Journal of Machine Learning Research 16 (2015) 3455-3467 Submitted 9/14; Revised 2/15; Published 12/15

Agnostic Learning of Disjunctions on Symmetric Distributions

Vitaly Feldman[*] VITALY@POST.HARVARD.EDU
IBM Research - Almaden
San Jose, CA

Pravesh Kothari[†] KOTHARI@CS.UTEXAS.EDU
Department of Computer Science
The University of Texas at Austin
Austin, TX

Editor: Nathan Srebro

Abstract

We consider the problem of approximating and learning disjunctions (or equivalently, conjunctions) on symmetric distributions over $\{0,1\}^n$. Symmetric distributions are distributions whose PDF is invariant under any permutation of the variables. We prove that for every symmetric distribution $\mathcal{D}$, there exists a set of $n^{O(\log(1/\epsilon))}$ functions $\mathbb{S}$, such that for every disjunction c, there is function p, expressible as a linear combination of functions in $\mathbb{S}$, such that p ϵ-approximates c in ℓ_1 distance on $\mathcal{D}$ or $\mathbf{E}_{x\sim\mathcal{D}}[|c(x)-p(x)|] \leq \epsilon$. This implies an agnostic learning algorithm for disjunctions on symmetric distributions that runs in time $n^{O(\log(1/\epsilon))}$. The best known previous bound is $n^{O(1/\epsilon^4)}$ and follows from approximation of the more general class of halfspaces (Wimmer, 2010). We also show that there exists a symmetric distribution $\mathcal{D}$, such that the minimum degree of a polynomial that 1/3-approximates the disjunction of all n variables in ℓ_1 distance on $\mathcal{D}$ is $\Omega(\sqrt{n})$. Therefore the learning result above cannot be achieved via ℓ_1-regression with a polynomial basis used in most other agnostic learning algorithms.

Our technique also gives a simple proof that for any product distribution $\mathcal{D}$ and every disjunction c, there exists a polynomial p of degree $O(\log(1/\epsilon))$ such that p ϵ-approximates c in ℓ_1 distance on $\mathcal{D}$. This was first proved by Blais et al. (2008) via a more involved argument.

Keywords: agnostic learning, symmetric distribution, polynomial approximation, regression, disjunction, conjunction, DNF, decision tree

1. Introduction

The goal of an agnostic learning algorithm for a concept class $\mathcal{C}$ is to produce, for any distribution on examples, a hypothesis h whose error on a random example from the distribution is close to the best possible by a concept from C. This model reflects a common empirical approach to learning, where few or no assumptions are made on the process that generates the examples and a limited space of candidate hypothesis functions is searched in an attempt to find the best approximation to the given data.

∗. Corresponding author.
†. Work done while the author was at IBM Research - Almaden.

Agnostic learning of disjunctions (or, equivalently, conjunctions) is a fundamental question in learning theory and a key step in learning algorithms for other concept classes such as DNF formulas and decision trees. Algorithms for this problem, such as the Set Covering Machine (Marchand and Shawe-Taylor, 2002), are also used in practical applications. There is no known efficient algorithm for the problem, in fact the fastest algorithm that does not make any distributional assumptions runs in $2^{\tilde{O}(\sqrt{n})}$ time (Kalai et al., 2008). Polynomial-time learnability is only known when the examples are very close to being consistent with some disjunction (Awasthi et al., 2010).

While the problem appears to be hard, strong hardness results are known only if the hypothesis is restricted to be a disjunction or a linear threshold function (Ben-David et al., 2003; Bshouty and Burroughs, 2006; Feldman et al., 2009, 2012), or for learning using ℓ_1-regression (Klivans and Sherstov, 2010). Weaker, quasi-polynomial lower bounds are known assuming hardness of learning sparse parities with noise (see Section 5) and, very recently, hardness of refuting random SAT formulas (Daniely and Shalev-Shwartz, 2014). It is also well-known that distribution-independent agnostic learning of disjunctions implies PAC learning of DNF expressions (Kearns et al., 1994). Finally, agnostic learning of disjunctions is known to be closely related to the problem of differentially-private release of answers to conjunctive queries (Gupta et al., 2011).

We consider this problem with an additional assumption that example points are distributed according to a symmetric or a product distribution. Symmetric and product distributions are two incomparable classes of distributions that generalize the well-studied uniform distribution. Theoretical study of learning over symmetric distributions was first done by Wimmer (2010) who gave $n^{O(1/\epsilon^4)}$ time agnostic learning algorithm for the class of halfspaces. Agnostic learning of disjunctions over symmetric distributions on $\{0,1\}^n$ also arises naturally in the well-studied problem of privately releasing answers to all short conjunction queries with low average error (Feldman and Kothari, 2014).

1.1 Our Results

We prove that disjunctions (and conjunctions) are learnable agnostically over any symmetric distribution in time $n^{O(\log(1/\epsilon))}$. This matches the well-known upper bound for the uniform distribution. Our proof is based on ℓ_1-approximation of any disjunction by a linear combination of functions from a fixed set of functions. Such approximation directly gives an agnostic learning algorithm via ℓ_1-regression based approach introduced by Kalai et al. (2008).

