta = 1/3$.*

Proof of Theorem 5.2. Suppose that the ground truth is a constant function, i.e., either $f^* \equiv 1$ or $f^* \equiv -1$, and that $m \leq 2^{d/10}/10$. Then, the adversary receives the m clean examples $\bar{S}_{\mathrm{inp}}$ and draws m additional i.i.d. examples S' from $\mathrm{Unif}(\{\pm 1\}^d)$, labeling them according to $-f^*$. The input dataset is $\bar{S}_{\mathrm{inp}} = \bar{S}_{\mathrm{cln}} \cup \bar{S}'$.

Note, first, that no algorithm can achieve error better than $1/2$ with probability at least $1/2$, since guessing f^* randomly is an optimal strategy because the distribution of the input dataset $\bar{S}_{\mathrm{inp}}$ is exactly the same in the two cases corresponding to $f^* \equiv -1$ and $f^* \equiv 1$. We will show that $Q \cdot \mathsf{opt}_{\mathrm{total}} = 0$ with high probability over the choice of $\bar{S}_{\mathrm{inp}}$.

Let $\mathbf{x}, \mathbf{x}'$ be two independent samples from $\mathrm{Unif}(\{\pm 1\}^d)$. We say that $\mathbf{x}$ and $\mathbf{x}'$ are comparable if $\mathbf{x} \leq \mathbf{x}'$ or $\mathbf{x}' \leq \mathbf{x}$, i.e., if $\mathbf{x}(i) \leq \mathbf{x}'(i)$ for all i or $\mathbf{x}'(i) \leq \mathbf{x}(i)$ for all i. Since the coordinates are independent, we have

$$\mathop{\mathbb{P}}_{\mathbf{x},\mathbf{x}'}[\mathbf{x}(i) \leq \mathbf{x}'(i)\,, \text{ for all } i \in [d]] = \prod_{i \in [d]} \mathop{\mathbb{P}}_{\mathbf{x},\mathbf{x}'}[\mathbf{x}(i) \leq \mathbf{x}'(i)] = (3/4)^d$$

Similarly, for the other direction we have that $\mathbb{P}_{\mathbf{x},\mathbf{x}'}[\mathbf{x}(i) \geq \mathbf{x}'(i)\,, \text{ for all } i \in [d]] \leq (3/4)^d$ and, overall, that $\mathbb{P}_{\mathbf{x},\mathbf{x}'}[\mathbf{x}, \mathbf{x}' \text{ are comparable}] \leq 2 \cdot (3/4)^d$.

We have a set $\bar{S}_{\mathrm{inp}}$ of $2m$ examples. It suffices to show that no pair of examples in S_{inp} is comparable, because, then, any labeling of these examples is consistent with some monotone function and, hence, $\mathsf{opt}_{\mathrm{total}} = 0$. We overall have at most $4m^2$ pairs and each of them is comparable with probability at most $2 \cdot (3/4)^d$. Therefore, by a union bound, the probability that there is a pair that is comparable is at most $8m^2(3/4)^d \leq 1/10$. □

6 Discussion

In this work, we lay the groundwork for a systematic study of efficient learning algorithms in the presence of contamination. For bounded contamination, we showed that low-degree appriximators suffice, nearly matching the best upper bounds for learning under label noise—a significantly milder noise model. Our results capture the challenging setting of heavy contamination as well, for classes with low sandwiching degree. Central to our approach is the *iterative polynomial filtering* algorithm, which is related to outlier removal procedures with a long history in the literature of robust learning and optimization (see [KLS09, DKS18a, DKK+19a, DKK+19b, GSSV24, KSV24c] and references therein). There are several interesting directions for future work:

1. *Universality:* Our algorithms require samples from the target unlabeled distribution $\mathcal{D}^*$. Relaxing this requirement and obtaining algorithms that work universally with respect to broad classes of distributions is an important question. This is known to be possible for several problems in learning with label noise [BOW10, ABL17, DKTZ20b], robust unsupervised learning [KS17a, KS17b, DHPT24], as well as testable learning [GKSV23, CKK+24b].

2. *Characterization of Efficient Learnability under Contamination:* We show that $\mathcal{L}_2$ polynomial approximation suffices for efficient BC-learning and $\mathcal{L}_1$ sandwiching suffices for efficient HC-learning. Determining whether the weaker notion of $\mathcal{L}_1$ approximation is sufficient for BC-learning or relaxing the sandwiching requirement for HC-learning are both interesting open directions for future work.

3. *Improved Error Guarantees:* Our algorithms obtain error guarantees that are proven to be optimal in general (see Proposition B.3). However, it is an open question whether one can achieve error $\eta + \epsilon$ for BC-learning or error $\frac{1}{2}(Q \cdot \mathsf{opt}_{\text{total}} + \mathsf{opt}_{\text{clean}}) + \epsilon$ for HC-learning for special classes and distributions, perhaps by allowing the learner to output randomized hypotheses (also known as probabilistic concepts).

4. *Heavy Contamination without Label Noise:* Our definition for learning with heavy contamination allows the adversary to add arbitrary labeled examples, as long as the resulting contaminated dataset can be approximately labeled by some classifier in the learned class up to a small number of mistakes. Note that the optimal classifier can also be chosen adversarially, in the sense that we do not require that the optimal classifier for the contaminated dataset be the same as the optimal classifier on the clean distribution. This property is crucially used in our lower bound for learning monotone functions with heavy contamination (Theorem 5.2): the hard instance is perfectly labeled by some monotone function, but this monotone function is far from the one that generates the clean labels.

