JMLR: Workshop and Conference Proceedings vol 40:1–19, 2015

A PTAS for Agnostically Learning Halfspaces

Amit Daniely AMIT.DANIELY@MAIL.HUJI.AC.IL

Abstract

We present a PTAS for agnostically learning halfspaces w.r.t. the uniform distribution on the d dimensional sphere. Namely, we show that for every $\mu > 0$ there is an algorithm that runs in time $\mathrm{poly}\left(d, \frac{1}{\epsilon}\right)$, and is guaranteed to return a classifier with error at most $(1+\mu)\mathrm{opt} + \epsilon$, where opt is the error of the best halfspace classifier. This improves on Awasthi, Balcan and Long Awasthi et al. (2014) who showed an algorithm with an (unspecified) constant approximation ratio. Our algorithm combines the classical technique of polynomial regression (e.g. Linial et al. (1989); Kalai et al. (2005)), together with the new localization technique of Awasthi et al. (2014).

Keywords: Agnostic learning, Uniform distribution, Halfspaces, Approximation algorithms, Polynomial approximation, Localization, Polynomial regression

1. Introduction

In the problem of agnostically learning halfspaces, the learner is given an access to examples drawn from a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$ and an accuracy parameter $\epsilon > 0$. It is required to output[1] a classifier $h : \mathbb{R}^d \to \{\pm 1\}$ whose error, $\mathrm{Err}_{\mathcal{D}}(h) := \Pr_{(x,y)\sim\mathcal{D}}\left(h(x) \neq y\right)$, is at most[2] $\mathrm{opt}+\epsilon$. Here, opt is the error of the best classifier of the from $h_w(x) = \mathrm{sign}(\langle w, x\rangle)$. The learner is *efficient* if it runs in time $\mathrm{poly}\left(d, \frac{1}{\epsilon}\right)$. We note that we consider the general, *improper*, setting where the learner have the freedom to return a hypothesis that is not a halfspace classifier.

Halfspaces are extremely popular in practical applications, and have been extensively studied in Machine Learning, Statistics and Theoretical Computer Science (see section 1.2). Unfortunately, from a worst case perspective, the problem seems very hard: Best known efficient algorithms have a terrible approximation ratio of $\tilde{\Omega}(d)$. In the case of *proper learning*, where the output hypothesis must be a halfspace, agnostic learning is known to be $\mathcal{NP}$-hard. Even learning with a constant *approximation ratio*, where the returned classifier should have error $\le \alpha \cdot \mathrm{opt} + \epsilon$, is $\mathcal{NP}$-hard. In fact, even approximation ratio of $2^{\log^{0.99}(d)}$ is $\mathcal{NP}$-hard. In the general (improper) case, agnostic learning of halfspaces, and even agnostic learning with a constant approximation ratio, have been showed hard under various complexity assumptions (see section 1.2). In light of that, it is just natural to consider agnostic learning under various restrictions on the distribution $\mathcal{D}$. A very natural and widely studied such restriction Klivans et al. (2009, 2002); Awasthi et al. (2014); Kalai et al. (2005) is that the marginal distribution, $\mathcal{D}_{\mathbb{R}^d}$, is uniform on the sphere S^{d-1}.

Even under the uniform distribution, no efficient algorithms are known, and there is also an evidence that the problem is hard Klivans and Kothari (2014). This lead researchers to consider *approximation algorithms*. The first approximation guarantee is due to Kalai et al. (2005), who

1. Throughout, we require our algorithms to succeed with a constant probability (that can be standardly amplified by repetition).
2. Note that opt might be > 0, namely, we consider the "agnostic PAC learning" model Kearns et al. (1994).

showed an efficient regression based algorithm with approximation ratio of $\alpha = O\left(\sqrt{\log\left(\frac{1}{\mathrm{opt}}\right)}\right)$. In an exciting recent work, Awasthi et al. (2014) introduced a new algorithmic technique, called *localization*, and showed an efficient algorithm with an unspecified constant approximation ratio. In this paper, we advance this line of work further, and show a Polynomial Time Approximation Scheme (PTAS). Namely, we show:

Theorem 1 (main) *For every $\mu > 0$, there is an efficient algorithm for agnostically learning halfspaces under the uniform distribution with an approximation ratio of $(1+\mu)$.*

As noted above, Klivans and Kothari (2014) showed that under a certain complexity assumption (hardness of learning sparse parity), there are no exact efficient algorithms (i.e., with approximation ratio $\alpha = 1$). In that case, our result is optimal.

Label Complexity: Our algorithm naturally fits to the *active learning* (e.g. Settles (2010)) setting. Often, a label is much more expensive than an example (e.g., in biology, it might be the case that we have to make an experiment to get a label). It is therefore useful to make economical use of labels. Our algorithm naturally have such property, as its *label complexity* (i.e., the number of labels it needs to see) is poly-logarithmic in $\frac{1}{\mathrm{opt}}$ (see theorem 5 for a more detailed statement).

Interpolation between approximation and exact algorithms: A more precise statement of our result is that there exists an algorithm with runtime $\mathrm{poly}\left(d^{\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}}, \frac{1}{\epsilon}\right)$ that returns a classifier with error $\leq (1+\mu)\mathrm{opt} + \epsilon$ for every $0 < \mu, \epsilon \leq 1$. Taking μ up to $\frac{\epsilon}{2}$ and replacing ϵ with $\frac{\epsilon}{2}$, the error bound is $\left(1+\frac{\epsilon}{2}\right)\mathrm{opt} + \frac{\epsilon}{2} \leq \mathrm{opt} + \epsilon$. Hence, we get an exact algorithm. The running time is $\mathrm{poly}\left(d^{\frac{\log^3\left(\frac{1}{\epsilon}\right)}{\epsilon^2}}\right)$, which almost matches the state of the art – $\mathrm{poly}\left(d^{\frac{1}{\epsilon^2}}\right)$ Kalai et al. (2005); Diakonikolas et al. (2010b).

Open questions: Obvious open questions are to extend our results to more distributions (uniform on $\{\pm 1\}^d$, permutation-invariant, product, log-concave, $\ldots$) and more problems (learning intersection of halfspaces, functions of halfspaces, ...). In addition, as opposed to previous approximation algorithms Awasthi et al. (2014); Kalai et al. (2005), our algorithm does not always return a halfspace classifier. A natural open question is therefore to find a *proper* PTAS.