A natural and commonly used set of basis functions is the set of all monomials on $\{0,1\}^n$ of some bounded degree. It is easy to see that on product distributions with constant bias, disjunctions longer than some constant multiple of $\log(1/\epsilon)$ are ϵ-close to the constant function 1. Therefore, polynomials of degree $O(\log(1/\epsilon))$ suffice for ℓ_1 (or ℓ_2) approximation on such distributions. This simple argument does not work for general product distributions. However it was shown by Blais et al. (2008) that the same degree (up to a constant factor) still suffices in this case. Their argument is based on the analysis of noise sensitivity under product distributions and implies additional interesting results.

Interestingly, it turns out that low-degree polynomials cannot be used to obtain the same result for all symmetric distributions: we show that there exists a symmetric distribution for which disjunctions are no longer ℓ_1-approximated by low-degree polynomials.

Theorem 1 *There exists a symmetric distribution $\mathcal{D}$ such that for $c = x_1 \vee x_2 \vee \cdots \vee x_n$, any polynomial p that satisfies $\mathbf{E}_{x\sim\mathcal{D}}[|c(x) - p(x)|] \leq 1/3$ is of degree $\Omega(\sqrt{n})$.*

To prove this, we consider the standard linear program to find the coefficients of a degree r polynomial that minimizes pointwise error with the disjunction c. The key idea is to observe that an optimal point for the dual can be used to obtain a distribution on which the ℓ_1 *error* of the best fitting polynomial p for c is same as the value of minimum *pointwise error* of any degree r polynomial with respect to c. When c is a symmetric function, one can further observe that the distribution so obtained is in fact symmetric. Combined with the degree lower bound for uniform approximation by polynomials by Klivans and Sherstov (2010), we obtain the result. The details of the proof appear in Section 3.1.

Our approximation for general symmetric distributions is based on a proof that for the special case of the uniform distribution on S_r (the points from $\{-1,1\}^n$ with Hamming weight r), low-degree polynomials still work, namely, for any disjunction c, there is a polynomial p of degree at most $O(\log(1/\epsilon))$ such that the ℓ_1 error $\mathbf{E}_{x\sim S_r}[|c(x) - p(x)|] \leq \epsilon$.

Theorem 2 *For $r \in \{0,\ldots,n\}$, let S_r denote the set of points in $\{0,1\}^n$ that have exactly r 1's and let $\mathcal{D}_r$ denote the uniform distribution on S_r. For every disjunction c and $\epsilon > 0$, there exists a polynomial p of degree at most $O(\log(1/\epsilon))$ such that $\mathbf{E}_{\mathcal{D}_r}[|c(x) - p(x)|] \leq \epsilon$.*

This result can be easily converted to a basis for approximating disjunctions over arbitrary symmetric distributions. All we need is to partition the domain $\{0,1\}^n$ into layers as $\cup_{0\leq r\leq n} S_r$ and use a (different) polynomial for each layer. Formally, the basis now contains functions of the form $\mathrm{IND}(r)\cdot\chi$, where IND is the indicator function of being in layer of Hamming weight r and χ is a monomial of degree $O(\log(1/\epsilon))$. We note that a related strategy, of constructing a collection of functions, one for each layer of the cube was used by Wimmer (2010) to give an $n^{O(1/\epsilon^4)}$ time agnostic learning algorithm for the class of halfspaces on symmetric distributions. However, his proof technique is based on an involved use of representation theory of the symmetric group and is not related to ours.

Our proof technique also gives a simpler proof for the result of Blais et al. (2008) that implies approximation of disjunction by low-degree polynomials on all product distributions.

Theorem 3 *For any disjunction c and product distribution $\mathcal{D}$ on $\{0,1\}^n$, there is a polynomial p of degree $O(\log(1/\epsilon))$ such that $\mathbf{E}_{x\sim\mathcal{D}}[|c(x) - p(x)|] \leq \epsilon$.*

1.2 Applications

Theorem 2 together with a standard application of ℓ_1 regression (Kalai et al., 2008) yields an agnostic learning algorithm for the class of disjunctions running in time $n^{O(\log(1/\epsilon))}$.

Corollary 4 *There is an algorithm that agnostically learns the class of disjunctions on arbitrary symmetric distributions on $\{0,1\}^n$ in time $n^{O(\log(1/\epsilon))}$.*

This learning algorithm was extended to the class of all coverage functions, and then applied to the well-studied problem of privately releasing answers to all short conjunction queries with low average error (Feldman and Kothari, 2014).

It was shown by Kalai et al. (2009) and Feldman (2010) that agnostic learning of conjunctions over a distribution D in time $T(n, 1/\epsilon)$ implies learning of DNF formulas with s terms over D in time $\mathsf{poly}(n, 1/\epsilon) \cdot T(n, (4s/\epsilon))$. Further, under the same conditions distribution-specific agnostic boosting (Kalai and Kanade, 2009; Feldman, 2010) implies that there exists an agnostic learning algorithm for decision trees with s leaves running in time $\mathsf{poly}(n, 1/\epsilon) \cdot T(n, s/\epsilon)$. Therefore we obtain quasi-polynomial learning algorithms for DNF formulas and decision trees over symmetric distributions.

Corollary 5
1. *DNF formulas with s terms are PAC learnable with error ϵ in time $n^{O(\log(s/\epsilon))}$ over all symmetric distributions;*
2. *Decision trees with s leaves are agnostically learnable with excess error ϵ in time $n^{O(\log(s/\epsilon))}$ over all symmetric distributions.*

We also observe that any algorithm that agnostically learns the class of disjunction on the uniform distribution in time $n^{o(\log(\frac{1}{\epsilon}))}$ would yield a faster algorithm for the notoriously hard problem of Learning Sparse Parities with Noise. This is implicit in prior work (Kalai et al., 2008; Feldman, 2012) and we provide additional details in Section 5.

