JMLR: Workshop and Conference Proceedings vol 35:1–22, 2014

# Distribution-Independent Reliable Learning

**Varun Kanade** VKANADE@EECS.BERKELEY.EDU
*University of California, Berkeley*

**Justin Thaler** JTHALER@SEAS.HARVARD.EDU
*Simons Institute for the Theory of Computing at UC Berkeley*

## Abstract

We study several questions in the *reliable agnostic* learning framework of Kalai et al. (2009), which captures learning tasks in which one type of error is costlier than other types. A positive reliable classifier is one that makes no false positive errors. The goal in the *positive reliable* agnostic framework is to output a hypothesis with the following properties: (i) its false positive error rate is at most $\epsilon$, (ii) its false negative error rate is at most $\epsilon$ more than that of the best positive reliable classifier from the class. A closely related notion is *fully reliable* agnostic learning, which considers *partial classifiers* that are allowed to predict "unknown" on some inputs. The best fully reliable partial classifier is one that makes no errors and minimizes the probability of predicting "unknown", and the goal in fully reliable learning is to output a hypothesis that is almost as good as the best fully reliable partial classifier from a class.

For distribution-independent learning, the best known algorithms for PAC learning typically utilize polynomial threshold representations, while the state of the art agnostic learning algorithms use point-wise polynomial approximations. We show that *one-sided polynomial approximations*, an intermediate notion between polynomial threshold representations and point-wise polynomial approximations, suffice for learning in the reliable agnostic settings. We then show that majorities can be fully reliably learned and disjunctions of majorities can be positive reliably learned, through constructions of appropriate one-sided polynomial approximations. Our fully reliable algorithm for majorities provides the first evidence that fully reliable learning may be strictly easier than agnostic learning. Our algorithms also satisfy strong attribute-efficiency properties, and in many cases they provide smooth tradeoffs between sample complexity and running time.

**Keywords:** reliable learning, agnostic learning, attribute-efficiency, polynomial approximations

## 1. Introduction

In many learning tasks, one type of error is costlier than other types. For example, when detecting spam messages, an important mail marked as spam (a false positive) is a major problem, whereas false negatives are only a minor nuisance. On the other hand, in settings such as detecting failures in an electric network, false negatives may be very harmful. In yet other settings, it may be better to refrain from making a prediction at all, rather than make a wrong one, e.g., when detecting medical ailments. Following Kalai et al. (2012), we call these kinds of tasks *reliable learning*. Closely related tasks have been widely studied in the statistics and machine learning literature. We discuss some of this work later; here, we simply note that the work of Kalai et al. and the present work depart from much of the extant literature by emphasizing computational considerations, i.e., by focusing on "fast" algorithms, and guarantees with respect to the *zero-one* loss.

Kalai et al. (2012) introduced a formal framework to study reliable learning in the *agnostic* setting, which is a challenging model that captures the problem of learning in the presence of adversarial classification noise. In particular, the goal of an agnostic learning algorithm is to produce a hypothesis that has error that is at most $\epsilon$ higher than the best from a certain class. A false positive error occurs when the true label is negative, but the hypothesis predicts positive. Analogously, a false negative error occurs when the true label is positive, but the hypothesis predicts negative.

The best positive reliable classifier from a class is one that make no false positive errors and minimizes false negative errors. In the *positive reliable* learning setting, the goal of a learning algorithm is to output a hypothesis with the following properties: (i) its false positive error rate is at most $\epsilon$, (ii) its false negative error rate is at most $\epsilon$ more than that of the best positive reliable classifier from the class. The notion of *negative reliable* learning is identical with the roles of false positive and false negatives reversed.

Kalai et al. (2012) also introduced the notion of full reliability. A partial classifier is one that is allowed to sometimes predict "?" or *unknown*. The best partial classifier from a class is one that makes no errors and minimizes the probability of predicting ?. In the *fully reliable* learning setting, the goal of the learning algorithm is to output a hypothesis $h : X \to \{-1, ?, +1\}$ such that (i) the error of $h$ is at most $\epsilon$, (ii) the probability that $h$ predicts '?' is at most $\epsilon$ more than the best partial classifier from the class.

**Our Contributions**: In this work, we focus on distribution-independent reliable learning, and our main technical contribution is to give new reliable learning algorithms for a variety of concept classes. We now place our reliable learning algorithms in the context of prior work on PAC and agnostic learning.

The *threshold degree* of a Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ is the least degree of a real polynomial that agrees in sign with $f$ at all inputs $x \in \{-1, 1\}^n$. The *approximate degree* (with error parameter $\epsilon$) of $f$ is the least degree of a real polynomial that point-wise approximates $f$ to error at most $\epsilon$. It is well-known that concept classes with low threshold degree can be efficiently learned in Valiant's PAC model under arbitrary distributions; indeed, threshold degree upper bounds underlie the fastest known PAC learning algorithms for a variety of fundamental concept classes, including DNF and read-once formulae (Klivans and Servedio, 2004). Meanwhile, concept classes with low approximate degree can be efficiently learned in the agnostic model, a connection that has yielded the fastest known algorithms for distribution-independent agnostic learning (Kalai et al., 2005).

We show that concept classes with low *one-sided* approximate degree can be efficiently learned in the *reliable* agnostic model. Here, one-sided approximate degree is an intermediate notion that lies between threshold degree and approximate degree; we defer a formal definition to Section 2.2. One-sided approximate degree was introduced in its present form by Bun and Thaler (2013a) (see also (Sherstov, 2014)), though equivalent *dual* formulations had been used in several prior works (Gavinsky and Sherstov, 2010; Sherstov, 2013a; Bun and Thaler, 2013b). Our learning algorithm is similar to the $L_1$ regression algorithm of Kalai et al. (2005); however, the analysis of our algorithm is more delicate. Specifically, due to asymmetry in the type of errors considered in the reliable setting, our analysis requires the use of *two* loss functions. On one side, we use the *hinge* loss rather than $L_1$ loss, since one-sided polynomial approximations may be unbounded, and on the other, we use a non-convex Lipschitz approximation to the *zero-one* loss.

We identify important concept classes, such as majorities and intersections of majorities, whose one-sided approximate degree is strictly smaller than its approximate degree. Consequently, we obtain reliable (in the case of majorities, even *fully reliable*) agnostic learning algorithms that are strictly more efficient than the fastest known algorithms in the standard agnostic setting. Our fully reliable learning algorithm for majorities gives the first indication that fully reliable learning may be strictly easier than agnostic learning. Finally, we show how to obtain smooth *tradeoffs* between sample complexity and runtime of algorithms for agnostically learning conjunctions, and for positive reliably learning DNF formulae.

In more detail, we summarize our new algorithmic results as follows (for simplicity, we omit dependence on the error parameter $\epsilon$ of the learning algorithm from this overview). We give:

- A simple $\mathrm{poly}(n)$ time algorithm for positive reliable learning of disjunctions.
- A $2^{\tilde{O}(\sqrt{n})}$ time algorithm for fully reliable learning of majorities. In contrast, no $2^{o(n)}$-time algorithm for agnostically learning majorities is known in the arbitrary-distribution setting.
- A $2^{\tilde{O}(\sqrt{n \log m})}$ time algorithm for positive (respectively, negative) reliable learning of disjunctions (respectively, conjunctions) of $m$ majorities.
- For any $d > n^{1/2}$, a $n^{O(d)}$-time algorithm with sample complexity $n^{O(n/d)}$ for agnostically learning conjunctions, and for positive reliably learning $\mathrm{poly}(n)$-term DNFs.

All of our algorithms also satisfy very strong *attribute-efficiency* properties: if the function being learned depends on only $k \ll n$ of the $n$ input variables, then the sample complexity of the algorithm depends only logarithmically on $n$, though the dependence on $k$ may be large. We defer a detailed statement of these properties until Section 3.

**Related Work**: The problem of reliable classification can be expressed as minimizing a loss function with different costs for false negative and false positive errors (see e.g., (Domingos, 1999; Elkan, 2001)). Reliable learning is also related to the Neyman-Pearson criterion from classical statistics — where it has been shown that the optimal strategy to minimize one type of errors, subject to the other type being bounded, is to threshold the ratio of the likelihoods (Neyman and Pearson, 1933). However, the main problem is *computational*; in general the loss functions with different costs from these prior works are not convex and the resulting optimization problems are intractable. The work of Kalai et al. (2012) and the present work departs from the prior work in that we focus on algorithms with both provable guarantees on their generalization error with respect to the *zero-one* loss, and bounds on their computational complexity, rather than focusing purely on statistical efficiency.