   On the other hand, even in the realizable HC setting—where the clean labels are given by some concept $f^* \in \mathcal{C}$ and the adversary is required to add points of the form $(\mathbf{x}, f^*(\mathbf{x}))$—it is not clear when efficient learning is possible. This is because for many classes, the assumptions on the distribution of $\mathbf{x}$ are crucial, even in the absence of label noise. It is therefore an interesting open question whether improved HC-learning guarantees can be achieved in the realizable setting.

References

[ABL17] 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. 1, 2, 6, 22

[Bal93] Keith Ball. The reverse isoperimetric problem for gaussian measure. *Discrete Comput. Geom.*, 10(4):411–420, dec 1993. 33

[Baz09] Louay MJ Bazzi. Polylogarithmic independence can fool dnf formulas. *SIAM Journal on Computing*, 38(6):2220–2272, 2009. 15

[BCO+15] Eric Blais, Clément L Canonne, Igor C Oliveira, Rocco A Servedio, and Li-Yang Tan. Learning circuits with few negations. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*, pages 512–527. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2015. 35

[BDH+20] Ainesh Bakshi, Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, Sushrut Karmalkar, and Pravesh K Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *2020 ieee 61st annual symposium on foundations of computer science (focs)*, pages 149–159. IEEE, 2020. 1

[BDLS17] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212. PMLR, 2017. 1

[BEK02] Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. Pac learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002. Algorithmic Learning Theory. 1

[BGMMS21] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021. 1

[BK21] Ainesh Bakshi and Pravesh K Kothari. List-decodable subspace recovery: Dimension independent error in polynomial time. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1279–1297. SIAM, 2021. 1, 2

[Bob01] Sergey G Bobkov. Some generalizations of prokhorov's results on khinchin-type inequalities for polynomials. *Theory of Probability & Its Applications*, 45(4):644–647, 2001. 18

[BOW10] Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. *Machine learning*, 80(2):273–294, 2010. 1, 22

[Bra08] Mark Braverman. Polylogarithmic independence fools ac 0 circuits. *Journal of the ACM (JACM)*, 57(5):1–10, 2008. 33

[BS95] Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995. 6

[BT96] Nader H Bshouty and Christino Tamon. On the fourier spectrum of monotone functions. *Journal of the ACM (JACM)*, 43(4):747–770, 1996. 33, 34

[CDG19] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the thirtieth annual ACM-SIAM symposium on discrete algorithms*, pages 2755–2771. SIAM, 2019. 1

[CG18] Yu Cheng and Rong Ge. Non-convex matrix completion against a semi-random adversary. In *Conference On Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1362–1394. PMLR, 2018. 6

[CKK+24a] Gautam Chandrasekaran, Adam Klivans, Vasilis Kontonis, Raghu Meka, and Konstantinos Stavropoulos. Smoothed analysis for learning concepts with low intrinsic dimension. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 876–922. PMLR, 2024. 6

[CKK+24b] Gautam Chandrasekaran, Adam Klivans, Vasilis Kontonis, Konstantinos Stavropoulos, and Arsen Vasilyan. Efficient discrepancy testing for learning with distribution shift. *Advances in Neural Information Processing Systems*, 37:137263–137308, 2024. 3, 7, 22

[CKLS25] Gautam Chandrasekaran, Adam Klivans, Lin Lin Lee, and Konstantinos Stavropoulos. Learning neural networks with distribution shift: Efficiently certifiable guarantees. In *The Thirteenth International Conference on Learning Representations*, 2025. 7, 30

[CMY20] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau. List decodable mean estimation in nearly linear time. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 141–148. IEEE, 2020. 1

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th annual ACM SIGACT symposium on theory of computing*, pages 47–60, 2017. 1, 2

[Dan16] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 105–117, 2016. 3, 6

[DGTZ18] Constantinos Daskalakis, Themis Gouleakis, Chistos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018. 2

[DGTZ19] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on learning theory*, pages 955–960. PMLR, 2019. 2

[DHK+10] Ilias Diakonikolas, Prahladh Harsha, Adam Klivans, Raghu Meka, Prasad Raghavendra, Rocco A. Servedio, and Li-Yang Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, page 533–542, New York, NY, USA, 2010. Association for Computing Machinery. 34

[DHPT24] Ilias Diakonikolas, Samuel B Hopkins, Ankit Pensia, and Stefan Tiegel. Sos certifiability of subgaussian distributions and its algorithmic applications. *arXiv preprint arXiv:2410.21194*, 2024. 1, 2, 22

[DK22] Ilias Diakonikolas and Daniel Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, pages 4258–4282. PMLR, 2022. 6

[DK23] Ilias Diakonikolas and Daniel M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023. 1

[DKK+19a] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. 5, 22

[DKK+19b] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019. 22

[DKK+22a] Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. List-decodable sparse mean estimation via difference-of-pairs filtering. *Advances in Neural Information Processing Systems*, 35:13947–13960, 2022. 1

[DKK+22b] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with general massart noise under the gaussian distribution. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 874–885, 2022. 1, 6

[DKK+23] Ilias Diakonikolas, Daniel Kane, Vasilis Kontonis, Sihan Liu, and Nikos Zarifis. Efficient testable learning of halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 36, 2023. 4, 7

[DKLZ24] Ilias Diakonikolas, Daniel Kane, Sihan Liu, and Nikos Zarifis. Testable learning of general halfspaces with adversarial label noise. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1308–1335. PMLR, 30 Jun–03 Jul 2024. 7