Dachman-Soled et al. (2015) recently showed that ℓ_1 approximation by polynomials is necessary and sufficient condition for agnostic learning over a product distribution (at least in the statistical query framework of Kearns (1998)). Our agnostic learning algorithm (Theorem 4) and lower bound for polynomial approximation (Theorem 1) demonstrate that this equivalence does not hold for non-product distributions.

2. Preliminaries

We use $\{0,1\}^n$ to denote the n-dimensional Boolean hypercube. Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For $S \subseteq [n]$, we denote by $\mathsf{OR}_S : \{0,1\}^n \to \{0,1\}$, the monotone Boolean disjunction on variables with indices in S, that is, for any $x \in \{0,1\}^n$, $\mathsf{OR}_S(x) = 0 \Leftrightarrow \forall i \in S \;\; x_i = 0$.

One can define norms and errors with respect to any distribution $\mathcal{D}$ on $\{0,1\}^n$. Thus, for $f : \{0,1\}^n \to \mathbb{R}$, we write the ℓ_1 and ℓ_2 norms of f as $\|f\|_1 = \mathbf{E}_{x\sim\mathcal{D}}[|f(x)|]$ and $\|f\|_2 = \sqrt{\mathbf{E}[f(x)^2]}$ respectively. The ℓ_1 and ℓ_2 error of f with respect to g are given by $\|f - g\|_1$ and $\|f - g\|_2$ respectively.

2.1 Agnostic Learning

The agnostic learning model is formally defined as follows (Haussler, 1992; Kearns et al., 1994).

Definition 6 *Let $\mathcal{F}$ be a class of Boolean functions and let $\mathcal{D}$ be any fixed distribution on $\{0,1\}^n$. For any distribution $\mathcal{P}$ over $\{0,1\}^n \times \{0,1\}$, let opt$(\mathcal{P}, \mathcal{F})$ be defined as: opt$(\mathcal{P}, \mathcal{F}) = \inf_{f\in\mathcal{F}} \mathbf{E}_{(x,y)\sim\mathcal{P}}[|y - f(x)|]$. An algorithm $\mathcal{A}$, is said to agnostically learn $\mathcal{F}$ on $\mathcal{D}$ if for every* excess error $\epsilon > 0$ *and any distribution $\mathcal{P}$ on $\{0,1\}^n \times \{0,1\}$ such that the marginal of $\mathcal{P}$ on*

$\{0,1\}^n$ *is* $\mathcal{D}$*, given access to random independent examples drawn from* $\mathcal{P}$*, with probability at least* $\frac{2}{3}$*,* $\mathcal{A}$ *outputs a hypothesis* $h : \{0,1\}^n \to [0,1]$*, such that* $\mathbf{E}_{(x,y)\sim\mathcal{P}}[|h(x) - y|] \leq opt(\mathcal{P}, \mathcal{F}) + \epsilon$*.*

It is easy to see that given a set of t examples $\{(x^i, y^i)\}_{i\leq t}$ and a set of m functions $\phi_1, \phi_2, \ldots, \phi_m$ finding coefficients $\alpha_1, \ldots, \alpha_m$ which minimize

$$\sum_{i\leq t} \left| \sum_{j\leq m} \alpha_j \phi_j(x^i) - y^i \right|$$

can be formulated as a linear program. This LP is referred to as Least-Absolute-Error (LAE) LP or Least-Absolute-Deviation LP, or ℓ_1 linear regression. As observed by Kalai et al. (2008), ℓ_1 linear regression gives a general technique for agnostic learning of Boolean functions.

Theorem 7 *Let* $\mathcal{C}$ *be a class of Boolean functions,* $\mathcal{D}$ *be distribution on* $\{0,1\}^n$ *and* $\phi_1, \phi_2, \ldots, \phi_m : \{0,1\}^n \to \mathbb{R}$ *be a set of functions that can be evaluated in time polynomial in* n*. Assume that there exists* Δ *such that for each* $f \in \mathcal{C}$*, there exist reals* $\alpha_1, \alpha_2, \ldots, \alpha_m$ *such that*

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}} \left[\left| \sum_{i\leq m} \alpha_i \phi_i(x) - f(x) \right| \right] \leq \Delta.$$

Then there is an algorithm that for every $\epsilon > 0$ *and any distribution* $\mathcal{P}$ *on* $\{0,1\}^n \times \{0,1\}$ *such that the marginal of* $\mathcal{P}$ *on* $\{0,1\}^n$ *is* $\mathcal{D}$*, given access to random independent examples drawn from* $\mathcal{P}$*, with probability at least* $2/3$*, outputs a function* h *such that*

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{P}}[|h(x) - y|] \leq \Delta + \epsilon.$$

The algorithm uses $O(m/\epsilon^2)$ *examples, runs in time polynomial in* n*,* m*,* $1/\epsilon$ *and returns a linear combination of* ϕ_i*'s.*

The output of this LP is not necessarily a Boolean function but can be converted to a Boolean function with disagreement error of $\Delta+2\epsilon$ using "$h(x) \geq \theta$" function as a hypothesis for an appropriately chosen θ (Kalai et al., 2008).