We note that the question of learning where the hypothesis is allowed to predict "don't know" or "?" has also been studied in several previous works, in contexts that are somewhat different from ours. For example, Freund et al. (2004) studied how one could tradeoff accuracy vs. the probability of predicting predicting *unknown*, when the available training data was limited. El-Yaniv and Wiener (2010) also present a mathematical formulation of predicting with unknowns, but mostly focus on the *realizable* (i.e., non-agnostic) case; moreover, they are mainly interested in settings where available data is limited. The work of Bartlett and Wegkamp (2008) is similar in spirit to ours, however they show generalization properties of a related *surrogate* loss function, which may not provide the appropriate bounds on the *zero-one* type losses considered in this paper. The work of Bshouty and Burroughs (2005) is closely related to ours, though they mainly focus on proving lower bounds for *proper learning*. Also, their results deal with multiplicative approximations rather than additive, which are less interesting in the context of agnostic learning as pointed out by Kalai et al. (2005). Finally, the phrase “reliable learning” was first introduced by Rivest and Sloan (1988)—however, the learning model considered in that paper had much more power than just access to random examples.

Kalai et al. (2012) showed that any concept class that is agnostically learnable under a fixed distribution is also learnable in the reliable agnostic learning models under the same distribution. Furthermore, they showed that if a class $C$ is agnostically learnable, the class of disjunctions of concepts in $C$ is positive reliably learnable (and the class of conjunctions of concepts in $C$ is negative reliably learnable). Finally, they showed that if $C$ is both positive and negative reliably learnable, then it is also fully reliably learnable. Using these general reductions, Kalai et al. showed that the class of polynomial-size DNF formulae is positive reliable learnable under the uniform distribution in polynomial time with membership queries (it also follows from their reductions and the agnostic learning algorithm of Kalai et al. (2005) described below that DNF formulae can be positive reliably learned in the distribution-independent setting in time $2^{\tilde{O}(\sqrt{n})}$). Agnostically learning DNFs under the uniform distribution remains a notorious open problem, and thus their work gave the first indication that positive (or negative) reliable learning may be easier than agnostic learning.

Kalai et al. (2005) put forth an algorithm for agnostic learning based on $L_1$-*regression*. Our reliable learning algorithms based on one-sided approximate degree upper bounds is inspired by and generalizes their work. Klivans and Sherstov (2010) subsequently established strong *limitations* on the $L_1$-regression approach of Kalai et al. (2005), proving lower bounds on the size of *any* set of “feature functions” that can point-wise approximate the concept classes of majorities and conjunctions. Their work implies that substantially new ideas will be required to obtain a $2^{o(n)}$-time distribution-independent agnostic learning algorithm for majorities, or a $2^{o(\sqrt{n})}$ time algorithm for agnostically learning conjunctions.

Finally, lower bounds on one-sided approximate degree have recently been used in several works to establish strong limitations on the power of existing algorithms for PAC learning (Bun and Thaler, 2013a; Sherstov, 2014; Bun and Thaler, 2013b; Gavinsky and Sherstov, 2010; Sherstov, 2013a). In this paper, we do the opposite: we use one-sided approximate degree upper bounds to give new, more efficient learning algorithms in the reliable agnostic setting.

**Organization**: In Section 2, we review the definitions of agnostic learning, and positive, negative and fully reliable learning. In Section 3, we first give a very simple polynomial time algorithm for positive reliable learning of disjunctions, before showing that appropriate one-sided polynomial approximations for function classes result in efficient reliable learning algorithms. In Section 4, we give constructions of one-sided approximating polynomials for (conjunctions and disjunctions of) low-weight halfspaces, as well as for DNF and CNF formulae. In Section 5, we show how tradeoffs may be obtained for some of our results between sample complexity and running time. We end with a discussion and some directions for future work. Due to space restrictions, we discuss limitations of our techniques in Appendix C.

## 2. Preliminaries and Definitions

Let $X = \{-1, 1\}^n$ denote the instance space. Let $C$ denote a concept class of functions from $X \to \{-1, 1\}$. We will use the convention that $+1$ is TRUE and $-1$ is FALSE.[1] For ease of

[1] This is contrary to the usual convention in the analysis of Boolean functions. However, our definitions would appear a lot more counter-intuitive in the standard notation.

notation, we will keep the parameter $n$, corresponding to the length of input vectors, implicit in the discussion. In the agnostic setting, the data may come from an arbitrary joint distribution on examples and labels. Let $D$ denote a distribution over $X \times \{-1, 1\}$. Let $h : X \to \{-1, 1\}$ be a Boolean function, let $\mathrm{err}(h, D) = \Pr_{(x,y)\sim D}[h(x) \neq y]$, denote the error of $h$ with respect to $D$, and let $\mathsf{EX}(D)$ denote the example oracle which when queried returns $(x, y) \sim D$. Since the algorithms presented in this paper typically do not run in polynomial time, we do not impose such a condition in the definitions of learnability. We will explicitly mention the running time and sample complexity in all our results.

Throughout the paper, we use $\tilde{O}$ to hide factors polylogarithmic in $n$ and $\log(1/\epsilon)$. We also define $\mathrm{sgn}(t) = -1$ for $t \leq 0$ and 1 otherwise. The notion of point-wise approximating polynomials and the results required to prove generalization bounds for our algorithms are provided in Appendix A.

**Definition 1 (Agnostic Learning (Haussler, 1992; Kearns et al., 1994))** *A concept class $C$ is agnostically learnable if there exists a learning algorithm that for any distribution $D$ over $X \times \{-1, 1\}$, any $\epsilon, \delta > 0$, with access to example oracle $\mathsf{EX}(D)$, outputs a hypothesis $h$, such that with probability at least $1 - \delta$, $\mathrm{err}(h, D) \leq \mathrm{opt} + \epsilon$, where $\mathrm{opt} = \min_{c \in C} \mathrm{err}(c, D)$.*

### 2.1. Reliable Learning

We review the various notions of reliable agnostic learning proposed by Kalai et al. (2012). As in the case of agnostic learning, the data comes from an arbitrary joint distribution $D$ over $X \times \{-1, 1\}$. For a Boolean function, $h : X \to \{-1, 1\}$, define the false positive error ($\mathrm{false}_+$) and the false negative error ($\mathrm{false}_-$) with respect to $D$ as follows:

$$\mathrm{false}_+(h, D) = \Pr_{(x,y)\sim D}[h(x) = 1 \wedge y = -1]; \qquad \mathrm{false}_-(h, D) = \Pr_{(x,y)\sim D}[h(x) = -1 \wedge y = +1]$$

Let $C$ denote the concept class of interest for learning. For a distribution $D$, define the following:

$$C^+(D) = \{c \in C \mid \mathrm{false}_+(c, D) = 0\}; \qquad C^-(D) = \{c \in C \mid \mathrm{false}_-(c, D) = 0\}$$

We call the concepts in $C^+$ (respectively, $C^-$) positive (respectively, negative) reliable with respect to $D$.[2] Positive reliable learning requires that the learning algorithm produce a hypothesis that makes (almost) no false positive errors, while simultaneously minimizing false negative errors. Likewise, in the case of negative reliable learning, the learning algorithm must output a hypothesis that makes (almost) no false negative errors, while simultaneously minimizing false positive errors. Below, we formally define positive reliable learning; the definition of negative reliable learning is symmetric and is provided in Appendix A for completeness.

**Definition 2 (Positive Reliable Learning (Kalai et al., 2012))** *A concept class $C$ is positive reliably learnable if there exists a learning algorithm that for any distribution $D$ over $X \times \{-1, 1\}$, and any $\epsilon, \delta > 0$, when given access to the example oracle $\mathsf{EX}(D)$, outputs a hypothesis $h$ that satisfies the following with probability at least $1 - \delta$, (i) $\mathrm{false}_+(h, D) \leq \epsilon$, and (ii) $\mathrm{false}_-(h, D) \leq \mathrm{opt}^+ + \epsilon$, where $\mathrm{opt}^+ = \min_{c \in C^+(D)} \mathrm{false}_-(c, D)$. We refer to $\epsilon$ as the error parameter of the learning algorithm.*

[2] It is easily ensured that $C^+$ and $C^-$ are non-empty by insisting that the constant classifiers that predict $+1$ and $-1$ everywhere are in $C$. This is the case in natural classes such as disjunctions, conjunctions and thresholds.

Kalai et al. (2012) also define a notion of *fully reliable learning*. Here, the learning algorithm may output a *partial classifier* $h : X \to \{-1, ?, +1\}$, and must make (almost) no errors, while simultaneously minimizing the probability of abstaining from prediction, i.e., outputting ?. Again, recall that we are in the agnostic setting, and let $D$ be an arbitrary distribution over $X \times \{-1, 1\}$. For some partial classifier, $h : X \to \{-1, ?, +1\}$, let $\mathrm{err}(h, D) = \Pr_{(x,y)\sim D}[h(x) = -y]$ denote the error, and $?(h, D) = \Pr_{(x,y)\sim D}[h(x) =?]$ denote the uncertainty of $h$. From a concept class $C$, each pair of concepts defines a partial classifier, $c_p = (c_+, c_-)$, defined as: $c_p(x) = c_+(x)$, if $c_+(x) = c_-(x)$, and $c_p(x) =?$ otherwise. Let $C^f(D) = \{c_p = (c_+, c_-) \mid \mathrm{err}(c_p, D) = 0\}$ denote the fully reliable partial classifiers with respect to distribution $D$. Formally, fully reliable learning is defined as:

**Definition 3 (Fully Reliable Learning (Kalai et al., 2012))** *A concept class $C$ is fully reliable learnable, if there exists a learning algorithm that for any distribution $D$ over $X \times \{-1, 1\}$, any $\epsilon, \delta > 0$, with access to the example oracle $\mathsf{EX}(D)$, outputs a partial hypothesis $h : X \to \{-1, ?, +1\}$, that satisfies the following with probability at least $1 - \delta$, (i) $\mathrm{err}(h, D) \leq \epsilon$, and (ii) $?(h, D) \leq \mathrm{opt}^? + \epsilon$, where $\mathrm{opt}^? = \min_{c_p \in C^f(D)} ?(c_p, D)$. We refer to $\epsilon$ as the error parameter of the learning algorithm.*

Kalai et al. (2012) showed the following simple result.