1.1. Algorithmic Components, The PTAS, and Proof Outline

Our algorithm and its analysis build on and combine various algorithmic and proof techniques that were previously used for learning halfspaces. This includes regression based algorithms (e.g. Shalev-Shwartz et al. (2011); Kalai et al. (2005)), polynomial approximations of the sign function (e.g. Shalev-Shwartz et al. (2011); Kalai et al. (2005); Diakonikolas et al. (2010a,b)) and localization techniques Awasthi et al. (2014). In this section we outline these techniques and the way we use them. Then, we present our PTAS, state its properties (theorem 5), and describe the course of the proof. The full proof is in sections 2 and A.

1.1.1. Some preliminaries

Noise tolerance is a measure to evaluate the performance of learning algorithms, that is essentially equivalent to the approximation ratio. Yet, we find it slightly more convenient for the technical

exposition. We say that a learning algorithm tolerates noise rate of $0 < f(\eta) < \eta$ (w.r.t. halfspaces) if, when running on input $0 < \eta < 1$, it guaranteed to return a hypothesis with error $\le \eta$, provided that $\mathrm{opt} \le f(\eta)$. We say that such an algorithm is *efficient* if it runs in time $\mathrm{poly}\left(d, \frac{1}{\eta}\right)$. We note that given a learning algorithm that tolerates noise rate of $\frac{\eta}{\alpha}$, for some $\alpha > 1$, it is not hard to construct an algorithm with approximation ratio of α, and the running time grows only by a factor of $\mathrm{poly}\left(\frac{1}{\epsilon}\right)$: Indeed, in order to return a hypothesis with error $\le \alpha \cdot \mathrm{opt} + \epsilon$, we can run the algorithm with $\alpha \cdot \mathrm{opt} \le \eta \le \alpha \cdot \mathrm{opt} + \epsilon$. We can find such an η by trying $\eta = k\epsilon$ for $k = 1, 2, \ldots, \lceil \frac{1}{\epsilon} \rceil$.
Notation. Let $\mathcal{D}$ be a distribution on a space X. For $Y \subset X$ we denote by $\mathcal{D}|_Y$ the restriction of $\mathcal{D}$ to $\mathcal{Y}$. If $\mathcal{D}$ is a distribution on $X \times \{\pm 1\}$ we denote by $\mathcal{D}_X$ the marginal distribution on X. If $\mathcal{D}$ is a distribution on S^{d-1} (resp. $S^{d-1} \times \{\pm 1\}$) and $w \in S^{d-1}$, we define the *projection of $\mathcal{D}$ on w* as follows: If $x \sim \mathcal{D}$ (resp. $(x, y) \sim \mathcal{D}$) then $\mathcal{D}_w$ is the distribution (on $[-1, 1]$) of the random variable $\langle w, x\rangle$. For a distribution $\mathcal{D}$ on a space X and a function $f : S^{d-1} \to \mathbb{R}$, we denote $\|f\|_{p,\mathcal{D}} = \left(\mathbb{E}_{x\sim\mathcal{D}}|f(x)|^p\right)^{\frac{1}{p}}$. We will sometimes abuse notation and use $\|f\|_{p,\mathcal{D}}$ instead of $\|f\|_{p,\mathcal{D}_{S^{d-1}}}$ even when $\mathcal{D}$ is a distribution on $S^{d-1} \times \{\pm 1\}$. We denote by $\theta(w, w^*) = \cos^{-1}(\langle w, w^*\rangle)$ the angle between two vectors $w, w^* \in S^{d-1}$. We will frequently use the fact that for uniform $x \in S^{d-1}$ we have $\Pr\left(h_{w^*}(x) \ne h_w(x)\right) = \frac{\theta(w,w^*)}{\pi}$. We denote by $\mathrm{POL}_{r,d}$ the space of d-variate polynomials of degree $\le r$. For $w \in S^{d-1}$ and $\gamma > 0$ we let $T_{d,\gamma}(w) := \{u \in S^{d-1} : |\langle w, u\rangle| \le \gamma\}$.

1.1.2. POLYNOMIAL ℓ_1-REGRESSION FOR CLASSIFICATION

The output of a classification algorithm is a (description of a) hypothesis $h : S^{d-1} \to \{\pm 1\}$. Often, the returned hypothesis is of the form $h(x) = \mathrm{sign}(f(x))$, for some real valued function $f : S^{d-1} \to \mathbb{R}$. To conveniently dealing with such hypotheses, we introduce some terminology. We denote the standard (zero-one) loss of f by $\mathrm{Err}_{\mathcal{D}}(f) = \mathrm{Err}_{\mathcal{D}}(\mathrm{sign}(f))$. We also consider the ℓ_1-*loss*, $\mathrm{Err}_{\mathcal{D},1}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}|f(x) - y|$. We note that for $f : S^{d-1} \to \mathbb{R}$, since $\frac{|\mathrm{sign}(z)-1|}{2} \le |z - 1|$ for all z, we have

$$\mathrm{Err}_{\mathcal{D}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \frac{|\mathrm{sign}(yf(x)) - 1|}{2} \le \mathbb{E}_{(x,y)\sim\mathcal{D}}|yf(x) - 1| = \mathbb{E}_{(x,y)\sim\mathcal{D}}|f(x) - y| = \mathrm{Err}_{\mathcal{D},1}(f)$$

Thus, by finding f with small ℓ_1-error we can find a good classifier. The motivation for moving from the 0-1 loss to the ℓ_1 loss is the convexity of the ℓ_1 loss, which enables the use of convex optimization. Concretely, for "nice enough" convex set, $\mathcal{F}$, of functions from S^{d-1} to $\mathbb{R}$, it is possible to efficiently find (both in terms of number of examples and time) $f \in \mathcal{F}$ with ℓ_1 error almost as small as $\min_{f\in\mathcal{F}} \mathrm{Err}_{\mathcal{D},1}(f)$. Now, for a classifier $h : S^{d-1} \to \{\pm 1\}$ we have

$$\begin{aligned} \mathrm{Err}_{\mathcal{D}}(f) \le \mathrm{Err}_{\mathcal{D},1}(f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}|f(x) - y| \\ &\le \mathbb{E}_{(x,y)\sim\mathcal{D}}|f(x) - h(x)| + \mathbb{E}_{(x,y)\sim\mathcal{D}}|h(x) - y| \\ &= \|f - h\|_{1,\mathcal{D}} + 2\,\mathrm{Err}(h) \end{aligned} \tag{1}$$