3. ℓ_1 Approximation on Symmetric Distributions

In this section, we show how to approximate the class of all disjunctions on any symmetric distribution by a linear combination of a small set of basis functions.

As discussed above, polynomials of degree $O(\log(1/\epsilon))$ can ϵ-approximate any disjunction in ℓ_1 distance on any product distribution. This is equivalent to using low-degree monomials as basis functions. We first show that this basis would not suffice for approximating disjunctions on symmetric distributions. Indeed, we construct a symmetric distribution on $\{0,1\}^n$, on which, any polynomial that approximates the monotone disjunction $c = x_1 \vee x_2 \vee \cdots \vee x_n$ within ℓ_1 error of $1/3$ must be of degree $\Omega(\sqrt{n})$.

3.1 Lower Bound on ℓ_1 Approximation by Low-Degree Polynomials

In this section we give the proof of Theorem 1.

Proof [of Thm. 1] Let $d : [n] \to \{0,1\}$ be the predicate corresponding to the disjunction $x_1 \vee x_2 \vee \cdots \vee x_n$, that is, $d(0) = 0$ and $d(i) = 1$ for each $i > 0$.

Consider a natural linear program to find a univariate polynomial f of degree at most d such that $\|d - f\|_\infty = \max_{0 \le i \le n} |d(i) - f(i)|$ is minimized:

$$\begin{aligned} &\min\ \epsilon \\ &s.t.\ \epsilon \ge |d(m) - \sum_{i=0}^{r} \alpha_i \cdot m^i| && \forall\ m \in \{0,\ldots,n\} \\ &\alpha_i \in \mathbb{R} && \forall\ i \in \{0,\ldots,r\}. \end{aligned}$$

This program (and its dual) often comes up in proving polynomial degree lower bounds for various function classes (for example, Sherstov, 2009). If $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$ is a solution for the program above that has value ϵ then $f(m) = \sum_{i=0}^{r} \alpha_i m^i$ is a degree r polynomial that approximates d within an error of at most ϵ at every point in $\{0,\ldots,n\}$. Klivans and Sherstov (2010) show that there exists an $r^* = \Theta(\sqrt{n})$, such that the optimal value of the program above for $r = r^*$ is $\epsilon^* \ge 1/3$. Standard manipulations can be used to produce the dual of the program:

$$\begin{aligned} &\max\ \sum_{m=0}^{n} \beta_m \cdot d(m) \\ &s.t.\ \sum_{m=0}^{n} \beta_m \cdot m^i = 0 && \forall\ i \in \{0,\ldots,r\} \\ &\sum_{m=0}^{n} |\beta_m| \le 1 \\ &\beta_m \in \mathbb{R} && \forall\ m \in \{0,\ldots,n\}. \end{aligned}$$

Let $\beta^* = \{\beta^*_m\}_{m \in \{0,\ldots,n\}}$ denote an optimal solution for the dual program with $r = r^*$. Then, by strong duality, the value of the dual is also ϵ^*. Observe that $\sum_{m=0}^{n} |\beta^*_m| = 1$, since otherwise we can scale up all the β^*_m by the same factor and increase the value of the program while still satisfying the constraints.

Let $\rho : \{0,\ldots,n\} \to [0,1]$ be defined by $\rho(m) = |\beta^*_m|$. Then ρ can be viewed as a density function of a distribution on $\{0,\ldots,n\}$ and we use it to define a symmetric distribution $\mathcal{D}$ on $\{-1,1\}^n$ as follows: $\mathcal{D}(x) = \rho(w(x))/\binom{n}{w(x)}$, where $w(x) = \sum_{i=1}^{n} x_i$ is the Hamming weight of point x. We now show that any polynomial p of degree r^* satisfies $\mathbf{E}_{x \sim \mathcal{D}}[|c(x) - p(x)|] \ge 1/3$.

We now extract a univariate polynomial f_p that approximates d on the distribution with the density function ρ using p. Let $p_{avg} : \{-1,1\}^n \to \mathbb{R}$ be obtained by averaging p over every layer. That is, $p_{avg}(x) = \mathbf{E}_{z \sim \mathcal{D}_{w(x)}}[p(z)]$, where $w(x)$ denotes the Hamming weight of x. It is easy to check that since c is symmetric, p_{avg} is at least as close to c as p in ℓ_1 distance.

Further, p_{avg} is a symmetric function computed by a multivariate polynomial of degree at most r^* on $\{0,1\}^n$. Thus, the function $f_p(m)$ that gives the value of p_{avg} on points of Hamming weight m can be computed by a univariate polynomial of degree r^*. Further,

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}}[|c(x)-p(x)|] \geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[|c(x)-p_{avg}(x)|] = \mathop{\mathbf{E}}_{m\sim\rho}[|d(m)-f_p(m)|].$$

Let us now estimate the error of f_p w.r.t d on the distribution ρ. Using the fact that f_p is of degree at most r^* and thus $\sum_{m=0}^{n} f_p(m)\cdot\beta_m = 0$ (enforced by the dual constraints), we have:

$$\begin{aligned}\mathop{\mathbf{E}}_{m\sim\rho}[|d(m)-f_p(m)|] &\geq \mathop{\mathbf{E}}_{m\sim\rho}[(d(m)-f_p(m))\cdot \mathsf{sign}(\beta^*_m)]\\ &= \sum_{m=0}^{n} d(m)\cdot\beta^*_m - \sum_{m=0}^{n} f_p(m)\cdot\beta^*_m\\ &= \epsilon^* - 0 = \epsilon^* \geq 1/3.\end{aligned}$$

Thus, the degree of any polynomial that approximates c on the distribution $\mathcal{D}$ with error of at most $1/3$ is $\Omega(\sqrt{n})$. ■

3.2 Upper Bound

In this section, we describe how to approximate disjunctions on any symmetric distribution by using a linear combination of functions from a set of small size. Recall that S_r denotes the set of all points from $\{0,1\}^n$ with weight r.