**Theorem 4 ((Kalai et al., 2012))** *If a concept class $C$ is positive and negative reliable learnable in time $T(n, \epsilon)$ and with sample complexity $S(n, \epsilon)$, then $C$ is fully reliable learnable in time $O(T(n, \epsilon/4))$ and sample complexity $O(S(n, \epsilon/4))$.*

### 2.2. One-sided Approximating Polynomials

We define the notion of one-sided approximating polynomials. The definitions as they are presented here essentially appeared in prior work of Bun and Thaler (2013a) (see also Sherstov (2014)), who only required the notion we refer to as positive one-sided approximate degree. Here, we explicitly distinguish between positive and negative one-sided approximations. Below, we define positive one-sided approximating polynomials; the symmetric notion of negative one-sided approximation is provided in Appendix A.

**Definition 5 (Positive One-Sided Approximating Polynomial)** *Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be a Boolean function. We say that a polynomial $p$ is a positive one-sided $\epsilon$-approximation for $f$ if $p$ satisfies the following two conditions.*

1. *For all $x \in f^{-1}(1)$, $p(x) \in [1 - \epsilon, \infty)$*
2. *For all $x \in f^{-1}(-1)$, $p(x) \in [-1 - \epsilon, -1 + \epsilon]$.*

Throughout, the degree of a multivariate polynomial refers to its total degree. We define the *positive and negative one-sided approximate degrees* of $f$, denoted $\widetilde{\deg}_{+,\epsilon}(f)$ and $\widetilde{\deg}_{-,\epsilon}(f)$ respectively, to be the minimum degree of a positive (respectively, negative) one-sided $\epsilon$-approximating polynomial $p$ for $f$. We define $\widetilde{\deg}_+ := \widetilde{\deg}_{+,1/3}$ and $\widetilde{\deg}_- := \widetilde{\deg}_{-,1/3}$, and refer to these quantities as the *positive and negative one-sided approximate degrees* of $f$ without qualification.

For a polynomial $p : \{-1, 1\}^n \to \{-1, 1\}$, we define its *weight* to be the sum of the absolute values of its coefficients and denote it by $\mathrm{weight}(p)$. Let $C$ be a concept class of Boolean functions;

```
Input: Sample $\langle (x_i, y_i) \rangle_{i=1}^m$ from $D^m$

1. Let $h = x[1] \vee \bar{x}[1] \vee \cdots x[n] \vee \bar{x}[n]$ be the disjunction that include all literals

2. For every $(x_i, y_i)$ such that $y_i = -1$, for $j = 1, \ldots, n$, modify $h$ by dropping the literal $x[j]$
   if $x_i[j] = 1$ and the literal $\bar{x}[j]$ if $x_i[j] = -1$

3. Output $h$
```

Figure 1: Algorithm: Positive Reliable Learning Disjunctions

we say that $C$ is positive one-sided $\epsilon$-approximated by degree $d$ and weight $W$ polynomials, if the following is true: for every $c \in C$, there exists a polynomial $p$ of weight at most $W$ and degree at most $d$, such that $p$ is a positive one-sided $\epsilon$-approximation of $c$. An analogous definition can be made for the negative one-sided $\epsilon$-approximation of a concept class.

## 3. Learning Algorithms

We first present a very simple algorithm for positive reliable learning disjunctions (see Fig. 1). This algorithm is essentially the same as that for PAC learning disjunctions (see Kearns and Vazirani (1994, Chap. 1)); it ignores all the positive examples, and finds a disjunction that is maximally positive and classifies all the negative examples correctly. The proof of Theorem 6 is simple and is provided in Appendix B.1.

**Theorem 6** *The algorithm in Fig. 1 positive reliably learns the class of disjunctions for some $m$ in $O(n/\epsilon^2)$, where $m$ is the number of labeled examples that the algorithm takes as input. Here, the dependence on $\delta$ is hidden.*

It is unlikely, however, that such simple reliable learning algorithms exist for richer classes. We now prove our main learning result—we describe a generic algorithm that positive reliably learns any concept class that can be positive one-sided approximated by degree $d$ and weight $W$ polynomials. The weight $W$ controls the sample complexity of the learning algorithm, and the degree $d$ controls the running time. For many natural classes, the resulting algorithm has strong attribute-efficient properties, since the weight of the approximating polynomial typically depends only on the number of *relevant* attributes.

Our algorithm extends the $L_1$-regression technique of Kalai et al. (2005) for agnostic learning, but we require a more detailed analysis. In the case of positive-reliable learning, it is required that the hypothesis output by the algorithm makes almost no false positive errors — this is enforced as constraints in a linear program. To control the false negative errors of the hypothesis, we have the objective function of the linear program minimize the *hinge loss*, which is analogous to the $L_1$ loss, but the penalty is only enforced when the prediction disagrees in sign with the true label. To bound the generalization error of the output hypothesis, we use bounds on the Rademacher complexity of the approximating polynomials (see Appendix A.5 for details).

**Theorem 7** *Let $C$ be a concept class that is positive (negative) one-sided $\epsilon$-approximated by polynomials of degree $d$ and weight $W$. Then, $C$ can be positive (negative) reliably learned by an algorithm with the following properties:*

1. *The running time of the learning algorithm is polynomial in $n^d$ and $1/\epsilon$.*
2. *The sample complexity is $m = \max\{\frac{512}{\epsilon^4} \cdot W^2 d \log(2n), \frac{64}{\epsilon^2}(W+1)^2 \log\left(\frac{1}{\delta}\right)\}$*
3. *The hypothesis output by the algorithm can be evaluated at any $x \in X$ in time $O(n^d)$.*

**Proof** We only prove the theorem for the case of positive reliable learning. The case of negative reliable learning is entirely symmetric.

**Description of Algorithm.** Suppose $D$ is an arbitrary distribution over $X \times \{-1, 1\}$ and let $\mathcal{S} = \langle (x_i, y_i) \rangle_{i=1}^m$ be a sample drawn according to $D$. The learning algorithm first solves the following mathematical program.

$$\boxed{\begin{array}{ll} \underset{p\,:\,\deg(p)\le d}{\text{minimize}} & \displaystyle\sum_{i:y_i=+1} (1 - p(x_i))_+ \\ \text{subject to} & \\ & p(x_i) \le -1 + \epsilon \quad \forall i \text{ such that } y_i = -1 \\ & \text{weight}(p) \le W. \end{array}}$$

Here $(a)_+$ denotes $a$ if $a > 0$ and $0$ otherwise. This program is similar to one used in the $L_1$-regression algorithm for agnostic learning introduced by Kalai et al. (2005). The variables of the program are the $\sum_{j=0}^{d} \binom{n}{j} = O(n^d)$ coefficients of the polynomial $p(x)$. The above mathematical program is then easily implemented as a linear program.

Let $p$ denote an optimal solution to the linear program. Define $\text{chop}(a) = a$ if $a \in [-1, 1]$ and $\text{chop}(a) = \text{sgn}(a)$ otherwise. The hypothesis output by the algorithm will be a randomized Boolean function, defined as follows:

$$h(x) = \begin{cases} +1 & \text{with probability}(1 + \text{chop}(p(x)))/2 \\ -1 & \text{with probability}(1 - \text{chop}(p(x)))/2 \end{cases}$$

**Running Time.** Since the above program can be implemented as a linear program with $O(n^d)$ variables and $O(m + n^d)$ constraints, the running time to produce the output polynomial $p$ is $\text{poly}(m, n^d)$. Note that the polynomial $p$ defines the output hypothesis $h$ completely, except for the randomness used by $h$. For any $x$, $h(x)$ can be evaluated in time $O(n^d)$ by a randomized Turing machine. Remark 8 explains how $h$ can be converted to a deterministic hypothesis.

**Generalization Error.** We will use two loss functions in our analysis. Define $\ell_+ : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}^+$ as follows:

$$\ell_+(y', +1) = 0,$$

$$\ell_+(y', -1) = \begin{cases} 0 & y' \le -1 + \epsilon \\ \frac{1}{\epsilon}(y' + 1 - \epsilon) & -1 + \epsilon < y' \le -1 + 2\epsilon \\ 1 & -1 + 2\epsilon < y' \end{cases}$$

Clearly $\ell_+$ is bounded between $[0,1]$ always and also it is $1/\epsilon$-Lipschitz. For a function, $f : X \to \mathbb{R}$, let $\mathcal{L}_+(f)$ denote the expected loss of $f$ under $D$ and the loss function $\ell_+$, and similarly let $\hat{\mathcal{L}}_+(f)$ denote the empirical loss of $f$ under $\ell_+$.