Thus, if we minimize the ℓ_1-loss over a collection of functions that is large enough to contain a good ℓ_1-approximation of the best halfspace classifier, we can find a function whose ℓ_1-error, and therefore also the 0-1 error, is almost as good as the 0-1 error of the best halfspace classifier. Methods that follow the above spirit have been extensively studied in computational learning theory. Concretely, Kalai et al. (2005) suggested the following algorithm: First, find $P \in \mathrm{POL}_{r,d}$ that minimizes the empirical ℓ_1-error on the given sample[3]. Then, find a classifier that makes the least number of errors on the given sample, among all classifiers of the form $x \mapsto \text{sign}\left(P(x) - a\right)$ for $a \in \mathbb{R}$. We note that the second step is required in order to overcome the factor of 2 in equation (1). They used that algorithm to show:

Theorem 2 *Kalai et al. (2005) There is an algorithm with runtime* $\text{poly}\left(d^r, \frac{1}{\epsilon}\right)$ *such that, for every distribution* $\mathcal{D}$ *on* $S^{d-1} \times \{\pm 1\}$ *and every* $h : S^{d-1} \to \{\pm 1\}$, *it returns* $P \in \mathbf{POL}_{r,d}$ *with* $\text{Err}_{\mathcal{D}}(P) \le \text{Err}_{\mathcal{D}}(h) + \min_{P' \in \text{POL}_{r,d}} \|h - P'\|_{1,\mathcal{D}} + \epsilon$.

1.1.3. LEARNING HALFSPACES USING SIGN APPROXIMATIONS

To use theorem 2 for learning halfspaces, we need to prove the existence of low degree polynomials P such that $\|h - P\|_{1,\mathcal{D}}$ is small, where h is a halfspace classifier. As explained below, this is naturally done by approximating the *sign function*, $\text{sign}(x) = \begin{cases} 1 & x > 0 \\ -1 & x \le 0 \end{cases}$, with respect to an appropriate proximity measure.

Suppose that $w^* \in S^{d-1}$ defines the optimal halfspace and let $\mathcal{D}_{w^*}$ be the projection of $\mathcal{D}$ on w^*. For a univariate polynomial $p \in \text{POL}_{r,1}$, consider the d-variate polynomial $P \in \text{POL}_{r,d}$ given by $P(x) = p(\langle w^*, x\rangle)$. We have

$$\begin{aligned} \|P - h_{w^*}\|_{1,\mathcal{D}} &= \text{E}_{x \sim \mathcal{D}_{S^{d-1}}}\left[|p(\langle w^*, x\rangle) - \text{sign}(\langle w^*, x\rangle)|\right] \\ &= \text{E}_{x \sim \mathcal{D}_{w^*}}\left[|p(x) - \text{sign}(x)|\right] = \|p - \text{sign}\,\|_{1,\mathcal{D}_{w^*}} \end{aligned} \tag{2}$$

Therefore, in order to find a good ℓ_1 approximation for h_{w^*} w.r.t. $\mathcal{D}$, we can find a good ℓ_1 approximation for sign w.r.t. $\mathcal{D}_{w^*}$.

Approximating the sign function is a central component in many papers about halfspaces Birnbaum and Shalev-Shwartz (2012); Diakonikolas et al. (2010a,a); Kalai et al. (2005); Shalev-Shwartz et al. (2011). These papers needed to find approximation of the sign function w.r.t. relatively well studied proximity measures, such as the ℓ_∞ norm, or the ℓ_1 and ℓ_2 norms w.r.t. the Gaussian distribution. Therefore, some of these papers used basis expansion methods (Fourier, Hermite, Chebyshev, ...). In this paper we need to find ℓ_1 approximation w.r.t. somewhat messier distributions. Therefore, we use a somewhat more flexible approach, similar to the one used in Diakonikolas et al. (2010a). We rely on techniques from approximation theory Davis (1975). In particular, our main tool for constructing polynomials will be the celebrated Jackson's theorem.

Theorem 3 (Jackson, Davis (1975)) *For every L-lipschitz function $f : [-1, 1] \to \mathbb{R}$ and $r \in \mathbb{N}$ there is a degree r polynomial p such that $\|p - f\|_{\infty,[-1,1]} \le \frac{6L}{r}$.*

1.1.4. LOCALIZATION

An additional algorithmic component we will use, except polynomial regression, is *localization in the instance and the hypotheses space* (e.g. (Bartlett et al., 2005; Awasthi et al., 2014)). The basic idea is the following. Suppose that $w^* \in S^{d-1}$ defines the optimal halfspace. Suppose furthermore that we have found (say, using some simple algorithm) a vector $w \in S^{d-1}$ that defines a halfspace with a relatively small error. The facts that the marginal distribution is uniform and $\text{Err}(h_w)$ is small have two relevant consequences:

3. I.e., if the sample is $(x_1, y_1), \ldots, (x_m, y_m) \in S^{d-1} \times \{\pm 1\}$, find $P \in \text{POL}_{r,d}$ that minimizes $\frac{1}{m}\sum_{i=1}^m |P(x_i) - y_i|$.

- We know that the optimal vector, w^*, is close to w.
- Hence, if $|\langle w, x\rangle|$ is large, then $h_{w^*}(x) = h_w(x)$ and therefore we know $h_{w^*}(x)$.