As we have seen above, symmetric distributions can behave very differently when compared to (constant bounded) product distributions. However, for the special case of the uniform distribution on S_r, denoted by $\mathcal{D}_r$, we show that for every disjunction c, there is a polynomial of degree $O(\log(1/\epsilon))$ that ϵ-approximates it in ℓ_1 distance on $\mathcal{D}_r$. As described in Section 1.1, one can stitch together polynomial approximations on each S_r to build a set of basis functions $\mathbb{S}$ such that every disjunction is well approximated by some linear combination of functions in $\mathbb{S}$. Thus, our goal is now reduced to constructing approximating polynomials on $\mathcal{D}_r$.

Proof [of Thm. 2] We first assume that c is monotone and without loss of generality $c = x_1\vee\cdots\vee x_k$. We will also prove a slightly stronger claim that $\mathbf{E}_{\mathcal{D}_r}[|c(x)-p(x)|] \leq \mathbf{E}_{\mathcal{D}_r}[(c(x)-p(x))^2] \leq \epsilon$ in this case. Let $d : \{0,\ldots,k\}\to\{0,1\}$ be the predicate associated with the disjunction, that is $d(i) = 1$ whenever $i \geq 1$. Note that $c(x) = d\left(\sum_{i\in[k]} x_i\right)$. Therefore our goal is to find a univariate polynomial f that approximates d and then substitute $p_f(x) = f\left(\sum_{i\in[k]} x_i\right)$. This substitution preserves the total degree of the polynomial. We break our construction into several cases based on the relative magnitudes of r, k and ϵ.

If $k \leq 2\ln(1/\epsilon)$, then the univariate polynomial that exactly computes the predicate d satisfies the requirements. Thus assume that $k > 2\ln(1/\epsilon)$. If $r > n-k$, then, c always takes the value 1 on S_r and thus the constant polynomial 1 achieves zero error. If on the

other hand, if $r \geq (n/k)\ln(1/\epsilon)$, then,

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_r}[c(x)=0] = \frac{\binom{n-k}{r}}{\binom{n}{r}} = \prod_{i=0}^{r-1}\left(1-\frac{k}{n-i}\right) \leq (1-k/n)^r \leq e^{-kr/n} \leq \epsilon.$$

In this case, the constant polynomial 1 achieves an ℓ_2^2 error of at most $\mathbf{Pr}_{x\sim\mathcal{D}_r}[c(x)=0]\cdot 1 \leq \epsilon$. Finally, observe that $r \leq (n/k)\ln(1/\epsilon)$ and $k > 2\ln(1/\epsilon)$ implies $r \leq n/2$. Thus, for the remaining part of the proof, assume that $r < \min\{n-k, (n/k)\ln(1/\epsilon), n/2\}$.

Consider the univariate polynomial $f : \{0,\ldots,k\} \to \mathbb{R}$ of degree t (for some t to be chosen later) that computes the predicate d exactly on $\{0,\ldots,t\}$. This polynomial is given by

$$f(w) = 1 - \frac{1}{t!}\prod_{i=1}^{t}(w-i) = \begin{cases} 1-\binom{w}{t} \text{ for } w>t \\ 1 \text{ for } 0<w\leq t \\ 0 \text{ for } w=0 \end{cases}$$

Let

$$\delta_j = \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_r}[|\{i \mid x_i=1\}| = j] = \frac{\binom{n-k}{r-j}\cdot\binom{k}{j}}{\binom{n}{r}}.$$

The ℓ_2^2 error of $p_f(x)$ on c satisfies,

$$||p_f - c||_2^2 = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[(c(x)-p_f(x))^2] = \sum_{j=t+1}^{k}\delta_j\cdot\binom{j}{t}^2.$$

We denote the RHS of this equality by $\|d-f\|_2^2$.

We first upper bound δ_j as follows:

$$\begin{aligned}
\delta_j = \frac{\binom{n-k}{r-j}\cdot\binom{k}{j}}{\binom{n}{r}} &= \frac{(n-k)!}{(n-k-r+j)!(r-j)!}\cdot\frac{k!}{(k-j)!j!}\cdot\frac{(n-r)!r!}{n!} \\
&= \frac{1}{j!}\cdot\frac{r!}{(r-j)!}\cdot\frac{k!}{(k-j)!}\cdot\frac{(n-r)!}{n!}\cdot\frac{(n-k)!}{(n-k-r+j)!} \\
&\leq \frac{1}{j!}\cdot(rk)^j\cdot\frac{(n-k)\cdot(n-k-1)\cdots(n-k-r+j+1)}{n\cdot(n-1)\cdots(n-r+1)} \\
&\leq \frac{1}{j!}\cdot(n\ln(1/\epsilon))^j\cdot\frac{1}{(n-r+j)\cdot(n-r+j-1)\cdots(n-r+1)},
\end{aligned}$$