Define $\ell_- : \mathbb{R} \times \{-1,1\} \to \mathbb{R}^+$ as follows:

$$\begin{aligned}\ell_-(y', -1) &= 0,\\ \ell_-(y', +1) &= (1-y')_+\end{aligned}$$

Let $p$ continue to denote an optimal solution to the linear program. Note that since $X = \{-1,1\}^n$, and $\mathrm{weight}(p) \leq W$, it holds that $|p(x)| \leq W$ for all $x \in X$. It follows that $\ell_-(p(x), b) \leq W + 1$ for all $x \in X$ and $b \in \{-1,+1\}$. Moreover, $\ell_-$ is easily seen to be 1-Lipschitz. For a function, $f : X \to \mathbb{R}$, let $\mathcal{L}_-(f)$ and $\hat{\mathcal{L}}_-(f)$ denote the expected and empirical loss of $f$ respectively under distribution $D$ and loss function $\ell_-$.

Recall that $C^+(D) = \{c \in C \mid \mathrm{false}_+(c) = 0\}$. Let $c^* \in C^+(D)$ be an optimal positive reliable classifier, i.e., $\mathrm{false}_-(c^*) = \min_{c \in C^+(D)} \mathrm{false}_-(c)$. Let $p^* \in \mathbb{P}_{d,W}$ be a positive one-sided $\epsilon$-approximating polynomial for $c^*$ whose existence is guaranteed by hypothesis. Note that since $p^*(x) \geq 1 - \epsilon$ for $x \in (c^*)^{-1}(1)$ and $p^*(x) \in [-1-\epsilon, -1+\epsilon]$ for $x \in (c^*)^{-1}(-1)$, the following is true:

$$\begin{aligned}\mathcal{L}_+(p^*) &= 0\\ \mathcal{L}_-(p^*) &\leq 2\,\mathrm{false}_-(c^*) + \epsilon\end{aligned}$$

Here, the inequality holds because $\ell_-(y', 1) = (1-y')_+$, which is between $2-\epsilon$ and $2+\epsilon$ when $p^*(x) \in [-1-\epsilon, -1+\epsilon]$. Thus, each $x$ on which $c^*$ makes a false negative error contributes approximately 2 to $\mathcal{L}_-(p)$; the extra $\epsilon$ accounts for the approximation error.

Fix a $\delta > 0$. Recall that $\mathbb{P}_{d,W}$ is the class of degree $d$ and weight $W$ polynomials. Then the Rademacher complexity, $\mathcal{R}_m(\mathbb{P}_{d,W}) \leq W\sqrt{(2d\log(2n))/m}$ (see (2) in App. A.5). Let $\alpha = (4/\epsilon)\mathcal{R}_m(\mathbb{P}_{d,W}) + 2(W+1)\sqrt{\frac{\log(1/\delta)}{2m}}$. Recall that $p$ is the polynomial output by running the linear program. Then the following holds with probability $1-\delta$:

$$\begin{aligned}\mathcal{L}_-(p) &\leq \hat{\mathcal{L}}_-(p) + \alpha && \text{Using Theorem 23}\\ &\leq \hat{\mathcal{L}}_-(p^*) + \alpha && \text{Since } p^* \text{ is a feasible solution}\\ &\leq \mathcal{L}_-(p^*) + 2\alpha && \text{Using Theorem 23}\\ &\leq 2\,\mathrm{false}_-(c^*) + 2\alpha + \epsilon. && (1)\end{aligned}$$

Similarly, using Theorem 23 and the fact that $\hat{\mathcal{L}}_+(p) = 0$, we have that $\mathcal{L}_+(p) \leq \alpha$.

We have the following:

$$\mathbb{E}_h[\mathrm{false}_+(h)] = \mathbb{E}_{(x,y)\sim D}\mathbb{E}_h[\mathbb{I}(y=-1)\mathbb{I}(h(x)=1)] = \mathbb{E}_{(x,y)\sim D}[\mathbb{I}(y=-1)\Pr(h(x)=1 \mid p(x))].$$

The inner probability is only over the randomness used by the hypothesis $h$. It follows from the definition of the randomized hypothesis $h$ and the loss function $\ell_+$, that $\Pr(h(x) = 1 \mid p(x)) \leq \ell_+(p(x), -1) + \epsilon/2$. This together with the fact that $\ell_+(p(x), +1) = 0$ for all $x$, and $\mathcal{L}_+(p) \leq \alpha$, gives us

$$\mathbb{E}_h[\mathrm{false}_+(h)] \leq \mathbb{E}_{(x,y)\sim D}[\epsilon/2 + \ell_+(p(x), y)] \leq \epsilon/2 + \mathcal{L}_+(p) \leq \epsilon/2 + \alpha.$$

Similarly, we have the following:

$$\mathbb{E}_h[\text{false}_-(h)] = \mathbb{E}_{(x,y)\sim D}\mathbb{E}_h[\mathbb{I}(y=+1)\mathbb{I}(h(x)=-1)] = \mathbb{E}_{(x,y)\sim D}[\mathbb{I}(y=+1)\Pr(h(x)=-1 \mid p(x))]$$

Again, the inner probability is only over the randomness of the hypothesis $h$. From the definitions of $\ell_-$ and $h$, it follows that $\Pr(h(x) = -1 \mid p(x)) \leq \ell_-(p(x), +1)/2$. Using this along with the fact that $\ell_-(p(x), -1) = 0$ for all $x$, and (1) we get

$$\mathbb{E}_h[\text{false}_-(h)] \leq \mathbb{E}_{(x,y)\sim D}[\frac{1}{2}\ell_-(p(x), y)] \leq \text{false}_-(c^*) + \alpha + \epsilon/2$$

Finally, it is easily verified that for the value of $m$ in the theorem statement, $\alpha \leq \epsilon/2$. This completes the proof of the theorem. ■

**Remark 8** *The randomized hypothesis $h$ can easily be converted to a deterministic one as follows: let $H(x) = \text{chop}(p(x))$. Note that $\mathbb{E}[h(x)] = H(x)$ for each $x$. Observe that according to the choices made in Theorem 7, $\mathcal{L}_+(H) \leq \epsilon/2$. Now by definition of the loss function, $\ell_+$, it is clear that any value of $t \in [-1+\epsilon, 1]$ satisfies, $\text{false}_+(\text{sgn}(H - t)) \leq \mathcal{L}_+(H) \leq \epsilon/2$. Also, we know that $\mathcal{L}_-(H) \leq 2\,\text{false}_-(c^*) + \alpha + \epsilon$ (see Equation (1)). Thus, $\mathbb{E}_t[\text{false}_-(\text{sgn}(H - t))] \leq \text{false}_-(c^*) + O(\epsilon)$, where $t \in [-1+\epsilon, 1]$ is chosen uniformly at random. Thus, by probabilistic method, there exists $t \in [-1+\epsilon, 1]$ that simultaneously satisfies, $\text{false}_-(\text{sgn}(H - t)) \leq \epsilon$ and $\text{false}_-(\text{sgn}(H - t)) \leq \text{false}_-(c^*) + O(\epsilon)$. A suitable value of $t$ can be obtained by taking a fresh sample of size $m = O(1/\epsilon^2)$ finding the smallest value $t^*$, such that $\frac{1}{m}\sum_{y_i=-1}\mathbb{I}(h_{t^*}(x_i) = +1) \leq \epsilon$. Then, a simple VC argument implies that $h_{t^*}$ is a deterministic hypothesis with the required properties. The dependence on $\delta$ (which is logarithmic) is hidden in the preceding argument.*

Theorem 7 satisfies a strong *attribute-efficiency* property. The sample complexity depends only logarithmically on $n$, and polynomially on the weight of the polynomial approximations, which can be much smaller then $n^d$. A similar statement can also be made for agnostic learning; this observation was already implicit in some prior work (see e.g., (Feldman et al., 2013)); we state this as a theorem for completeness. Instead of the mathematical program described in the proof of Theorem 7, to obtain Theorem 9, the $L_1$-regression algorithm of Kalai et al. (2005) is directly applied, with the added constraint that the weight of the approximating polynomial is at most $W$. The rest of the proof is similar, but simpler—we only use $\ell(y', y) = |y' - y|$ as the loss function in the analysis and is omitted.

**Theorem 9** *Let $C$ be a concept class of functions from $X \to \{-1, 1\}$, such that for every $c \in C$, there exists a polynomial $p$ of degree at most $d$ and weight at most $W$, such that for all $x \in X$, $|p(x) - c(x)| \leq \epsilon$. Then, $C$ can be agnostically learned with the following properties:*

1. *The running time of the learning algorithm is polynomial in $n^d$ and $1/\epsilon$.*

2. *The sample complexity is polynomial in $W$, $\log(n)$, $\log(1/\delta)$ and $1/\epsilon$.*

3. *The hypothesis output by the algorithm can be evaluated at any $x \in X$ in time $O(n^d)$.*

## 4. One-sided Polynomial Approximations

In this section, we construct both positive and negative one-sided polynomial approximations for low-weight halfspaces, as well as positive (respectively, negative) one-sided approximations for disjunctions (respectively, conjunctions) of low-weight halfspaces. The proof of the following theorem is provided in Appendix B.2.