These two properties enable us to "localize the learning" and concentrate only on hypotheses $h_{w'}$ with w' close to w, and on instances x with small $|\langle w, x\rangle|$. We will use this idea directly in our algorithm. In addition, we will use, as a black-box, the localization-based algorithm of Awasthi et al. (2014). Their algorithm starts with a crude approximation $w_1 \in S^{d-1}$ of the optimal halfspace w^*. Then, it finds w_2 that minimizes the *hinge loss* $\mathrm{E}_{\mathcal{D}|_{T\times\{\pm 1\}}}(1 - \langle w, yx\rangle)_+$ on the restriction of $\mathcal{D}$ to some small strip $T = \{x \in S^{d-1} \mid |\langle w, x\rangle| \le \gamma\}$. Then, it continue in this manner to find better and better w_i's. Awasthi, Balcan and Long used their algorithm to show:

Theorem 4 *Awasthi et al. (2014) There is an efficient learning algorithm with label complexity* $\mathrm{poly}\left(d, \log\left(\frac{1}{\eta}\right)\right)$ *that tolerates noise rate of* $\frac{\eta}{\alpha_0}$ *for some universal constant* $\alpha_0 > 1$*. Moreover, the algorithm is* proper*, that is, its output is a halfspace.*

1.1.5. THE PTAS AND ITS ANALYSIS

In a nutshell, our algorithm first find (step 1) a "rough estimation", w, of w^*. Then, it "localizes the learning" and apply more computation power (step 3), to a small strip T that is closed to h_w's decision boundary, and therefore, intuitively, we are less certain about h_w's prediction.

Algorithm 1 A PTAS for agnostically learning halfspaces w.r.t. the uniform distribution

Input: $0 < \eta \le 1$ and access to samples from a distribution $\mathcal{D}$ on $S^{d-1} \times \{\pm 1\}$.
Parameters: $r \in \mathbb{N}$, $\beta > 0$ and $\gamma > 0$.

1: Find, using Awasthi et al. (2014) (theorem 4), a vector $w \in S^{d-1}$ with $\mathrm{Err}_{\mathcal{D}}(h_w) \le \alpha_0 \eta$
2: Let $T = T_{d,\gamma}(w) := \{u \in S^{d-1} : |\langle w, u\rangle| \le \gamma\}$.
3: Find, using Kalai et al. (2005) (theorem 2), $P \in \mathbf{POL}_{r,d}$ with

$$\mathrm{Err}_{\mathcal{D}|_T}(P) \le \mathrm{Err}_{\mathcal{D}|_T}(h_{w^*}) + \min_{P' \in \mathrm{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} + \beta$$

where h_{w^*} is an optimal halfspace classifier w.r.t. $\mathcal{D}$.
4: With probability $\frac{1}{2}$ return h_w, and w.p. $\frac{1}{2}$ return $h(x) = \begin{cases} h_w(x) & |\langle w, x\rangle| > \gamma \\ \mathrm{sign}(P(x)) & |\langle w, x\rangle| \le \gamma \end{cases}$

Theorem 5 (main – detailed) *With appropriate choice of the parameters* r, β, γ *(depending on* $0 < \mu, \eta \le 1$*), algorithm 1 satisfies:*

- *It tolerates noise rate of* $(1-\mu)\eta$.
- *It runs in time* $\mathrm{poly}\left(d^{\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}}, \frac{1}{\eta}\right)$.
- *Its label complexity is* $\mathrm{poly}\left(d^{\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}}, \log\left(\frac{1}{\eta}\right)\right)$.

Proof outline. To prove theorem 5, we must show that we can choose the parameters so that the time and label complexity are as stated, and under the assumption that $\mathrm{Err}_{\mathcal{D}}(h_{w^*}) \le (1-\mu)\eta$, the error of the returned classifier satisfies $\mathrm{Err}_{\mathcal{D}}(h) \le \eta$. Below, we explain how we do that. We would naturally like to decompose the error into two parts:

$$\begin{aligned}
\mathrm{Err}_{\mathcal{D}}(h) &= \Pr_{(x,y)\sim\mathcal{D}}(x \notin T) \cdot \mathrm{Err}_{\mathcal{D}|_{T^c\times\{\pm1\}}}(h) + \Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_{T\times\{\pm1\}}}(h) \\
&= \Pr_{(x,y)\sim\mathcal{D}}(x \notin T) \cdot \mathrm{Err}_{\mathcal{D}|_{T^c\times\{\pm1\}}}(h_w) + \Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_{T\times\{\pm1\}}}(P) \qquad (3)
\end{aligned}$$

We first handle the former summand using a localization lemma (lemma 6 below). We show that for $\gamma = \Theta\left(\frac{\eta\sqrt{\log\left(\frac{1}{\mu}\right)}}{\sqrt{d}}\right)$, the probability that $h_w(x) \ne h_{w^*}(x)$ outside the strip T, is $\le \frac{\mu\eta}{2}$. Hence, on the complement of T, the returned classifier, that coincides with h_w, is as good as h_*, up to an additive error of $\frac{\mu\eta}{2}$. Concretely,

$$\Pr_{(x,y)\sim\mathcal{D}}(x \notin T) \cdot \mathrm{Err}_{\mathcal{D}|_{T^c\times\{\pm1\}}}(h_w) \le \Pr_{(x,y)\sim\mathcal{D}}(x \notin T) \cdot \mathrm{Err}_{\mathcal{D}|_{T^c\times\{\pm1\}}}(h_{w^*}) + \frac{\mu\eta}{2}\,. \qquad (4)$$

It remains to handle the latter summand in equation (4). It is enough to show that

$$\Pr_{(x,y)\sim T}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_{T\times\{\pm1\}}}(P) \le \Pr_{(x,y)\sim T}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_{T\times\{\pm1\}}}(h_{w^*}) + \frac{\mu\eta}{2} \qquad (5)$$

Indeed, in that case it follows from equations (3), (4) and (5) that

$$\begin{aligned}
\mathrm{Err}_{\mathcal{D}}(h) &\le \Pr_{(x,y)\sim\mathcal{D}}(x \notin T) \cdot \mathrm{Err}_{\mathcal{D}|_{T^c\times\{\pm1\}}}(h_{w^*}) + \Pr_{(x,y)\sim T}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_{T\times\{\pm1\}}}(h_{w^*}) + \mu\eta \\
&= \mathrm{Err}_{\mathcal{D}}(h_{w^*}) + \mu\eta \le (1-\mu)\eta + \mu\eta = \eta\,.
\end{aligned}$$

To prove equation (5) we first note that $\Pr_{(x,y)\sim\mathcal{D}}(x \in T) = \Theta\left(\eta\sqrt{\log\left(\frac{1}{\mu}\right)}\right)$. Hence, it is enough to show that for suitable choice of r and β, $\mathrm{Err}_{\mathcal{D}|_{T\times\{\pm1\}}}(P) \le \mathrm{Err}_{\mathcal{D}|_{T\times\{\pm1\}}}(h_{w^*}) + \frac{\mu}{C\sqrt{\log\left(\frac{1}{\mu}\right)}}$ for large enough constant $C > 0$. By theorem 2, it is enough to choose $\beta = \frac{\mu}{2C\sqrt{\log\left(\frac{1}{\mu}\right)}}$, and large enough r so that $\min_{P'\in\mathrm{POL}_{r,d}} \|h - P'\|_{1,\mathcal{D}|_{T\times\{\pm1\}}} \le \frac{\mu}{2C\sqrt{\log\left(\frac{1}{\mu}\right)}}$.