where, in the second to last inequality, we used that $r < n/k\ln(1/\epsilon)$ to conclude that $rk \leq (n\ln(1/\epsilon))$. Now, $r < n/2$ and thus $(n-r+1) > n/2$. Therefore,

$$\delta_j \leq \frac{2^j\cdot(n\ln(1/\epsilon))^j}{n^j\cdot j!} = \frac{(2\ln(1/\epsilon))^j}{j!},$$

and thus:

$$\|d-f\|_2^2 \leq \sum_{j=t+1}^{k}\binom{j}{t}^2\frac{(2\ln(1/\epsilon))^j}{j!}.$$

Set $t = 8e^2\ln(1/\epsilon)$. Using $j! > (j/e)^j > (t/e)^j$ for every $j \geq t+1$, we obtain:

$$\|d-f\|_2^2 \leq \sum_{j=t+1}^{k} 2^{2j} \cdot \left(\frac{2\ln(1/\epsilon)}{8e\ln(1/\epsilon)}\right)^j \leq \epsilon \cdot \sum_{j=t+1}^{\infty} 1/e^j \leq \epsilon. \tag{1}$$

To see that $\mathbf{E}_{\mathcal{D}_r}[|c(x)-p(x)|] \leq \mathbf{E}_{\mathcal{D}_r}[(c(x)-p(x))^2]$ we note that in all cases and for all x, $|p(x)-c(x)|$ is either 0 or ≥ 1. This completes the proof of the monotone case.

We next consider the more general case when $c = x_1 \vee x_2 \vee \cdots \vee x_{k_1} \vee \bar{x}_{k_1+1} \vee \bar{x}_{k_1+2} \vee \cdots \vee \bar{x}_{k_1+k_2}$. Let $c_1 = x_1 \vee x_2 \vee \cdots \vee x_{k_1}$ and $c_2 = \bar{x}_{k_1+1} \vee \bar{x}_{k_1+2} \vee \cdots \vee \bar{x}_{k_1+k_2}$ and $k = k_1 + k_2$. Observe that $c = 1-(1-c_1)\cdot(1-c_2) = c_1 + c_2 - c_1c_2$.

Let p_1 be a polynomial of degree $O(\log(1/\epsilon))$ such that $\|c_1 - p_1\|_1 \leq \|c_1 - p_1\|_2^2 \leq \epsilon/3$. Note that if we swap 0 and 1 in $\{0,1\}^n$ then c_2 will be equal to a monotone disjunction $\bar{c}_2 = x_{k_1+1} \vee x_{k_1+2} \vee \cdots \vee x_{k_1+k_2}$ and $\mathcal{D}_r$ will become $\mathcal{D}_{n-r}$. Therefore by the argument for the monotone case, there exists a polynomial $\bar{p}_2$ of degree $O(\log(1/\epsilon))$ such that $\|\bar{c}_2 - \bar{p}_2\|_1 \leq \epsilon/3$. By renaming the variables back we will obtain a polynomial p_2 of degree $O(\log(1/\epsilon))$ such that $\|c_2 - p_2\|_1 \leq \|c_2 - p_2\|_2^2 \leq \epsilon/3$. Now let $p = p_1 + p_2 - p_1p_2$. Clearly the degree of p is $O(\log(1/\epsilon))$. We now show that $\|c-p\|_1 \leq \epsilon$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|c(x)-p(x)|] &= \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|(1-c(x))-(1-p(x))|] \\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|(1-c_1)(1-c_2)-(1-p_1)(1-p_2)|] \\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|(1-c_1)(p_2-c_2)+(1-c_2)(p_1-c_1)-(c_1-p_1)(c_2-p_2)|] \\
&\leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|(1-c_1)(p_2-c_2)|] + \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|(1-c_2)(p_1-c_1)|] + \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|(c_1-p_1)(c_2-p_2)|] \\
&\leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|p_2-c_2|] + \mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[|p_1-c_1|] + \sqrt{\mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[(c_1-p_1)^2]\mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[(c_2-p_2)^2]} \\
&\leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon.
\end{aligned}$$

■

4. Polynomial Approximation on Product Distributions

In this section, we show that for every product distribution $\mathcal{D} = \prod_{i\in[n]} \mathcal{D}_i$, every $\epsilon > 0$ and every disjunction (or conjunction) c of length k, there exists a polynomial $p : \{0,1\}^n \to \mathbb{R}$ of degree $O(\log(1/\epsilon))$ such that p ϵ-approximates c in ℓ_1 distance on $\mathcal{D}$.

Proof [of Thm. 3] First, we note that without loss of generality we can assume that the disjunction c is equal to $x_1 \vee x_2 \vee \cdots \vee x_k$ for some $k \in [n]$. We can assume monotonicity since we can convert negated variables to un-negated variables by swapping the roles of 0 and 1 for that variable. The obtained distribution will remain product after this operation. Further we can assume that $k = n$ since variables with indices $i > k$ do not affect probabilities of variables with indices $\leq k$ or the value of $c(x)$.