[DKMR22] Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi, and Lisheng Ren. Cryptographic hardness of learning halfspaces with massart noise. In *Advances in Neural Information Processing Systems*, 2022. 6

[DKN10] Ilias Diakonikolas, Daniel M Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 11–20. IEEE, 2010. 34

[DKPZ21] Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. In *Conference on Learning Theory*, pages 1552–1584. PMLR, 2021. 2, 6, 20, 21, 35

[DKS18a] 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. 1, 2, 5, 6, 11, 22

[DKS18b] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018. 1, 2

[DKTZ20a] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory*, pages 1486–1513. PMLR, 2020. 6

[DKTZ20b] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33:18540–18549, 2020. 1, 6, 22

[DSFT+14] Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 498–511. SIAM, 2014. 2, 6, 21, 35

[Fer14] Dietmar Ferger. Optimal constants in the marcinkiewicz–zygmund inequalities. *Statistics & Probability Letters*, 84:96–101, 2014. 10

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 563–574. IEEE, 2006. 3

[FK01] Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001. 6

[FKV20] Vitaly Feldman, Pravesh Kothari, and Jan Vondrák. Tight bounds on l1 approximation and learning of self-bounding functions. *Theoretical Computer Science*, 808:86–98, 2020. 34

[GIF+23] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 1

[GKK23] 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. 3, 4, 7, 17, 18

[GKKM20] Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. *Advances in Neural Information Processing Systems*, 33:15859–15870, 2020. 2, 7, 17

[GKSV23] Aravind Gollakota, Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Tester-learners for halfspaces: Universal algorithms. *Advances in Neural Information Processing Systems*, 36, 2023. 4, 7, 22

[GKSV24] Aravind Gollakota, Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. An efficient tester-learner for halfspaces. *The Twelfth International Conference on Learning Representations*, 2024. 4, 7

[GKSV25] Surbhi Goel, Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testing noise assumptions of learning algorithms. *arXiv preprint arXiv:2501.09189*, 2025. 7

[GOWZ10] 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. 33

[GSSV24] Surbhi Goel, Abhishek Shetty, Konstantinos Stavropoulos, and Arsen Vasilyan. Tolerant algorithms for learning with arbitrary covariate shift. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 124979–125018. Curran Associates, Inc., 2024. 1, 2, 3, 4, 5, 6, 7, 11, 17, 18, 22, 30

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018. 1

[HM13] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375. PMLR, 2013. 1

[HS19] Prahladh Harsha and Srikanth Srinivasan. On polynomial approximations to AC^0. *Random Structures & Algorithms*, 54(2):289–303, 2019. 33

[IK22] Misha Ivkov and Pravesh K Kothari. List-decodable covariance estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1276–1283, 2022. 1, 2

[JLM+23] Arun Jambulapati, Jerry Li, Christopher Musco, Kirankumar Shiragur, Aaron Sidford, and Kevin Tian. Structured semidefinite programming for recovering structured preconditioners. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023. 2, 6

[Kan11a] Daniel M Kane. The gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. *computational complexity*, 20(2):389–412, 2011. 33, 34

[Kan11b] Daniel M. Kane. k-independent gaussians fool polynomial threshold functions, 2011. 34

[Kan13] Daniel M Kane. The correct exponent for the gotsman-linial conjecture. In *2013 IEEE Conference on Computational Complexity*, pages 56–64. IEEE, 2013. 33, 34

[Kan14] Daniel M Kane. The average sensitivity of an intersection of half spaces. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 437–440, 2014. 33, 34

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, November 1998. 34

[Kha93] 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. 35

[KKK19] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. *Advances in neural information processing systems*, 32, 2019. 2

[KKM13] Daniel M. Kane, Adam R. Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 522–545. JMLR, 2013. 18, 34

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. 1, 2, 5, 13, 16, 17, 19, 21, 30, 34

[KL93] Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22(4):807–837, 1993. 1

[KLL+23] Jonathan Kelner, Jerry Li, Allen X. Liu, Aaron Sidford, and Kevin Tian. Semi-random sparse recovery in nearly-linear time. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2352–2398. PMLR, 12–15 Jul 2023. 2, 6

[KLL+24] Jonathan A. Kelner, Jerry Li, Allen Liu, Aaron Sidford, and Kevin Tian. Semi-random matrix completion via flow-based adaptive reweighting. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*, 2024. 6

[KLS09] Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009. 1, 2, 22

[KM13] Adam Klivans and Raghu Meka. Moment-matching polynomials. *arXiv preprint arXiv:1301.0820*, 2013. 18, 34

[KOS08] Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning geometric concepts via gaussian surface area. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550. IEEE, 2008. 1, 12, 33, 34

[KS17a] Pravesh K Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *arXiv preprint arXiv:1711.07465*, 2017. 1, 22

[KS17b] Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017. 1, 22

[KSV24a] Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Learning intersections of halfspaces with distribution shift: Improved algorithms and sq lower bounds. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2944–2978. PMLR, 30 Jun–03 Jul 2024. 7

[KSV24b] 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. 2, 3, 7, 17

[KSV24c] Adam R Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Learning constant-depth circuits in malicious noise models. *arXiv preprint arXiv:2411.03570*, 2024. 1, 2, 3, 5, 6, 8, 11, 22

[KZZ24] Alkis Kalavasis, Ilias Zadik, and Manolis Zampetakis. Transfer learning beyond bounded density ratios. *arXiv preprint arXiv:2403.11963*, 2024. 2