As we show, $r = O\left(\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}\right)$ suffices. To do that, by equation (2), it is enough to find a polynomial of degree $O\left(\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}\right)$ that approximates the sign function up to an ℓ_1-error of $\frac{\mu}{2C\sqrt{\log\left(\frac{1}{\mu}\right)}}$ w.r.t the distribution $(\mathcal{D}|_{T\times\{\pm1\}})_{w^*}$. This is done in section A, in three steps:

1. We first (section A.1) show how to find polynomials that approximate the sign function on all the points of a given segment $[-a, a]$, except the area that is very close to the origin, say $[-\epsilon, \epsilon]$. To this end, we invoke Jackson's theorem (theorem 3) to find a polynomial

that roughly (up to an error of, say, 0.1) approximates the sign function on the mentioned regime. Namely, we find a polynomial p of degree $O\left(\frac{a}{\epsilon}\right)$ that maps $[-a,-\epsilon]$ (resp. $[\epsilon, a]$) to $[-1.1,-0.9]$ (resp. $[0,9,1.1]$). To move from accuracy of 0.1 to accuracy of some small $\tau > 0$, we compose p with another polynomial r that maps $[-1.1,-0.9]$ (resp. $[0.9,1.1]$) to $[-1-\tau,-1+\tau]$ (resp. $[1-\tau,1+\tau]$). Using the Taylor expansion of the the *error function* $\mathrm{erf}(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$, we show that there exists such r of degree $O\left(\log\left(\frac{1}{\tau}\right)\right)$.

2. In the second step (section A.2), we find ℓ_1 approximations for distributions with strong tail bounds (namely, with density function bounded by $2\exp\left(-\frac{x^2}{32}\right)$ on a certain domain). Using step 1 we find polynomials that approximate the sign function in ℓ_∞ on a large area, and use the tail bounds and lemma 13 to neglect the ℓ_1 norm on the complement of that area.

3. In the last step (section A.3), using basic facts about high dimensional spherical geometry, we show that the distribution $(\mathcal{D}|_{T\times\{\pm 1\}})_{w^*}$ have strong enough tail bounds.

1.2. Related work

Upper bounds. Statistical aspects of learning halfspaces have been extensively studied (e.g. Vapnik (1998)). Halfspaces are efficiently learnable in the *realizable case*, when $\mathrm{opt} = 0$. This is done using the ERM algorithm Vapnik (1998) that efficiently find, using linear programming, a halfspace that makes no errors on the given sample. For agnostic, distribution free learning, the best known efficient algorithm Kearns and Li (1988) have an approximation ratio of $O(d)$, and the best known exact algorithm is the naive (exponential time) algorithm that go over all halfspaces and return the one with minimal error on the given sample. Under distributional assumptions, better algorithms are known. Under the uniform distribution, Kalai et al. (2005) and Awasthi et al. (2014) presented efficient algorithms with approximation ratios $\sqrt{\log\left(\frac{1}{\mathrm{opt}}\right)}$ and $O(1)$ respectively. The best known exact algorithm Kalai et al. (2005) runs in time $d^{O\left(\frac{1}{\epsilon^2}\right)}$ (as follows from Diakonikolas et al. (2010b)). For log-concave distributions, Klivans et al. (2009) and Awasthi et al. (2014) presented efficient algorithms with approximation ratios $O\left(\frac{\log\left(\frac{1}{\mathrm{opt}}\right)}{\mathrm{opt}^2}\right)$ and $O\left(\log^2\left(\frac{1}{\mathrm{opt}}\right)\right)$ respectively. The best known exact algorithm Kalai et al. (2005) runs in time $d^{f(\epsilon)}$. In learning halfspaces with margin[4] $\gamma > 0$, best known algorithms Long and Servedio (2011); Birnbaum and Shalev-Shwartz (2012) have approximation ratio of $\frac{1/\gamma}{\log(1/\gamma)}$, while the best known exact algorithm Shalev-Shwartz et al. (2011) runs in time $\left(\frac{1}{\epsilon}\right)^{O\left(\frac{\log(1/\gamma)}{\gamma}\right)}$.

Lower bounds. Hardness of (distribution free) agnostic learning of halfspaces is known to follow from several complexity assumptions including hardness of learning parity Kalai et al. (2005) (this result even rules out learning under the uniform distribution on $\{\pm 1\}^d$), hardness of the shortest vector problem Feldman et al. (2006), and hardness of refuting random K-SAT formulas Daniely and Shalev-Shwartz (2014). Hardness of learning sparse parity implies hardness of agnostic learning under the uniform distribution on S^{d-1} Klivans and Kothari (2014). Hardness of agnostic learning of

4. In this problem the distribution is supported in the unit ball, and the algorithm should compete with all classifiers that predict like a halfspace classifier h_w, except that they give no prediction (and therefore err) for instances that are within distance γ of the decision boundary of h_w.

halfspaces with a constant approximation ratio follows from a rather strong complexity assumption on the complexity of refuting random CSP instances Daniely et al. (2014a). For *proper* learning of halfspaces, super constant ($\Omega\left(2^{\log^{1-\tau}(d)}\right)$ for every $\tau > 0$) lower bounds on the best approximation ratio are known, assuming $\mathcal{NP} \neq \mathcal{RP}$ Arora et al. (1993); Guruswami and Raghavendra (2006); Feldman et al. (2006). Finally, lower bounds on concrete families of algorithms were studied in Ben-David et al. (2012); Daniely et al. (2014b)

2. Proof of theorem 5

For localization arguments, we will use the following lemma.