We first note that we can assume that $\mathbf{Pr}_{x\sim\mathcal{D}}[x = 0^k] > \epsilon$ since, otherwise, the constant polynomial 1 gives the desired approximation. Let $\mu_i = \mathbf{Pr}_{x_i\sim\mathcal{D}^i}[x_i = 1]$. Since c is a

symmetric function, its value at any $x \in \{0,1\}^k$ depends only on the Hamming weight of x that we denote by $w(x)$. Thus, we can equivalently work with the univariate predicate $d : \{0,1,\ldots,k\} \to \{0,1\}$, where $d(i) = 1$ for $i > 0$ and $d(0) = 0$.

As in the proof of Theorem 2, we will approximate d by a univariate polynomial f and then use the polynomial $p_f(x) = f(w(x))$ to approximate c.

Let $f : \{0,1,\ldots,k\} \to \mathbb{R}$ be the univariate polynomial of degree t that matches d on all points in $\{0,1,\ldots,t\}$. Thus,

$$f(w) = 1 - \frac{1}{t!} \cdot \prod_{i=1}^{t} (w-i) = \begin{cases} 1-\binom{w}{t} \text{ for } w > t \\ 1 \text{ for } 0<w\leq t \\ 0 \text{ for } w=0 \end{cases}$$

We have,

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_r}[(c(x) - p_f(x))^2] = \sum_{j=0}^{k} \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[w(x) = j] \cdot |d(j) - f(j)|$$

and we denote the RHS of this equation by $\|d - f\|_1$.

Then:

$$\begin{aligned} \|d - f\|_1 &= \sum_{j=t+1}^{k} \mathop{\mathbf{Pr}}_{\mathcal{D}}[w(x) = j] \cdot |1 - f(j)| \\ &= \sum_{j=t+1}^{k} \mathop{\mathbf{Pr}}_{\mathcal{D}}[w(x) = j] \cdot \binom{j}{t}. \end{aligned} \tag{2}$$

Let us now estimate $\mathbf{Pr}_{\mathcal{D}}[w(x) = j]$.

$$\begin{aligned} \mathop{\mathbf{Pr}}_{\mathcal{D}}[w(x) = j] &= \sum_{S\subseteq[n],\ |S|=j} \prod_{i\in S} \mu_i \cdot \prod_{i\notin S} (1-\mu_i) \\ &\leq \sum_{S\subseteq[n],\ |S|=j} \prod_{i\in S} \mu_i \end{aligned}$$

Observe that in the expansion of $(\sum_{i=1}^{k} \mu_i)^j$, the term $\prod_{i\in S} \mu_i$ occurs exactly $j!$ times. Thus,

$$\sum_{S\subseteq[n],\ |S|=j} \prod_{i\in S} \mu_i \leq \frac{(\sum_{i=1}^{k} \mu_i)^j}{j!}.$$

Set $\mu_{avg} = \frac{1}{k}\sum_{i=1}^{k} \mu_i$. We have:

$$\epsilon \leq \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[x = 0^k] = \prod_{i=1}^{k} (1-\mu_i) \leq \left(1 - \frac{1}{k} \cdot \sum_{i=1}^{k} \mu_i\right)^k = (1-\mu_{avg})^k.$$

Thus, $\mu_{avg} = c/k$ for some $c \leq 2\ln(1/\epsilon)$ whenever $k \geq k_0$ where k_0 is some universal constant. In what follows, assume that $k \geq k_0$. (Otherwise, we can use the polynomial of degree equal to k that exactly computes the predicate d on all points).

We are now ready to upper bound the error $\|d - f\|_1$. From Equation (2), we have:

$$\|d - f\|_1 = \sum_{j=t+1}^{k} \mathop{\mathbf{Pr}}_{\mathcal{D}}[w(x) = j] \cdot \binom{j}{t} \leq \sum_{j=t+1}^{k} \frac{(\sum_{i=1}^{k} \mu_i)^j}{j!} \cdot \binom{j}{t}$$
$$\leq \sum_{j=t+1}^{k} \binom{j}{t} \cdot \frac{(2\ln(1/\epsilon))^j}{j!}$$

Setting $t = 4e^2 \ln(1/\epsilon)$ and using the calculation from Equation (1) in the proof of Thm. 2, we obtain that the error $\|d - f\|_1 \leq \epsilon$. ■

5. Agnostic Learning of Disjunctions

Combining Thm. 7 with the results of the previous section (and the discussion in Section 1.1), we obtain an agnostic learning algorithm for the class of all disjunctions on product and symmetric distributions running in time $n^{O(\log(1/\epsilon))}$.

Corollary 8 (Cor. 4, restated) *There is an algorithm that agnostically learns the class of disjunctions on any product or symmetric distribution on $\{0,1\}^n$ with excess error of at most ϵ in time $n^{O(\log(1/\epsilon))}$.*

We now remark that any algorithm that agnostically learns the class of disjunctions (or conjunctions) on n inputs on the uniform distribution on $\{0,1\}^n$ in time $n^{o(\log(\frac{1}{\epsilon}))}$ would yield a faster algorithm for the notoriously hard problem of Learning Sparse Parities with Noise(SLPN). The reduction is based on the technique implicit in the work of Kalai et al. (2008) and Feldman (2012).

For $S \subseteq [n]$, we use χ_S to denote the parity of inputs with indices in S. Let $\mathcal{U}$ denote the uniform distribution on $\{0,1\}^n$. We say that random examples of a Boolean function f have noise of rate η if the label of a random example equals $f(x)$ with probability $1 - \eta$ and $1 - f(x)$ with probability η.