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993. 33, 34

[LRB+21] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 13–24. IEEE, 2021. 1

[RV23] Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. *Proceedings of the fifty-fifth annual ACM Symposium on Theory of Computing*, 2023. 4, 6, 17

[RY20] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM, 2020. 1, 2

[She25] Jie Shen. Efficient pac learning of halfspaces with constant malicious noise rate. In *36th International Conference on Algorithmic Learning Theory*, 2025. 1

[SKL17] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30, 2017. 1

[STW24] Lucas Slot, Stefan Tiegel, and Manuel Wiedmer. Testably learning polynomial threshold functions. In *38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. 4, 7, 34

[Tal17] Avishay Tal. Tight bounds on the Fourier spectrum of AC^0. In *32nd Computational Complexity Conference (CCC 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. 33

[Val85] 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. 1

[ZS22] Shiwei Zeng and Jie Shen. List-decodable sparse mean estimation. *Advances in Neural Information Processing Systems*, 35:24031–24045, 2022. 1, 2

A Additional Tools

In our proofs, we make use of the following results for hypercontractive distributions.

Lemma A.1 (Loewner Concentration, Lemma 3.8 in [CKLS25], Lemma B.1 in [GSSV24])**.** *Let $\mathcal{D}$ be some A-hypercontractive distribution in d dimensions and let S be a set of m i.i.d. examples from $\mathcal{D}$ where $m \geq (CAd)^{2\ell}(\log \frac{1}{\delta})^{4\ell+1}$ for some sufficiently large universal constant $C \geq 1$, $\delta \in (0,1)$ and $A, \ell, d \geq 1$. Then, with probability at least $1-\delta$ over S the following condition holds.*

$$\frac{1}{2}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[(p(\mathbf{x}))^2] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim S}[(p(\mathbf{x}))^2] \leq 2\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[(p(\mathbf{x}))^2]\,, \text{ for any } p \text{ of degree at most } \ell$$

Lemma A.2. *Let $\mathcal{D}$ be some A-hypercontractive distribution in d dimensions for some $A \geq 1$. Let S be a set of m i.i.d. samples from $\mathcal{D}$ and $\mathcal{P}_\beta(S)$ be the family of polynomials of degree at most ℓ such that $\mathbb{E}_{\mathbf{x}\sim S}[(p(\mathbf{x}))^2] \leq \beta$, where $\ell, \beta, d \geq 1$. Then, for any $\epsilon, \delta \in (0,1)$, if $m \geq (CAd)^{2\ell}(\log \frac{1}{\delta})^{4\ell+1}$, and $B \geq \sqrt{2\beta/\epsilon}\cdot(d+1)^{\ell/2}$ the following holds with probability at least $1-\delta$ over S:*

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[|p(\mathbf{x})| \leq B\,, \text{ for all } p \in \mathcal{P}_\beta(S)] \geq 1-\epsilon$$

Proof of Lemma A.2. We first apply Lemma A.1 to show that, with probability at least $1-\delta$ over S, for any $p \in \mathcal{P}_\beta(S)$ we have $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(p(\mathbf{x}))^2] \leq 2\,\mathbb{E}_{\mathbf{x}\sim S}[(p(\mathbf{x}))^2] \leq 2\beta$.

Consider now the moment matrix $\mathbf{M} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\psi_\ell(\mathbf{x})\psi_\ell(\mathbf{x})^\top]$, where $\psi_\ell(\mathbf{x})$ is the vector whose coordinates correspond to monomials of $\mathbf{x}$ of degree at most ℓ. Let $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ be the concise SVD of $\mathbf{M}$, i.e., $\mathbf{D}$ is diagonal with dimension equal to the rank of $\mathbf{M}$, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{U}\mathbf{U}^\top$ is the orthogonal projection on the column (or row) space of $\mathbf{M}$. Let $\varphi(\mathbf{x}) = \mathbf{D}^{-1/2}\mathbf{U}^\top\psi_\ell(\mathbf{x})$ and observe that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\varphi(\mathbf{x})\varphi(\mathbf{x})^\top] = \mathbf{I}$ and the dimension of φ is equal to the rank of $\mathbf{M}$, which is at most $(d+1)^\ell$. Here, we use the fact that all the entries of $\mathbf{M}$ are finite, due to hypercontractivity.

We may express any polynomial p in $\mathcal{P}_\beta(S)$ as a linear combination of the elements of φ, i.e., $p(\mathbf{x}) = \tilde{\mathbf{c}}_p^\top\varphi(\mathbf{x})$, where $\|\tilde{\mathbf{c}}_p\|_2^2 = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(p(\mathbf{x}))^2] \leq 2\beta$. Moreover, we have the following due to Markov's inequality and the fact that $\mathbb{E}[\|\varphi(\mathbf{x})\|_2^2] \leq (d+1)^\ell$.

$$\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}\Big[\|\varphi(\mathbf{x})\|_2^2 > \frac{B^2}{2\beta}\Big] \leq \frac{2\beta(d+1)^\ell}{B^2} \leq \epsilon$$

In the event that $\|\varphi(\mathbf{x})\|_2^2 \leq \frac{B^2}{2\beta}$, we have $|p(\mathbf{x})| \leq \|\tilde{\mathbf{c}}_p\|_2\|\varphi(\mathbf{x})\|_2 \leq \sqrt{2\beta}\cdot\frac{B}{\sqrt{2\beta}} = B$. □

We will also use the following theorem which is implicit in [KKMS08] and shows that polynomial approximation is sufficient to find a simple hypothesis with near-optimum error on a given dataset.