Lemma 6 (localization) *Let $w, w^* \in S^{n-1}$ and let $\mathcal{D}$ be a distribution of $S^{d-1} \times \{\pm 1\}$ such that $\mathcal{D}|_{S^{d-1}}$ is uniform.*

- *We have $\frac{\theta(w,w^*)}{\pi} \le \mathrm{Err}_{\mathcal{D}}(w) + \mathrm{Err}_{\mathcal{D}}(w^*)$.*
- *If $x \in S^{d-1}$ is a uniform vector, then for every $r > 0$,*

$$\Pr\left(h_w(x) \neq h_{w^*}(x) \text{ and } |\langle x, w\rangle| > r\cdot\theta(w,w^*)\right) \le \frac{4\cdot\theta(w,w^*)}{\pi}\exp\left(-\frac{1}{8}r^2 d\right)$$

Proof For the first part we note that $\Pr_{x\sim\mathcal{D}}\left(h_w(x) \neq h_{w^*}(x)\right) = \frac{\theta(w,w^*)}{\pi}$, while on the other hand,

$$\Pr_{x\sim\mathcal{D}}\left(h_w(x) \neq h_{w^*}(x)\right) \le \Pr_{(x,y)\sim\mathcal{D}}\left(h_w(x) \neq y\right) + \Pr_{(x,y)\sim\mathcal{D}}\left(h_{w^*}(x) \neq y\right) \ .$$

For the second part, let $V \subset \mathbb{R}^d$ be the 2-dimensional space spanned by w, w^*, let $P_V : \mathbb{R}^d \to V$ be the orthogonal projection V, and let $B \subset V$ be the ball of radius r around 0. We have

$$|\langle w^*, x\rangle - \langle w, x\rangle| = |\langle w^* - w, P_V(x)\rangle| \le \|w - w^*\|\cdot\|P_V(x)\| \le \theta(w,w^*)\cdot\|P_V(x)\| \ .$$

Therefore, if $P_V(x) \in B$ and $|\langle x, w\rangle| > r\cdot\theta(w,w^*)$ then $h_w(x) = h_{w^*}(x)$. It follows that

$$\begin{aligned}\Pr\left(h_w(x) \neq h_{w^*}(x) \text{ and } |\langle x, w\rangle| > r\cdot\theta(w,w^*)\right) &= \Pr\left(h_w(x) \neq h_{w^*}(x) \mid P_V(x) \notin B\right)\cdot\Pr\left(P_V(x) \notin B\right) \\ &= \frac{\theta(w,w^*)}{\pi}\cdot\Pr\left(P_V(x) \notin B\right) \ .\end{aligned}$$

Finally, let $e_1, e_2 \in V$ be an orthonormal basis. Note that if $|\langle x, e_1\rangle| \le \frac{r}{\sqrt{2}}$ and $|\langle x, e_2\rangle| \le \frac{r}{\sqrt{2}}$ then $P_V(x) \in B$. Hence, we have

$$\Pr\left(P_V(x) \notin B\right) \le \Pr\left(|\langle x, e_1\rangle| > \frac{r}{\sqrt{2}}\right) + \Pr\left(|\langle x, e_2\rangle| > \frac{r}{\sqrt{2}}\right) \le 4\exp\left(-\frac{1}{8}r^2 d\right) \ .$$

Here, the last inequality follows from the well known measure concentration bound according which for every $e \in S^{d-1}$ and $\sigma > 0$ we have $\Pr\left(|\langle x, e\rangle| \ge \sigma\right) \le 2\exp\left(-\frac{1}{4}\sigma^2 d\right)$. ■

To approximate h_{w^*}, we will find low degree ℓ_1 approximation of h_{w^*} w.r.t. $\mathcal{D}|_T$. Such approximations are given in the following two lemmas. The first is from Diakonikolas et al. (2010a) (see a proof in section A. For a stronger version, with $r = O\left(\frac{1}{\tau^2}\right)$, see Diakonikolas et al. (2010b)). The second lemma is established by approximating the sign function (as explained in section 1.1.3) and is given in section A.

Lemma 7 (uniform halfspaces approximation, Diakonikolas et al. (2010a)) *Let $\mathcal{D}$ be the uniform distribution on S^{d-1} and let $w^* \in S^{d-1}$. For every $\tau > 0$ there is $P \in \mathrm{POL}_{r,d}$, for $r = O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ such that $\|h_{w^*} - P\|_{1,\mathcal{D}} < \tau$.*

Lemma 8 (halfspaces approximation on a strip) *Let w, w^* be two vectors with $\theta = \theta(w, w^*)$ and let $\frac{1}{2} > \gamma > 0$. Let $\mathcal{D}$ be the distribution on S^{d-1} that is the restriction of the uniform distribution to $T_{d,\gamma}(w)$. Then, for every $0 < \tau < \frac{\sin(\theta)}{2\gamma\sqrt{d}}$ there is $P \in \mathrm{POL}_{r,d}$, for $r = O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ such that $\|h_{w^*} - P\|_{1,\mathcal{D}} < \tau$.*

Lastly, we will also rely on the following complexity analysis of algorithm 1.

Lemma 9 (complexity analysis) *The runtime of algorithm 1 is* $\mathrm{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\gamma}, \frac{1}{\eta}\right)$ *and the label complexity is* $\mathrm{poly}\left(d^r, \frac{1}{\eta}, \log\left(\frac{1}{\eta}\right)\right)$.

Proof The runtime of step 1 is $\mathrm{poly}\left(d, \frac{1}{\eta}\right)$, while the label complexity is $\mathrm{poly}\left(d, \log\left(\frac{1}{\eta}\right)\right)$. For step 3, we can apply the Kalai et al. (2005) algorithm on $\mathrm{poly}\left(d^r, \frac{1}{\eta}\right)$ examples and labels from the distribution $\mathcal{D}|_T$. We can get these many examples by sampling $\mathrm{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\Pr_{\mathcal{D}}(T\times\{\pm 1\})}\right)$ examples from $\mathcal{D}$ and keep and expose the labels of only the first $\mathrm{poly}\left(d^r, \frac{1}{\beta}\right)$ examples that fell in T. It is not hard to see that $\Pr_{\mathcal{D}}(T \times \{\pm 1\}) \geq \Omega\left(\min\left(\gamma\sqrt{d}, 1\right)\right)$. Hence, the runtime of step 3 is $\mathrm{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\gamma}\right)$. To summarize, the total runtime is $\mathrm{poly}\left(d^r, \frac{1}{\beta}, \frac{1}{\gamma}, \frac{1}{\eta}\right)$ and the label complexity is $\mathrm{poly}\left(d^r, \frac{1}{\beta}, \log\left(\frac{1}{\eta}\right)\right)$. ■

We are now ready to prove theorem 5.