**Theorem 10** *Let $h(x) = \mathrm{sgn}(w_0 + \sum_{i=1}^n w_i x_i)$ denote any halfspace, where $w_i$ are integers. Let $W = \sum_{i=0}^n |w_i|$ denote the* weight *of $h$. Both $\widetilde{\deg}_{+,\epsilon}(h)$ and $\widetilde{\deg}_{-,\epsilon}(h)$ are in $\tilde{O}\left(\sqrt{W \log(1/\epsilon)}\right)$, with the relevant approximating polynomials having weight at most $\exp\left(\tilde{O}\left(\sqrt{W \log(1/\epsilon)}\right)\right)$. In particular, the majority function $\mathrm{MAJ}(x) = \mathrm{sgn}(\sum_{i=1}^n x_i)$ has both positive and negative $\epsilon$-approximating polynomials of degree at most $\tilde{O}(\sqrt{n \log(1/\epsilon)})$ and weight at most $\exp\left(\tilde{O}(\sqrt{n \log(1/\epsilon)})\right)$*

**Remark 11** *By adapting standard symmetrization arguments (cf. Buhrman et al. (1999)), the $\tilde{O}(\sqrt{n \log(1/\epsilon)})$ upper bound on $\widetilde{\deg}_{+,\epsilon}(\mathrm{MAJ})$ is easily seen to be tight up to factors hidden by the $\tilde{O}$ notation.*

Theorems 7 and 10 imply Corollary 12, which combined with Theorem 4 implies Corollary 13.

**Corollary 12** *The concept class of majorities, defined as the collection of the majority functions on each of the $2^n$ subsets of the variables, can be positive or negative reliably agnostically learned with error parameter $\epsilon$ in time $2^{\tilde{O}\left(\sqrt{n \log(1/\epsilon)}\right)}$.*

**Corollary 13** *The concept class of Majorities on $n$ variables can be fully reliably learned with error parameter $\epsilon$ in time $2^{\tilde{O}\left(\sqrt{n \log(1/\epsilon)}\right)}$.*

We now consider significantly more expressive concept classes: *disjunctions and conjunctions* of majorities. Theorem 14 is proved in Appendix B.2 and Corollary 15 follows immediately from Theorems 7, 10 and 14.

**Theorem 14** *Consider $m$ functions $f_1 \dots f_m$. Fix a $d > 0$, and suppose that each $f_i$ has a positive one-sided $(\epsilon/m)$-approximating polynomial of degree at most $d$ and weight at most $W$. Then $\mathrm{OR}_m(f_1, \dots, f_m)$ has a positive one-sided $\epsilon$-approximating polynomial of degree at most $d$ and weight at most $m \cdot W$.*

*Similarly, if each $f_i$ has a negative one-sided $(\epsilon/m)$-approximating polynomial of degree at most $d$ and weight at most $W$, then $\mathrm{AND}_m(f_1, \dots, f_m)$ has a negative one-sided $\epsilon$-approximating polynomial of degree at most $d$ and weight at most $m \cdot W$.*

**Corollary 15** *Disjunctions of $m$ Majorities can be positive reliably learned with error parameter $\epsilon$ in time $2^{\tilde{O}(\sqrt{n \log(m/\epsilon)})}$. Conjunctions of $m$ Majorities can also be negative reliably learned in the same time bound.*

## 5. Trading off Runtime for Sample Complexity

**Standard Agnostic Learning of Conjunctions**: Kalai et al. (2005) showed how to use $L_1$-regression to agnostically learn conjunctions on $n$ variables in time $2^{\tilde{O}(\sqrt{n \log(1/\epsilon)})}$. However, the sample complexity of the algorithm can also be as large as $2^{\tilde{O}(\sqrt{n \log(1/\epsilon)})}$. This result relies on the existence of $\epsilon$-approximating polynomials for the $n$-variate AND function of degree $\tilde{O}(\sqrt{n \log(1/\epsilon)})$.

Theorem 9 gives an avenue for obtaining better sample complexity, at the cost of increased runtime: if we can show that any conjunction on $n$ variables can be $\epsilon$-approximated by a degree $d$ polynomial of weight $W \ll 2^{\sqrt{n \log(1/\epsilon)}}$, then the $L_1$-regression algorithm will have sample complexity only $\mathrm{poly}(d, W)$ and runtime $n^{O(d)}$. Thus, in order to obtain tradeoffs between runtime and sample complexity for algorithms that agnostically learn conjunctions, it suffices to understand what are the achievable tradeoffs between degree and weight of $\epsilon$-approximating polynomials for the AND function.

In fact, this question is already well-understood in the case of constant $\epsilon$: letting $\mathrm{AND}_n$ denote the AND function on $n$ variables, Servedio et al. (2012) implicitly showed that for any $\sqrt{n} < d$ and any $\epsilon = \Theta(1)$, there exists an $\epsilon$-approximating polynomial for the $\mathrm{AND}_n$ function of degree $d$ and weight $\mathrm{poly}(n) \cdot 2^{\tilde{O}(n/d)}$. In fact, this construction is essentially optimal, matching a lower bound for constant $\epsilon$ proved in the same paper (see also (Bun and Thaler, 2013a, Lemma 20)). Theorem 16 (proved in Appendix B.3) extends the ideas of Servedio et al. (2012) to handle sub-constant values of $\epsilon$, as a result we obtain Corollary 17 even when $\epsilon = o(1)$.

**Theorem 16** *Fix a $d > \tilde{\Omega}\left(\sqrt{n \log n} \log(1/\epsilon)\right)$. There exists an (explicit) $\epsilon$-approximating polynomial for $\mathrm{AND}_n$ of degree $d$ and weight $2^{\tilde{O}(n \log(1/\epsilon)/d)}$.*

**Corollary 17** *For any $d > \tilde{\Omega}\left(\sqrt{n \log n} \log(1/\epsilon)\right)$ and $\epsilon$, the class of conjunctions on $n$ variables can be agnostically learned to error $\epsilon$ in time $n^{O(d)}$, with sample complexity $2^{\tilde{O}(n \log(1/\epsilon)/d)}$.*

**Positive Reliable Learning of DNFs**: As discussed earlier, the reductions of Kalai et al. (2012), combined with the agnostic learning algorithm for conjunctions due to Kalai et al. (2005), imply that DNFs can be positive reliably learned in time $2^{(\tilde{O}(\sqrt{n}))}$. However, the sample complexity of the resulting algorithm may be as large as its runtime. Here, we give an algorithm for positive reliable learning of DNFs that has smaller sample complexity, at the cost of larger runtime. Corollary 19 follows from Theorem 18 (proved in Appendix B.3).

**Theorem 18** *For any DNF $F$ of size $m$ and width (i.e., maximum term length) at most $w$, and any $d > \tilde{\Omega}\left(\sqrt{w \log w} \log(1/\epsilon)\right)$, there exists an (explicit) positive one-sided $\epsilon$-approximating polynomial for $F$ of degree $d$ and weight $2^{\tilde{O}(w \log(m/\epsilon)/d)}$. Similarly, any CNF $F$ of size $m$ and width at most $w$ has a negative one-sided $\epsilon$-approximation with the same weight and degree bounds.*

**Corollary 19** *For any $d > \tilde{\Omega}\left(\sqrt{w \log w} \log(1/\epsilon)\right)$, the concept class of DNFs of size $m$ and width at most $w$ can be positive reliably learned in time $n^{O(d)}$, using at most $2^{\tilde{O}(w \log(m/\epsilon)/d)}$ samples. The class of CNFs of size $m$ and width at most $w$ can be negative reliably learned with the same efficiency guarantees.*

## 6. Discussion

We have shown that concept classes with low one-sided approximate degree can be efficiently learned in the reliable agnostic model. As we have seen, one-sided approximate degree is an intermediate notion that lies between threshold degree and approximate degree; we have identified important concept classes, such as majorities and intersections of majorities, whose one-sided approximate degree is strictly smaller than its approximate degree. Consequently, we have obtained reliable (in some cases, even *fully reliable*) agnostic learning algorithms that are strictly more efficient than the fastest known agnostic ones. We have thereby given the first evidence that even fully reliable agnostic learning may be strictly easier than agnostic learning.

The notion of one-sided polynomial approximation has only been introduced very recently (Bun and Thaler (2013a)), and previously had only been used to prove lower bounds. By giving the first *algorithmic* application of one-sided polynomial approximations, our work lends further credence to the notion that these approximations are fundamental objects worthy of further study in their own right. Just as threshold degree and approximate degree have found applications (both positive and negative) in many domains outside of learning theory, we hope that one-sided polynomial approximations will as well. Identifying such applications is a significant direction for further work.