Theorem A.3 ($\mathcal{L}_1$ polynomial regression [KKMS08])**.** *Let $\bar{S}$ be a set of labeled examples and $\mathcal{C}$ be a concept class such that for each $f \in \mathcal{C}$ there is some polynomial p of degree at most ℓ such that $\mathbb{E}_{\mathbf{x}\sim S}[|f(\mathbf{x}) - p(\mathbf{x})|] \leq \epsilon$. Then, the degree-$\ell$ $\mathcal{L}_1$ polynomial regression algorithm of [KKMS08] outputs, in time $\mathrm{poly}(|\bar{S}|d^\ell/\epsilon)$, a degree-$\ell$ polynomial threshold function h such that $\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}}[y \neq h(\mathbf{x})] \leq \min_{f\in\mathcal{C}}\mathbb{P}_{(\mathbf{x},y)\sim\bar{S}}[y \neq f(\mathbf{x})] + \epsilon$.*

We also give the definition of $\mathcal{L}_1$ approximating polynomials, which are sufficient for agnostic learning, but it is not clear whether they suffice for BC-learning. We instead use the notion of $\mathcal{L}_2$ approximating polynomials (Definition 4.1).

Definition A.4 ($\mathcal{L}_1$ Polynomial Approximators)**.** For $\epsilon \in (0,1)$, we say that a class $\mathcal{C} \subseteq \{\mathcal{X} \to \{\pm 1\}\}$ has ϵ-approximate degree $\ell = \ell(\epsilon)$ with respect to some distribution $\mathcal{D}^*$ over $\mathcal{X}$ if for any $f \in \mathcal{C}$ there is a polynomial p of degree at most $\ell(\epsilon)$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}^*}[|f(\mathbf{x}) - p(\mathbf{x})|] \leq \epsilon$.

The difference between $\mathcal{L}_1$ and $\mathcal{L}_2$ approximators is that the former enjoy the approximation property with respect to the $\mathcal{L}_1$ norm, while the latter with respect to the $\mathcal{L}_2$ norm. Achieving $\mathcal{L}_2$ approximation is, in general, a stronger assumption, because the $\mathcal{L}_1$ norm is upper bounded by the $\mathcal{L}_2$ norm, but not vice versa. Nevertheless, for most interesting classes that $\mathcal{L}_1$ approximators are known, we also have $\mathcal{L}_2$ approximators. This is because of analytic advantages of the $\mathcal{L}_2$ norm.

B Error Lower Bounds for HC-Learning

In order to prove our lower bounds, we provide a definition for the adversary in the context of heavy contamination.

Definition B.1 (Contamination Strategy)**.** An algorithm $\mathcal{M}$ is called a (randomized) Q-HC strategy if, upon receiving a labeled set $\bar{S}_{\text{cln}}$, it outputs a set $\bar{S}_{\text{inp}}$ of size $|\bar{S}_{\text{inp}}| \leq Q|\bar{S}_{\text{cln}}|$ such that $\bar{S}_{\text{inp}} \subseteq \bar{S}_{\text{cln}}$. For $m \in \mathbb{N}$ and a labeled distribution $\bar{\mathcal{D}}$, we denote with $\mathcal{M}(\bar{\mathcal{D}}^m)$ the distribution of the output $\bar{S}_{\text{inp}}$ of $\mathcal{M}$ on a set of m i.i.d. samples $\bar{S}_{\text{cln}}$ from $\bar{\mathcal{D}}$. Note that $\bar{S}_{\text{inp}}$ depends on the random choice of $\bar{S}_{\text{cln}}$, as well as the randomness of $\mathcal{M}$.

We first prove the following lower bound, which holds for any non-trivial concept class.

Proposition B.2. *Let $\mathcal{C}$ be any concept class such that there is some $f \in \mathcal{C}$ for which $-f \in \mathcal{C}$. For any $Q \in \{2, 3, \dots\}$, any unlabeled distribution $\mathcal{D}^*$, any $\rho \in [0, 1/2)$, $\epsilon, \delta > 0$, and any $m \geq \frac{\log(2/\delta)}{2\epsilon^2}$, there is a Q-HC strategy $\mathcal{M}$ and a distribution $\bar{\mathcal{D}}$ whose marginal is $\mathcal{D}^*$ such that, if we let $\bar{\mathcal{D}}'$ be the distribution of $(\mathbf{x}, -y)$, where $(\mathbf{x}, y) \sim \bar{\mathcal{D}}$, then the following hold.*

1. *$\mathcal{M}(\bar{\mathcal{D}}^m) = \mathcal{M}((\bar{\mathcal{D}}')^m)$.*
2. *$\min_{f \in \mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})] = \min_{f \in \mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}'}[y \neq f(\mathbf{x})] = \rho$.*
3. *With probability at least $1 - \delta$ over $\bar{S}_{\text{inp}} \sim \mathcal{M}(\bar{\mathcal{D}}^m)$, we have $Q \cdot \mathsf{opt}_{\text{total}} \leq 1 - \rho + \epsilon$, where $\mathsf{opt}_{\text{total}} = \min_{f \in \mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{S}_{\text{inp}}}[y \neq f(\mathbf{x})]$.*

Hence, no HC-learner for $\mathcal{C}$ with respect to $\mathcal{D}^*$ can achieve error less than $\frac{1}{2}(Q \cdot \mathsf{opt}_{\text{total}} + \mathsf{opt}_{\text{clean}} - \epsilon)$ with probability more than $\frac{1}{2} - \delta$, even when $\mathsf{opt}_{\text{clean}} = \rho$.