Proof (of theorem 5) We will first deal with the case that $\eta > \frac{1}{2(1+\alpha_0)}$. In that case we won't use localization, that is we will choose $\gamma = 1$ (in that case our algorithm is essentially the algorithm of Kalai et al. (2005)). We will choose $\beta = \frac{\mu\eta}{2}$, and $r = O\left(\frac{\log^2(1/(\mu\eta))}{(\mu\eta)^2}\right) = O\left(\frac{\log^2(1/\mu)}{\mu^2}\right)$ that is large enough so that $\min_{P' \in \mathrm{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} \leq \frac{\mu\eta}{2}$ (this is possible according to lemma 7). It that case, the algorithm will, w.p. $\frac{1}{2}$, return the hypothesis $\mathrm{sign}(P)$ for the polynomial P that was found in step 3. We have $\mathrm{Err}_{\mathcal{D}}(P) \leq \mathrm{Err}_{\mathcal{D}}(h_{w^*}) + \frac{\mu\eta}{2} + \frac{\mu\eta}{2}$. By assumption, $\mathrm{Err}_{\mathcal{D}}(h_{w^*}) \leq (1-\mu)\eta$. Hence, $\mathrm{Err}_{\mathcal{D}}(P) \leq \eta$, as required. It also follows from lemma 9 that the runtime and label complexity are $\mathrm{poly}\left(d^{\frac{\log^2(1/\mu)}{\mu^2}}\right)$ (note that η is bounded from below by a constant) as stated .

Next, we deal with the case that $\eta \leq \frac{1}{2(1+\alpha_0)}$. We will show how to choose $r = \Theta\left(\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}\right)$, $\beta = \theta\left(\frac{\mu}{\sqrt{\log\left(\frac{1}{\mu}\right)}}\right)$ and $\gamma = \Theta\left(\frac{\eta\sqrt{\log\left(\frac{1}{\mu}\right)}}{\sqrt{d}}\right)$ for which the algorithm will have the desired properties. Also, by lemma 9, for such a choice of parameters, the runtime and label complexity are as stated.

Let w^* be the vector defining the optimal halfspace. By assumption, $\mathrm{Err}_{\mathcal{D}}(h_{w^*}) \le (1-\mu)\eta$. Let w be the vector found in step 1, and let P be the polynomial found in step 3. We first claim that we can assume w.l.o.g. that

$$\frac{\theta}{\pi} := \frac{\theta(w, w^*)}{\pi} \ge \mu\eta \; . \tag{6}$$

Indeed, otherwise, we will have

$$\begin{aligned} \mathrm{Err}_{\mathcal{D}}(h_w) &\le \mathrm{Err}_{\mathcal{D}}(h_{w^*}) + \Pr_{(x,y)\sim\mathcal{D}}(h_w(x) \ne h_{w^*}(x)) \\ &= \mathrm{Err}_{\mathcal{D}}(h_{w^*}) + \frac{\theta}{\pi} \le (1-\mu)\eta + \mu\eta < \eta \end{aligned}$$

and in that case the algorithm will return, w.p. $\frac{1}{2}$, a hypothesis with error $\le \eta$, as required.

Let $h(x) = \begin{cases} h_w(x) & |\langle w, x\rangle| > \gamma \\ \mathrm{sign}(P(x)) & |\langle w, x\rangle| \le \gamma \end{cases}$. It is enough to show that $\mathrm{Err}_{\mathcal{D}}(h) \le \eta$. Let $T = T_{d,\gamma}(w) := \{u \in S^{d-1} : |\langle w, u\rangle| \le \gamma\}$. The error of h is

$$\begin{aligned} \mathrm{Err}_{\mathcal{D}}(h) &= \Pr_{(x,y)\sim\mathcal{D}}(h_w(x) \ne y \text{ and } |\langle w, x\rangle| > \gamma) + \Pr_{(x,y)\sim\mathcal{D}}(\mathrm{sign}(P(x)) \ne y \text{ and } |\langle w, x\rangle| \le \gamma) \\ &\le \Pr_{(x,y)\sim\mathcal{D}}(h_w(x) \ne h_{w^*}(x) \text{ and } |\langle w, x\rangle| > \gamma) + \Pr_{(x,y)\sim\mathcal{D}}(h_{w^*}(x) \ne y \text{ and } |\langle w, x\rangle| > \gamma) \\ &\quad + \Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_T}(P) \end{aligned} \tag{7}$$

By the first part of lemma 6 we have

$$\frac{\theta}{\pi} \le \mathrm{Err}_{\mathcal{D}}(h_w) + \mathrm{Err}_{\mathcal{D}}(h_{w^*}) \le (1+\alpha_0)\eta \; . \tag{8}$$

By the second part of lemma 6 we have

$$\begin{aligned} \Pr_{(x,y)\sim\mathcal{D}}(h_w(x) \ne h_{w^*}(x) \text{ and } |\langle w, x\rangle| > \gamma) &\le 4(1+\alpha_0)\eta \exp\left(-\frac{1}{8}\left(\frac{\gamma}{\theta}\right)^2 d\right) \\ &\le 4(1+\alpha_0)\eta \exp\left(-\frac{1}{8}\left(\frac{\gamma}{(1+\alpha_0)\pi\eta}\right)^2 d\right) \end{aligned}$$

Now, by an appropriate choice of $\gamma = \Theta\left(\frac{\eta\sqrt{\log\left(\frac{1}{\mu}\right)}}{\sqrt{d}}\right)$, we get

$$\Pr_{(x,y)\sim\mathcal{D}}(h_w(x) \ne h_{w^*}(x) \text{ and } |\langle w, x\rangle| > \gamma) \le \frac{\mu\eta}{2} \; . \tag{9}$$