Problem 1 (Learning Sparse Parities with Noise) *For $\eta \in (0, 1/2)$ and $k \leq n$ the problem of learning k-sparse parities with noise η is the problem of finding (with probability at least $2/3$) the set $S \subseteq [n], |S| \leq k$, given access to random examples with noise of rate η of parity function χ_S.*

The fastest known algorithm for learning k-sparse parities with noise η is a recent breakthrough result of Valiant (2012) which runs in time $O(n^{0.8k}\mathsf{poly}(\frac{1}{1-2\eta}))$.

Kalai et al. (2008) and Feldman (2012) prove hardness of agnostic learning of majorities and conjunctions, respectively, based on correlation of concepts in these classes with parities. We state below this general relationship between correlation with parities and reduction to SLPN given by Feldman et al. (2013).

Lemma 9 *Let $\mathcal{C}$ be a class of Boolean functions on $\{0,1\}^n$. Suppose, there exist $\gamma > 0$ and $k \in \mathbb{N}$ such that for every $S \subseteq [n]$, $|S| \leq k$, there exists a function, $f_S \in \mathcal{C}$, such*

that $|\mathbf{E}_{x\sim\mathcal{U}}[f_S(x)\chi_S(x)]| \geq \gamma(k)$. *If there exists an algorithm* $\mathcal{A}$ *that learns the class* $\mathcal{C}$ *agnostically with excess error* ϵ *in time* $T(n, \frac{1}{\epsilon})$ *then, there exists an algorithm* $\mathcal{A}'$ *that learns* k*-sparse parities with noise* $\eta < 1/2$ *in time* $\mathsf{poly}(n, \frac{1}{(1-2\eta)\gamma(k)}) + 2T(n, \frac{2}{(1-2\eta)\gamma(k)})$.

The correlation between a disjunction and a parity is easy to estimate.

Lemma 10 *For any* $S \subseteq [n]$, $|\mathbf{E}_{x\sim\mathcal{U}}[\mathsf{OR}_S(x)\chi_S(x)]| = \frac{1}{2^{|S|-1}}$.

We thus immediately obtain the following corollary.

Theorem 11 *Suppose there exists an algorithm that learns the class of Boolean disjunctions over the uniform distribution agnostically with excess error of* $\epsilon > 0$ *in time* $T(n, \frac{1}{\epsilon})$. *Then there exists an algorithm that learns* k*-sparse parities with noise* $\eta < \frac{1}{2}$ *in time* $\mathsf{poly}(n, \frac{2^{k-1}}{1-2\eta}) + 2T(n, \frac{2^{k-1}}{1-2\eta})$. *In particular, if* $T(n, \frac{1}{\epsilon}) = n^{o(\log(1/\epsilon))}$, *then, there exists an algorithm to solve* k*-SLPN in time* $n^{o(k)}$.

Thus, any algorithm that is asymptotically faster than the one from Cor. 4 yields a faster algorithm for k-SLPN.

References

P. Awasthi, A. Blum, and O. Sheffet. Improved guarantees for agnostic learning of disjunctions. In *Proceedings of COLT*, pages 359–367, 2010.

S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

E. Blais, R. O'Donnell, and K. Wimmer. Polynomial regression under arbitrary product distributions. In *Proceedings of COLT*, pages 193–204, 2008.

N. Bshouty and L. Burroughs. Maximizing agreements and coagnostic learning. *Theoretical Computer Science*, 350(1):24–39, 2006.

D. Dachman-Soled, V. Feldman, L.-Y. Tan, A. Wan, and K. Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of SODA*, 2015.

A. Daniely and S. Shalev-Shwartz. Complexity theoretic limitations on learning DNF's. *CoRR*, abs/1404.3378, 2014.

V. Feldman. Distribution-specific agnostic boosting. In *Proceedings of Innovations in Computer Science (ICS)*, pages 241–250, 2010.

V. Feldman. A complete characterization of statistical query learning with applications to evolvability. *Journal of Computer System Sciences*, 78(5):1444–1459, 2012.

V. Feldman and P. Kothari. Learning coverage functions and private release of marginals. In *Proceedings of COLT*, 2014.

V. Feldman, P. Gopalan, S. Khot, and A. Ponuswami. On agnostic learning of parities, monomials and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

V. Feldman, P. Kothari, and J. Vondrák. Representation, approximation and learning of submodular functions using low-rank decision trees. In *Prooceedings of COLT*, pages 30:711–740, 2013.

A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of STOC*, 2011.

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

A. Kalai and V. Kanade. Potential-based agnostic boosting. In *Proceedings of NIPS*, pages 880–888, 2009.

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

A. Kalai, V. Kanade, and Y. Mansour. Reliable agnostic learning. In *Proceedings of COLT*, 2009.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

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

M. Marchand and J. Shawe-Taylor. The set covering machine. *Journal of Machine Learning Research*, 3:723–746, 2002.

A. Sherstov. Approximate inclusion-exclusion for arbitrary symmetric functions. *Computational Complexity*, 18(2):219–247, 2009.

G. Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *Proceedings of FOCS*, 2012.

K. Wimmer. Agnostically learning under permutation invariant distributions. In *Proceedings of FOCS*, pages 113–122, 2010.