Proof of Proposition B.2. Let $\mathcal{D} = \mathcal{D}^*$. Consider the following two labeled distributions. First, $\bar{\mathcal{D}}_1$ is a distribution over $\mathcal{X} \times \{\pm 1\}$ whose marginal on $\mathcal{X}$ is $\mathcal{D}$ and the labels are generated as follows:

$$y_1 = \begin{cases} f(\mathbf{x}), \text{ with probability } 1 - \rho \\ -f(\mathbf{x}), \text{ with probability } \rho \end{cases}$$

Consider also a distribution $\bar{\mathcal{D}}_2$ whose marginal is also $\mathcal{D}$ but the labels are generated as follows:

$$y_2 = \begin{cases} -f(\mathbf{x}), \text{ with probability } 1 - \rho \\ f(\mathbf{x}), \text{ with probability } \rho \end{cases}$$

The adversary receives a set $\bar{S}_{\text{cln}}$ of m i.i.d. samples from $\bar{\mathcal{D}} \in \{\bar{\mathcal{D}}_1, \bar{\mathcal{D}}_2\}$ and does the following.

1. First, the adversary computes the value $m_1 = \sum_{(\mathbf{x},y)\in\bar{S}_{\text{inp}}} \mathbb{1}\{y = f(\mathbf{x})\}$, as well as the value $m_2 = \sum_{(\mathbf{x},y)\in\bar{S}_{\text{inp}}} \mathbb{1}\{y = -f(\mathbf{x})\}$.
2. If $m_1 \geq m_2$, then the adversary adds $M - m$ examples drawn by $\mathcal{D}$ and labeled by $-f(\mathbf{x})$.
3. If $m_1 < m_2$, then the adversary adds $m_2 - m_1$ examples drawn by $\mathcal{D}$ and labeled by $f(\mathbf{x})$, as well as $M - 2m_2$ examples drawn by $\mathcal{D}$ and labeled by $-f(\mathbf{x})$.

Part 1. Let $\xi = m_i$, where $i \in \{1, 2\}$ such that $\bar{\mathcal{D}} = \bar{\mathcal{D}}_i$. We have that ξ follows the binomial distribution with parameters $(m, 1-\rho)$, and is independent from the value of i.

Given any realization m_i of ξ with $m_i > m/2$, we can equivalently form the input examples $\bar{S}_{\text{inp}}$ by drawing M i.i.d. samples from $\mathcal{D}$ and labeling ξ of them according to $f(\mathbf{x})$ and $M - \xi$ of them according to $-f(\mathbf{x})$. We have that $\mathcal{M}(\bar{\mathcal{D}}_1^m)|_{\xi > m/2} = \mathcal{M}(\bar{\mathcal{D}}_2^m)|_{\xi > m/2}$.

Given any realization m_i of ξ with $m_i \leq m/2$, we can equivalently form the input examples $\bar{S}_{\text{inp}}$ by drawing M i.i.d. samples from $\mathcal{D}$ and labeling $m - \xi$ of them according to $f(\mathbf{x})$ and $M - m + \xi$ of them according to $-f(\mathbf{x})$. We, therefore, have $\mathcal{M}(\bar{\mathcal{D}}_1^m)|_{\xi \leq m/2} = \mathcal{M}(\bar{\mathcal{D}}_2^m)|_{\xi \leq m/2}$.

Overall, since ξ does not depend on i, we have $\mathcal{M}(\bar{\mathcal{D}}_1^m) = \mathcal{M}(\bar{\mathcal{D}}_2^m)$.

Part 2. Follows immediately from the fact that $f, -f \in \mathcal{C}$ and the definition of $\bar{\mathcal{D}} = \bar{\mathcal{D}}_1$, $\bar{\mathcal{D}}' = \bar{\mathcal{D}}_2$.

Part 3. Regarding the error benchmark, we have $Q \cdot \mathsf{opt}_{\text{total}} = \frac{\max\{\xi, m-\xi\}}{m} \leq 1 - \rho + \epsilon$, whenever $|\xi - (1-\rho)m| \leq \epsilon m$, which happens with probability at least $1 - \delta$, according to the following inequality.

$$\mathbb{P}[|\xi - (1-\rho)m| > \epsilon m] \leq 2\exp(-2\epsilon^2 m) \leq \delta$$

The bound follows from an application of Hoeffding's inequality, since ξ is binomial of mean $(1-\rho)m$. It follows that with probability at least $1 - \delta$, the algorithm outputs a hypothesis with error at least $\frac{1}{2}(Q \cdot \mathsf{opt}_{\text{total}} + \mathsf{opt}_{\text{clean}} - \epsilon)$.