Our work does raise several open questions specific to one-sided polynomial approximations. Here we highlight two. We have shown that halfspaces of weight at most $W$ have one-sided approximate degree $\tilde{O}(W^{1/2})$, and yet there exist halfspaces with one-sided approximate degree $\Omega(n)$. However, the (non-explicit) halfspace from Sherstov (2013b) that we used to demonstrate the $\Omega(n)$ lower bound has weight $2^{\Omega(n)}$ (see Appendix C). Is it possible that all halfspaces of weight $2^{O(n^{1-\delta})}$ for some $\delta > 0$ always have one-sided approximate degree $o(n)$? We also showed how to obtain tradeoffs between the weight and degree of one-sided polynomial approximations for DNFs. Is it possible to obtain similar tradeoffs for majorities?

## Acknowledgments

VK is supported by a Simons Postdoctoral Fellowship. JT is supported by a research fellowship from the Simons Institute for the Theory of Computing. The research was carried out while the authors were at the Simons Institute for the Theory of Computing at the University of California, Berkeley.

## References

Scott Aaronson and Yaoyun Shi. Quantum lower bounds for the collision and the element distinctness problems. *J. ACM*, 51(4):595–605, 2004.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

Nader H Bshouty and Lynn Burroughs. Maximizing agreements with one-sided error with applications to heuristic learning. *Machine Learning*, 59(1-2):99–123, 2005.

Harry Buhrman, Richard Cleve, Ronald de Wolf, and Christof Zalka. Bounds for small-error and zero-error quantum algorithms. In *FOCS*, pages 358–368. IEEE Computer Society, 1999.

Mark Bun and Justin Thaler. Hardness amplification and the approximate degree of constant-depth circuits. 2013a.

Mark Bun and Justin Thaler. Dual lower bounds for approximate degree and markov-bernstein inequalities. In Fedor V. Fomin, Rusins Freivalds, Marta Z. Kwiatkowska, and David Peleg, editors, *ICALP (1)*, volume 7965 of *Lecture Notes in Computer Science*, pages 303–314. Springer, 2013b. ISBN 978-3-642-39205-4.

E.W. Cheney. *Introduction to Approximation Theory*. AMS Chelsea Publishing Series. AMS Chelsea Pub., 1982. ISBN 9780821813744.

Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *KDD'99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, New York, NY, USA, 1999. ACM.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI'01: Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

Vitaly Feldman, Pravesh Kothari, and Jan Vondrák. Representation, approximation and learning of submodular functions using low-rank decision trees. In *Proceedings of the Conference on Learning Theory (COLT)*, 2013.

Yoav Freund, Yishay Mansour, and Robert E. Schapire. Generalization bounds for averaged classifiers. *Annals of Statistics*, 32(4):1698–1722, 2004.

Dmitry Gavinsky and Alexander A. Sherstov. A separation of np and conp in multiparty communication complexity. *Theory of Computing*, 6(1):227–245, 2010.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. ISSN 0890-5401.

Jeff Kahn, Nathan Linial, and Alex Samorodnitsky. Inclusion-exclusion: Exact and approximate. *Combinatorica*, 16(4):465–477, 1996.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *FOCS*, pages 11–20. IEEE Computer Society, 2005. ISBN 0-7695-2468-0.

Adam Tauman Kalai, Varun Kanade, and Yishay Mansour. Reliable agnostic learning. *Journal of Computer and System Sciences*, 8(5), 2012. Special Issue on Learning Theory. Earlier version appeared in Conference of Learning Theory 2009.

Michael Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. In *Machine Learning*, pages 341–352, 1994.

Michael J. Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

Adam R. Klivans and Rocco A. Servedio. Learning DNF in time $2^{\tilde{o}(n^{1/3})}$. *J. Comput. Syst. Sci.*, 68(2):303–318, 2004.

Adam R. Klivans and Alexander A. Sherstov. Lower bounds for agnostic learning via approximate rank. *Computational Complexity*, 19(4):581–604, 2010.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests for statistical hypotheses. *Philos. Trans. R. So. Lond. Ser. A Contain. Pap. Math. Phys. Character*, 231:281–337, 1933.

Ronald L. Rivest and Robert H. Sloan. Learning complicated concepts reliably and usefully. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 69–79, 1988.

T.J. Rivlin. *An Introduction to the Approximation of Functions*. Blaisdell book in numerical analysis and computer science. Dover Publications, 1981. ISBN 9780486640693.

Rocco A. Servedio, Li-Yang Tan, and Justin Thaler. Attribute-efficient learning and weight-degree tradeoffs for polynomial threshold functions. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *COLT*, volume 23 of *JMLR Proceedings*, pages 14.1–14.19. JMLR.org, 2012.

A. A. Sherstov. Breaking the Minsky-Papert barrier for constant-depth circuits. In *STOC*, 2014.

Alexander A. Sherstov. Approximating the and-or tree. *Theory of Computing*, 9(20):653–663, 2013a.

Alexander A. Sherstov. Optimal bounds for sign-representing the intersection of two halfspaces by polynomials. *Combinatorica*, 33(1):73–96, 2013b.

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

## Appendix A. Definitions and Preliminaries

Here, we provide the definitions and notation omitted from Section 2.

### A.1. PAC Learning

For a distribution $\mu$ over $X$, let $\mathrm{err}(h, (\mu, c)) = \Pr_{x\sim\mu}[c(x) \neq h(x)]$, denote the error of hypothesis $h$ with respect to concept $c$ and distribution $\mu$. Let $\mathsf{EX}(c, \mu)$ denote the example oracle, which when queried returns a pair $(x, c(x))$, where $x$ is drawn from distribution $\mu$, and $c$ is a concept in $C$.

**Definition 20 (PAC Learning (Valiant, 1984))** *A concept class $C$ is probably approximately correct (PAC) learnable if there exists a learning algorithm that for any $c \in C$, any distribution $\mu$ over $X$, any $\epsilon, \delta > 0$, with access to an example oracle $\mathsf{EX}(c, \mu)$, outputs a hypothesis $h$, such that with probability at least $1 - \delta$, $\mathrm{err}(h, (\mu, c)) \leq \epsilon$.*

### A.2. Negative Reliable Learning

**Definition 21 (Negative Reliable Learning (Kalai et al., 2012))** *A concept class $C$ is negative reliably learnable, if there exists a learning algorithm that for any distribution $D$ over $X \times \{-1, 1\}$, any $\epsilon, \delta > 0$, with access to the example oracle $\mathsf{EX}(D)$, outputs a hypothesis $h$, that satisfies the following with probability at least $1 - \delta$,*

1. $\mathrm{false}_-(h, D) \leq \epsilon$
2. $\mathrm{false}_+(h, D) \leq \mathrm{opt}^- + \epsilon$, *where* $\mathrm{opt}^- = \min\limits_{c \in C^-(D)} \mathrm{false}_+(c, D)$

*We refer to $\epsilon$ as the* error parameter *of the learning algorithm.*

### A.3. Approximating Polynomials

If $p : \{-1, 1\}^n \to \mathbb{R}$ is a real polynomial, $\deg(p)$ will denote the total degree of $p$. Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be a Boolean function. We say that a polynomial $p : \{-1, 1\}^n \to \{-1, 1\}$ is an $\epsilon$-approximation for $f$ if $|p(x) - f(x)| \leq \epsilon$ for all $x \in \{-1, 1\}^n$. We let $\widetilde{\deg}_\epsilon(f)$ denote the least degree of an $\epsilon$-approximation for $f$. We define $\widetilde{\deg}(f) = \widetilde{\deg}_{1/3}(f)$ and refer to the *approximate degree* of $f$ without qualification. The constant $1/3$ is arbitrary and is chosen by convention.

### A.4. Negative One-Sided Approximating Polynomials

**Definition 22 (Negative One-Sided Approximating Polynomial)** *Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be a Boolean function. We say that a polynomial $p$ is a negative one-sided $\epsilon$-approximation for $f$ if $p$ satisfies the following two conditions.*

1. *For all $x \in f^{-1}(-1)$, $p(x) \in (-\infty, -1 + \epsilon]$*
2. *For all $x \in f^{-1}(1)$, $p(x) \in [1 - \epsilon, 1 + \epsilon]$.*

### A.5. Generalization Bounds

We review the basic results required to bound the generalization error of our algorithms for reliable agnostic learning. Let $\mathcal{F} : X \to \mathbb{R}$ be a function class. Let $\epsilon_1, \ldots, \epsilon_n$ independently take values in $\{-1, +1\}$ with equal probability, and let the variables $x_1, \ldots, x_n$ be chosen i.i.d. from some distribution $\mu$ over $X$. Then the Rademacher complexity of $\mathcal{F}$, denoted $\mathcal{R}_m(\mathcal{F})$, is defined as:

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{m} f(x_i)\epsilon_i\right],$$