Implication to HC Learning. We have $\mathcal{M}(\bar{\mathcal{D}}_1^m) = \mathcal{M}(\bar{\mathcal{D}}_2^m)$. Let $\mathcal{A}$ be any, potentially randomized algorithm, and let $h := \mathcal{A}(\mathcal{M}(\bar{\mathcal{D}}_1^m)) = \mathcal{A}(\mathcal{M}(\bar{\mathcal{D}}_2^m))$ be the random variable corresponding to its output, where $h : \mathcal{X} \to \{\pm 1\}$. Consider the event that h is such that: $\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}_1}[h(\mathbf{x}) \neq y] < 1/2$. Then, we have:

$$\mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}_2}[h(\mathbf{x}) \neq -y] = \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}_1}[h(\mathbf{x}) \neq -y] = 1 - \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}_1}[h(\mathbf{x}) \neq y] > 1/2$$

Note that the same is true even if h is a randomized classifier. Overall, the error of $\mathcal{A}$ on either $\bar{\mathcal{D}}_1$ or $\bar{\mathcal{D}}_2$ is $1/2$ with probability at least $1/2$. Finally, due to part 3, with probability at least $1 - \delta$, we have $1/2 \geq \frac{1}{2}(Q \cdot \mathsf{opt}_{\text{total}} + \mathsf{opt}_{\text{clean}} - \epsilon)$, which implies the desired result. □

For finite feature spaces and the particular choice of the class of all functions, we obtain a sharper information-theoretic lower bound on the error. Note that this lower bound captures the problem of HC-learning halfspaces over $\{\pm 1\}$ in one dimension.

Proposition B.3. *Suppose that $|\mathcal{X}| < \infty$ and $\mathcal{C}$ is any concept class from $\mathcal{X}$ to $\{\pm 1\}$. Then, on any value of Q in $\{2, 3, \dots\}$, any (distribution-specific) HC-learner will output a hypothesis with error 1 with probability at least $2^{-|\mathcal{X}|}$ on some instance where $Q \cdot \mathsf{opt}_{\text{total}} = 1$. Moreover, if $\mathcal{C} = \{\pm 1\}^{\mathcal{X}}$, then the instance satisfies $\mathsf{opt}_{\text{clean}} = 0$.*

Proof of Proposition B.3. The adversary receives the clean dataset $\bar{S}_{\text{cln}}$ and creates a duplicate $\bar{S}'_{\text{cln}}$, where all the labels are flipped but the feature vectors are unchanged. Then, the adversary draws $(Q-2) \cdot m$ i.i.d. examples S' from the marginal distribution $\mathcal{D}$ on $\mathcal{X}$ of the clean distribution $\bar{\mathcal{D}}$ and labels them according to f_1. The input dataset is $\bar{S}_{\text{cln}} \cup \bar{S}'_{\text{cln}} \cup \bar{S}'$.

Note that the input dataset is completely independent from the clean labels, since $\bar{S}_{\text{cln}} \cup \bar{S}'_{\text{cln}}$ can be constructed equivalently by choosing an arbitrary function $g : \mathcal{X} \to \{\pm 1\}$ to label the points in $\bar{S}_{\text{cln}}$ and then label the points in $\bar{S}'_{\text{cln}}$ according to $-g$.

Therefore, we may assume that the ground truth labels are generated by the function f^*, where $-f^*$ is the most likely output of the HC-learner on a dataset of the form $\bar{S}_{\mathrm{cln}} \cup \bar{S}'_{\mathrm{cln}} \cup \bar{S}'$. Since there are at most $2^{|\mathcal{X}|}$ possible functions, $\mathcal{A}$ must output $-f^*$ with probability at least $2^{-|\mathcal{X}|}$. When $\mathcal{A}$ outputs $-f^*$, the error is 1.

Finally, note that f_1 achieves error on $\bar{S}_{\mathrm{inp}}$ equal to $1/Q$, so $\mathsf{opt}_{\mathrm{total}} = 1/Q$ and $Q \cdot \mathsf{opt}_{\mathrm{total}} = 1$.

When $\mathcal{C} = \{\pm 1\}^{\mathcal{X}}$, we have $f^* \in \mathcal{C}$ and, hence, $\mathsf{opt}_{\mathrm{clean}} := \min_{f \in \mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})] = 0$. □

C Approximation Theory Results

C.1 Low-Degree Approximators

Our structural Theorem 4.2 can be combined with results from polynomial approximation theory in order to obtain end-to-end results on learning with nasty noise, including the results of Table 2.

Intersections of Halfspaces. For the class of k-halfspace intersections, [KOS08] showed that the ϵ-approximate degree with respect to the standard Gaussian distribution $\mathcal{N}_d$ (see Definition 4.1) is $\ell(\epsilon) \leq O(\log(k)/\epsilon^2)$ and [Kan14] obtains the same result with respect to the uniform distribution Unif_d over the hypercube.

Monotone Functions and Convex Sets. For monotone functions, the ϵ-approximate degree with respect to Unif_d was bounded by $\ell(\epsilon) \leq O(\sqrt{d}/\epsilon)$ in [BT96]. For convex sets, [KOS08] showed that the ϵ-approximate degree with respect to the Gaussian distribution can be bounded via the Gaussian Surface Area (GSA). Combined with the result of [Bal93] on the GSA of convex sets, the corresponding bound is $\ell(\epsilon) \leq O(\sqrt{d}/\epsilon^2)$.

Polynomial Threshold Functions. For the class of degree-k PTFs, the ϵ-approximate degree with respect to the standard Gaussian was bounded by $\ell(\epsilon) \leq O(k^2/\epsilon^2)$ in [Kan11a]. In [Kan13], it was shown that $\ell(\epsilon) \leq \frac{k^{O(k^2)}(\log 1/\epsilon)^{O(k \log k)}}{\epsilon^2}$ with respect to the uniform over $\{\pm 1\}^d$.

Low-depth Circuits. For the class of AC^0 circuits of size s and depth t, the ϵ-approximate degree is known to be at most $\ell(\epsilon) = O(\log(s))^{t-1} \log(1/\epsilon)$, due to the seminal work of [LMN93] and subsequent improvements [Tal17].