Rademacher complexities have been widely used in the statistical learning theory literature to obtain bounds on generalization error. Here, we only cite results that are directly relevant to our work. Suppose $D$ is some distribution over $X \times \{-1, 1\}$. Let $\ell : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}^+$ be a loss function. For a function, $f : X \to \mathbb{R}$, the expected loss is given by $\mathcal{L}(f) = \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)]$. For a sample, $\langle (x_i, y_i) \rangle_{i=1}^m$, let $\hat{\mathcal{L}}(f) = \frac{1}{m} \sum_{i=1}^m \ell(f(x_i), y_i)$ denote the empirical loss. Bartlett and Mendelson (2002) proved the following result:

**Theorem 23 ((Bartlett and Mendelson, 2002))** *Let $\ell$ be a Lipschitz loss function (with respect to its first argument) with Lipschitz parameter $L$, and suppose that $\ell$ is bounded above by $B$. Then for any $\delta > 0$, with probability at least $1 - \delta$ (over the random sample draw), simultaneously for all $f \in \mathcal{F}$,*

$$|\mathcal{L}(f) - \hat{\mathcal{L}}(f)| \le 4L\mathcal{R}_m(\mathcal{F}) + 2B\sqrt{\frac{\log(1/\delta)}{2m}},$$

*where $\mathcal{R}_m(\mathcal{F})$ is the Rademacher complexity of the function class $\mathcal{F}$, and $m$ is the sample size.*

Finally, let $X = \{-1, 1\}^n$ and let $\mathbb{P}_{d,W}$ be the class of $n$-variate polynomials of degree at most $d$ and weight at most $W$. Observe that for $x \in X$, $\|x\|_\infty \le 1$. Note that we can view $p(x)$ as a linear function in an expanded feature space of dimension $n^d$, and the 1-norm of $p$ in such a space is bounded by $W$. Kakade et al. (2008) proved the following result:

**Theorem 24 ((Kakade et al., 2008))** *Let $X$ be an $n$ dimensional instance space and $\mathcal{W} = \{w \mid w(x) \mapsto w \cdot x\}$ be a class of linear functions, such that for each $x \in X$, $\|x\|_\infty \le 1$, and for each $w \in \mathcal{W}$, $\|w\|_1 \le W$, then, $\mathcal{R}_m(\mathcal{W}) \le W\sqrt{\frac{2\log(2n)}{m}}$.*

In our setting, the above implies that the Rademacher complexity of $\mathbb{P}_{d,W}$ is bounded as follows:

$$\mathcal{R}_m(\mathbb{P}_{d,W}) \le W\sqrt{\frac{2d\log(2n)}{m}}. \qquad (2)$$

## Appendix B. Omitted Proofs

### B.1. Omitted Proofs from Section 3

**Proof** (of Theorem 6) Let DISJ denote the class of disjunctions and let $D$ be the distribution over $X \times \{-1, 1\}$. It is known that VC-DIM(DISJ) $= n$, and hence for some $m = O(n/\epsilon^2)$, the following is true for every $c \in \text{DISJ}$:

$$|\,\text{false}_+(c; D) - \frac{1}{m} \sum_{i:y_i=-1} \mathbb{I}(c(x_i) = +1)| \le \epsilon/2,$$
$$|\,\text{false}_-(c; D) - \frac{1}{m} \sum_{i:y_i=+1} \mathbb{I}(c(x_i) = -1)| \le \epsilon/2.$$

Recall that $\text{DISJ}^+(D)$ denotes the positive reliable disjunctions for distribution $D$. Let $c_+^* \in \text{DISJ}^+(D)$ be such that $\text{false}_-(c_+^*) = \min_{c \in \text{DISJ}^+(D)} \text{false}_-(c)$. Both $h$ and $c_+^*$ classify all the negative examples in the sample correctly; since $h$ is chosen to have the largest number of literals subject to this property, it is the case that $(1/m) \sum_{i:y_i=+1} \mathbb{I}(h(x_i) = -1) \le \sum_{i:y_i=+1} \mathbb{I}(c_+^*(x_i) = -1)$. Then, we have

$$\text{false}_+(h) \le \frac{1}{m} \sum_{i:y_i=-1} \mathbb{I}(h(x_i) = +1) + \epsilon/2, = 0 + \epsilon/2 \le \epsilon$$
$$\text{false}_-(h) \le \frac{1}{m} \sum_{i:y_i=+1} \mathbb{I}(h(x_i) = -1) + \epsilon/2.$$
$$\le \frac{1}{m} \sum_{i:y_i=+1} \mathbb{I}(c_+^*(x_i) = -1) + \epsilon/2 \le \text{false}_-(c_+^*) + \epsilon$$

■

### B.2. Omitted Proofs from Section 4

**Proof** (of Theorem 10) We begin with the case of constant $\epsilon$; i.e., we first show that for $\epsilon = 1/4$, $\widetilde{\deg}_{+,\epsilon}(h)$ and $\widetilde{\deg}_{-,\epsilon}(h)$ are in $O(W^{1/2})$. We use the following standard properties of the Chebyshev polynomials (cf. the standard texts Cheney (1982) and Rivlin (1981)).

**Fact 1** *The $d$'th Chebyshev polynomial of the first kind, $T_d(t) : \mathbb{R} \to \mathbb{R}$ has degree $d$ and satisfies*

$$|T_d(t)| \le 1 \textit{ for all } -1 \le t \le 1. \tag{3}$$
$$2 \le T_{\lceil a \rceil}(1 + 1/a^2) \textit{ for all } a \ge 1. \tag{4}$$
$$T_d(t) \textit{ is non-decreasing on the interval } [1, \infty]. \tag{5}$$
$$\textit{All coefficients of } T_d \textit{ are bounded in absolute value by } 3^d. \tag{6}$$

Let $d = \lceil W^{1/2} \rceil$. Consider the univariate polynomial $G(t) = T_d(2t/W + 1)$. Then $G$ satisfies the following properties.

$$G(t) \in [-1, 1] \text{ for all } t \in [-W, 0]. \tag{7}$$
$$G(t) \ge 2 \text{ for all } t \in [1, \infty]. \tag{8}$$

Indeed, Property 7 follows from Property 3, while Property 8 follows from Properties 4 and 5.

Now consider the univariate polynomial $P(t) = G(t)^4/4 - 1$. It is straightforward to check that

$$P(t) \in [-3/4, 1] \text{ for all } t \in [-W, 0]. \tag{9}$$
$$P(t) \geq 3 \text{ for all } t \in [1, \infty]. \tag{10}$$

Finally, consider the $n$-variate polynomial $p : \{-1,1\}^n \to \mathbb{R}$ defined via

$$p(x) = P(w_0 + \sum_{i=1}^{n} w_i x_i).$$

Combining the fact that $\sum_{i=0}^{n} |w_i| \leq W$ with Properties 9 and 10, we see that $p$ is a positive one-sided $1/4$-approximation for $h$. Moreover, $\deg(p) \leq \deg(P) = O(W^{1/2})$, and the weight of $p$ is at most $W^{O(\sqrt{W})}$. Similarly, $-p(-x)$ is a negative one-sided $1/4$-approximation for $h$. This completes the proof for $\epsilon = 1/4$.

The construction for $\epsilon = o(1)$ is somewhat more complicated. For any $k \geq 1$ and any $W$, Kahn et al. (1996) construct a univariate polynomial $S_k$ satisfying the following properties:

$$\deg(S_k) \leq k. \tag{11}$$
$$S_k(t) \geq 1 \text{ for all } t \geq W. \tag{12}$$
$$S_k(t) \leq \exp\left(-\Omega(k^2/W \log W)\right) \text{ for all } t \in \{0, \ldots, W-1\}. \tag{13}$$
$$\text{All coefficients of } S_k(t) \text{ are bounded in absolute value by } W^{O(k)}. \tag{14}$$

For completeness, we give the details of this construction and a proof of Properties 11-14 in Appendix D.

For any $\epsilon > 0$, let $k = \lceil (W \log W \log(1/\epsilon))^{1/2} \rceil$, and let $q : \{-1,1\}^n \to \mathbb{R}$ denote the $n$-variate polynomial defined via

$$q(x) = S_k\left(W + w_0 + \sum_{i=1}^{n} w_i x_i\right).$$

It is then straightforward to check that $q$ is a positive one-sided $\epsilon$-approximation for $h$ of degree at most $k = \tilde{O}\left(\sqrt{W \log(1/\epsilon)}\right)$ and weight at most $W^{\tilde{O}(k)}$. Similarly, $-q(-x)$ is a negative one-sided $\epsilon$-approximation for $h$. This completes the proof. ■

**Proof** (of Theorem 14) We prove the statement about $\mathrm{OR}_m(f_1, \ldots, f_m)$; the statement about $\mathrm{AND}_m(f_1, \ldots, f_m)$ is analogous. Let $p_i$ be a positive one-sided $(\epsilon/m)$-approximating polynomial for $f_i$. Then $p = -1 + \sum_{i=1}^{m}(1 + p_i)$ is a positive one-sided $\epsilon$-approximating polynomial for $f$. Moreover, the degree of $p$ is at most $\max_i\{\deg(p_i)\} \leq d$, while the weight of $p$ is at most $m \cdot W$. This completes the proof. ■

### B.3. Omitted Proofs from Section 5

**Proof** (of Theorem 16) We write $\mathrm{AND}_n$ as an "and-of-ands", where the outer AND has fan-in $t$, and the inner ANDs each have fan-in $n/t$, where we choose $t$ such that $t/\log t = n^2 \log(1/\epsilon)/d^2$. That is, we write $\mathrm{AND}_n(x) = \mathrm{AND}_t(\mathrm{AND}_{n/t}(x^{(1)}), \ldots, \mathrm{AND}_{n/t}(x^{(t)}))$, where $x^{(i)} = (x_{n\cdot(i-1)/t+1}, \ldots, x_{n\cdot i/t})$ denotes the $i$th "block" of variables in $x$. Note that $t \leq n$ by the assumption that $d > \tilde{\Omega}\left(\sqrt{n \log n} \log(1/\epsilon)\right)$.