C.2 Sandwiching Approximators

For the more challenging problem of learning with heavy contamination, we require the existence of sandwiching approximators (see Theorem 4.4), and show that arbitrary (non-sandwiching) approximators do not suffice (see Theorem 5.2).

Decision Trees of Halfspaces. For the class of depth-t, size-s decision trees of halfspaces, [GOWZ10] showed that the ϵ-sandwiching degree with respect to either the standard Gaussian or the uniform on the hypercube is $\ell(\epsilon) \leq \tilde{O}(\frac{t^4 s^2}{\epsilon^2})$.

Low-Depth Circuits. The ϵ-sandwiching degree of AC^0 circuits of depth t and size s is known to be at most $\ell(\epsilon) \leq O(\log(s))^{O(t)} \log(1/\epsilon)$, due to the celebrated results of [Bra08, Tal17, HS19].

Polynomial Threshold Functions. The ϵ-sandwiching degree of degree-k PTFs over the Gaussian was bounded by $\ell(\epsilon) \leq O_k(\epsilon^{-4k\cdot 7^k})$ in [STW24], based on the pseudorandom generator of [Kan11b]. Here, $O_k(\cdot)$ is hiding a multiplicative factor that scales as an arbitrary function of k. For $k = 2$, the sandwiching degree with respect to the uniform distribution over the hypercube is known to be at most $\ell(\epsilon) \leq O(1/\epsilon^9)$ due to [DKN10].

Functions of Halfspaces over Log-Concave Measures. For arbitrary functions of k halfspaces, the ϵ-sandwiching degree with respect to any log-concave distribution was shown to be at most $\ell(\epsilon) \leq \exp((\log(\log(k)/\epsilon))^{O(k)}/\epsilon^4)$ by [KM13, KKM13].

D Complexity of Learning with Adversarial Label Noise

The computational complexity of agnostic learning (or equivalently learning with adversarial label noise) is much better understood than the complexity of learning with contamination. In this setting, the learner receives i.i.d. examples from some labeled distribution $\bar{\mathcal{D}}$ whose marginal on the feature space $\mathcal{X}$ is well-behaved (i.e., Gaussian or uniform) and is, otherwise, arbitrary. The goal is to output, with high probability, a hypothesis whose error is at most $\mathsf{opt} + \epsilon$, where $\mathsf{opt} = \min_{f\in\mathcal{C}} \mathbb{P}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}[y \neq f(\mathbf{x})]$, for some target concept class $\mathcal{C}$.

Concept Class	Target Marginal	Upper Bounds	Lower Bounds
Intersections of k Halfspaces	$\mathcal{N}(0, \mathbf{I}_d)$	$d^{\tilde{O}(\log(k)/\epsilon^4)}$	$d^{\Omega(\sqrt{\log k}/\epsilon)}$
Depth-t, Size-d Boolean Circuits	$\mathrm{Unif}\{\pm 1\}^d$	$d^{O(\log d)^{t-1}\cdot \log 1/\epsilon}$	$d^{(\log d)^{\Omega(t)}} (\epsilon = O(1))$
Degree-k PTFs	$\mathcal{N}(0, \mathbf{I}_d)$	$d^{\tilde{O}(k^2/\epsilon^4)}$	$d^{\Omega(k^2/\epsilon^2)}$
Monotone Functions	$\mathrm{Unif}\{\pm 1\}^d$	$2^{\tilde{O}(\sqrt{d}/\epsilon^2)}$	$2^{\Omega(\sqrt{d}/\epsilon)}$

Table 4: Upper and lower bounds on the time complexity of agnostic learning up to excess error ϵ and failure probability $\delta = 0.01$. The lower bounds either hold for algorithms in the Statistical Query model [Kea98] or for any algorithm under appropriate cryptographic assumptions.

Upper Bounds. On the upper bound side, [KKMS08] showed that any class with bounded $\mathcal{L}_1$ approximation degree admits dimension-efficient agnostic learners via low-degree $\mathcal{L}_1$ polynomial regression. Several prior and subsequent works provided such bounds for many fundamental classes, based on Fourier or Hermite analysis [LMN93, BT96, DHK+10, Kan11a, Kan13, Kan14], as well as geometric properties of Gaussian spaces [KOS08]. Note that the usual way to bound the $\mathcal{L}_1$ approximation degree, is to first bound the $\mathcal{L}_2$ approximation degree and then use Cauchy-Schwarz inequality. This is because bounding the $\mathcal{L}_2$ approximation is usually more analytically convenient. However, this approach does not always give tight results for the degree in terms of the approximation error. In [FKV20], it was shown that $\mathcal{L}_1$ approximation bounds over the hypercube can be obtained directly for any function class with bounded noise sensitivity.

Lower Bounds. The complexity of agnostic learning with respect to Gaussian marginals was shown to be characterized in terms of the $\mathcal{L}_1$ approximation degree for algorithms that fall in the statistical query (SQ) framework by [DKPZ21], implying the lower bounds of Table 4 for intersections of halfspaces and polynomial threshold functions. For the uniform distribution on the hypercube, a similar characterization was given by [DSFT$^+$14]. Moreover, [BCO$^+$15] showed an information-theoretic lower bound for learning monotone functions even in the realizable setting, implying the lower bound of Table 4 for monotone functions. For AC^0 circuits, [Kha93] provided a cryptographic lower bound.