We obtain an $\epsilon$-approximating polynomial $p$ for $\mathrm{AND}_n$ as follows. Kahn et al. (1996) gave an explicit $\epsilon$-approximating polynomial $p_t$ for $\mathrm{AND}_t$ of degree $d' = O(\sqrt{t \log t} \log(1/\epsilon))$. It is an immediate consequent of Parseval's inequality that $p_t$ has weight at most $t^{d'/2}$. We will also need the following standard fact.

**Fact 2** *The real polynomial $q : \{-1,1\}^{n/t} \to \{-1,1\}$ defined via $q(y_1, \ldots y_{n/t}) = 2\prod_{i=1}^{n/t} \frac{1+y_i}{2} - 1$ computes $\mathrm{AND}_{n/t}(x)$. Moreover, $q$ has degree at most $n/t$ and weight at most $3$.*

Finally, we define $p(x) = p_t(q(x^{(1)}), \ldots q(x^{(t)}))$. Notice that $p$ has degree at most $d' \cdot n/t = O\left(\sqrt{t \log t} \log(1/\epsilon) \cdot n/t\right) = O\left(n\sqrt{\log t} \log(1/\epsilon)/t\right) = O(d)$ and weight at most $t^{O(d')} = 2^{\tilde{O}(n \log(1/\epsilon)/d)}$ as claimed. ■

**Proof** (of Theorem 18) We prove the result for DNFs; the case of CNFs is analogous. Let $C_i$ denote the $i$th clause of $F$. Since $C_i$ has width at most $w$, Theorem 16 implies the existence of an $\epsilon/m$-approximating polynomial $p_i$ for $C_i$ of degree $d$ and weight at most $2^{\tilde{O}(w \log(m/\epsilon)/d)}$. Then $p = -1 + \sum_{i=1}^m (1 + p_i)$ is a positive one-sided $\epsilon$-approximating polynomial for $F$. Moreover, the degree of $p$ is at most $\max_i\{\deg(p_i)\} \leq d$, while the weight of $p$ is at most $m \cdot 2^{\tilde{O}(w \log(m/\epsilon)/d)} = 2^{\tilde{O}(w \log(m/\epsilon)/d)}$. This completes the proof. ■

## Appendix C. Limitations of Our Techniques

### C.1. On Halfspaces

Theorem 10 establishes that all *low-weight* halfspaces (i.e., weight $o(n^{2-\delta})$ for some $\delta > 0$) can be (both positive and negative) reliably learned in time $2^{o(n)}$. It is reasonable to ask whether we can reliably learn *all* halfspaces in time $2^{o(n)}$ using our techniques. Unfortunately, the answer is no.

**Theorem 25** *There exists a halfspace $h$ for which $\widetilde{\deg}_{+,1/8}(h)$ and $\widetilde{\deg}_{-,1/8}(h)$ are both $\Omega(n)$.*

**Proof** We prove the statement about $\widetilde{\deg}_{+,1/4}$, as the case of $\widetilde{\deg}_{-,1/4}$ is similar.

Given a Boolean function $h : \{-1,1\}^n \to \{-1,1\}$, let $g_h : \{-1,1\}^{2n} \to \{-1,1\}$ denote the function $g(x,y) = h(x_1) \cap h(x_2)$, where $x_1, x_2 \in \{-1,1\}^n$. That is, $g$ computes the intersection of two copies of $h$, where the two copies are applied to disjoint sets of input variables. Sherstov (2013b) proved that there exists a halfspace $h$ such that $\deg_{\pm}(g) = \Omega(n)$. Here, $\deg_{\pm}(g)$ denotes the least degree of a real polynomial $p$ that agrees in sign with $g$ at all Boolean inputs. Notice $\deg_{\pm}(g) \leq \deg_{+,\epsilon}(g)$ for any function $g$ and any $\epsilon < 1$.

Combining Sherstov's lower bound with Theorem 14 implies that $\Omega(n) = \deg_{\pm}(g) \leq \widetilde{\deg}_{+,1/4}(g) \leq \deg_{+,1/8}(h)$. This completes the proof. ■

### C.2. On DNFs

All polynomial-sized DNFs can be positive reliably learned in time and sample complexity $2^{\tilde{O}(\sqrt{n})}$, and Corollary 19 shows how to obtain smooth tradeoffs between runtime and sample complexity for this learning task. It is natural to ask whether DNFs can be *negative* reliably learned with similar efficiency using our techniques. Unfortunately, this is not the case. Bun and Thaler (2013a), extending a seminal lower bound of Aaronson and Shi (2004), showed that there is a polynomial-sized DNF $f$ (more specifically, $f$ is the negation of the ELEMENT DISTINCTNESS function) satisfying $\widetilde{\deg}_-(f) = \Omega((n/\log n)^{2/3})$; thus, our techniques cannot negative reliably learn polynomial-sized DNFs in time better than $\exp\left(\tilde{O}\left(n^{2/3}\right)\right)$.

While Bun and Thaler's is the best-known lower bound on the negative one-sided approximate degree of any polynomial-sized DNF – indeed, up to polylogarithmic factors, it is the best-known lower bound for any function in $\mathrm{AC}^0$ – no $o(n)$ upper bound is known for the negative one-sided approximate degree of polynomial-sized DNFs.

## Appendix D. Missing Details For Theorem 10

For any $k > 0$, Kahn et al. (1996) define the polynomial $S_k(t)$ as follows (in the below, $a, b$, and $r$ are parameters that Kahn et al. ultimately set to $a = \Theta(k/\log W)$, $b = \Theta(k^2/(W \log W))$, and $r = k - a - b$).

$$S_k(t) = C^{-1} \cdot \left( \prod_{i=0}^{a} (t-i) \cdot \prod_{j=W-b}^{W} (t-j) \right) \cdot T_r\left(\frac{t-a}{W-b-a}\right), \tag{15}$$

where $C = \left(\prod_{i=0}^{a}(W-i) \cdot \prod_{j=W-b}^{W}(W-j)\right) \cdot T_r\left(\frac{W-a}{W-b-a}\right)$ is a normalization constant chosen so that $S_k(W) = 1$, and as usual $T_r$ denotes the $r$'th Chebyshev polynomial of the first kind.

We now verify that $S_k$ satisfies Properties 11-14, which we restate here for the reader's convenience.

Property 11: $\deg(S_k) \leq k$.

Property 12: $S_k(t) \geq 1$ for all $t \geq W$.

Property 13: $S_k(t) \leq \exp\left(-\Omega(k^2/W \log W)\right)$ for all $t \in \{0, \ldots, W-1\}$.

Property 14: All coefficients of $S_k(t)$ are bounded in absolute value by $W^{O(k)}$.

Property 11 is immediate from the definition of $S_k$ and the choice of $r = k - a - b$.

To see that Property 12 holds, we note that $S_k(W) = 1$. The property will therefore follow if we can prove that

$$S_k(t) \geq S_k(W) \text{ for all } t \geq W. \tag{16}$$

To establish Equation (16), note first that $T_r$ is non-decreasing on the interval $[1, \infty]$ (cf. Property 5). Second, notice that $\frac{t-a}{W-b-a}$ is an increasing function on $[W, \infty]$, and is also larger than 1 on this interval. Thus, $T_r(\frac{t-a}{W-b-a})$ is a non-decreasing function in $t$ for $t \in [W, \infty]$. Finally, it is an easy observation that $\prod_{i=0}^{a}(t-i) \cdot \prod_{j=W-b}^{W}(t-j)$ is a non-decreasing function in $t$ on the interval

$t \in [W, \infty]$. Thus, $\prod_{i=0}^{a}(t-i) \cdot \prod_{j=W-b}^{W}(t-j) \cdot T_r(\frac{t-a}{W-b-a})$ is a non-decreasing function of $t$ on the same interval, and Equation (16) follows.

Property 13 is immediate from the analysis of Kahn et al. (1996). To see that Property 14 holds, note that $\prod_{i=0}^{a}(t-i) \prod_{j=W-b}^{W}(t-j)$ is a polynomial in $t$ with coefficients all bounded in absolute value by $W^{a+b} \leq W^k$, while $T_r(\frac{t-a}{W-b-a})$ is also a polynomial in $t$, with coefficients bounded in absolute value by $(3+a)^r \leq W^k$ (cf. Property 6).