We next deal with the term $\Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_T}(P)$. Since $\gamma = \Theta\left(\frac{\eta\sqrt{\log\left(\frac{1}{\mu}\right)}}{\sqrt{d}}\right)$ we have that

$$\Pr_{(x,y)\sim\mathcal{D}}(x \in T) = O\left(\eta \cdot \sqrt{\log\left(\frac{1}{\mu}\right)}\right) \tag{10}$$

Also, by equation (8) and the assumption that $\eta \le \frac{1}{2(\alpha_0+1)}$, we have that $0 \le \theta \le \frac{\pi}{2}$. For this regime, $\sin(\theta) \ge \frac{2\theta}{\pi}$. Hence, by equation (6) we have

$$\frac{\sin(\theta)}{2\gamma\sqrt{d}} \ge \frac{\theta}{\pi\gamma\sqrt{d}} \ge \frac{\mu\eta}{\gamma\sqrt{d}} = \Theta\left(\frac{\mu}{\sqrt{\log(1/\mu)}}\right) \tag{11}$$

By equations (10) and (11) we can choose $\beta = \frac{\mu}{4C\sqrt{\log\left(\frac{1}{\mu}\right)}}$, where $C > 0$ is a universal constant that is large enough so that

$$\beta < \frac{\sin(\theta)}{2\gamma\sqrt{d}} \text{ and } 2\beta \cdot \Pr_{(x,y)\sim\mathcal{D}}(x \in T) \le \frac{\mu\eta}{2} \tag{12}$$

By equation 12 and lemma 8 we can choose $r = \Theta\left(\frac{\log^2\left(\frac{1}{\beta}\right)}{\beta^2}\right) = \Theta\left(\frac{\log^3\left(\frac{1}{\mu}\right)}{\mu^2}\right)$ such that

$$\min_{P' \in \mathrm{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} \le \beta$$

in that case we have

$$\mathrm{Err}_{\mathcal{D}|_T}(P) \le \mathrm{Err}_{\mathcal{D}|_T}(h_{w^*}) + \min_{P' \in \mathrm{POL}_{r,d}} \|h_{w^*} - P'\|_{1,\mathcal{D}|_T} + \beta \le \mathrm{Err}_{\mathcal{D}|_T}(h_{w^*}) + 2\beta\ .$$

Hence,

$$\begin{aligned}
\Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_T}(P) &\le \Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_T}(h_{w^*}) + \Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot 2\beta \\
&\le \Pr_{(x,y)\sim\mathcal{D}}(x \in T) \cdot \mathrm{Err}_{\mathcal{D}|_T}(h_{w^*}) + \frac{\mu\eta}{2} \\
&= \Pr_{(x,y)\sim\mathcal{D}}(h_{w^*}(x) \ne y \text{ and } |\langle w, x\rangle| \le \gamma) + \frac{\mu\eta}{2} \qquad (13)
\end{aligned}$$

By equations (7), (9) and (13) we conclude that

$$\begin{aligned}
\mathrm{Err}_{\mathcal{D}}(h) &\le \frac{\mu\eta}{2} + \Pr_{(x,y)\sim\mathcal{D}}(h_{w^*}(x) \ne y \text{ and } |\langle w, x\rangle| > \gamma) \\
&\quad + \Pr_{(x,y)\sim\mathcal{D}}(h_{w^*}(x) \ne y \text{ and } |\langle w, x\rangle| \le \gamma) + \frac{\mu\eta}{2} \\
&= \mathrm{Err}_{\mathcal{D}}(h_{w^*}) + \mu\eta \le (1-\mu)\eta + \mu\eta = \eta\ .
\end{aligned}$$

■

Acknowledgments

Amit Daniely is a recipient of the Google Europe Fellowship in Learning Theory, and this research is supported in part by this Google Fellowship. The author thanks Pranjal Awasthi, Adam Klivans, Nati Linial, and Shai Shalev-Shwartz for valuable discussions and comments.

References

Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 724–733. IEEE, 1993.

Pranjal Awasthi, Maria-Florina Balcan, and Phil Long. The power of localization for efficiently learning linear separators with noise. In *STOC*, 2014.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

S. Ben-David, D. Loker, N. Srebro, and K. Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *ICML*, 2012.

I. Ben-Eliezer, S. Lovett, and A. Yadin. Polynomial threshold functions: Structure, approximation and pseudorandomness. *Unpublished manuscript*, 2009.

A. Birnbaum and S. Shalev-Shwartz. Learning halfspaces with the zero-one loss: Time-accuracy tradeoffs. In *NIPS*, 2012.

Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning dnf's. In *Arxiv preprint arXiv:1404.3378 v1*, 2014.

Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *STOC*, 2014a.

Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. The complexity of learning halfspaces using generalized linear methods. In *COLT*, 2014b.

Philip J Davis. *Interpolation and approximation*. Courier Dover Publications, 1975.

Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010a.

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

V. Feldman, P. Gopalan, S. Khot, and A.K. Ponnuswami. New results for learning noisy parities and halfspaces. In *In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.

V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of the 47th Foundations of Computer Science (FOCS)*, 2006.

A. Kalai, A.R. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Foundations of Computer Science (FOCS)*, 2005.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. pages 267–280, May 1988. SIAM Journal on Computing.

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

Adam Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space. In *RANDOM*, 2014.

Adam R Klivans, Ryan O'Donnell, and Rocco Servedio. Learning intersections and thresholds of halfspaces. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 177–186. IEEE, 2002.

A.R. Klivans, P.M. Long, and R.A. Servedio. Learning halfspaces with malicious noise. *The Journal of Machine Learning Research*, 10:2715–2740, 2009.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. In *FOCS*, pages 574–579, October 1989.

P.M. Long and R.A. Servedio. Learning large-margin halfspaces with more malicious noise. In *NIPS*, 2011.

Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66, 2010.

S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40:1623–1646, 2011.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Appendix A. Polynomial approximation of the sign function

In this section we will find ℓ_1 approximation of halfspaces. In particular, we will prove lemmas 8 and 7.

A.1. Approximation in "truncated L^∞"

Lemma 10 *Let $a, \gamma, \tau > 0$. There exist a polynomial p of degree $O\left(\frac{1}{\gamma} \cdot \log\left(\frac{1}{\tau}\right)\right)$ such that*

- *For $x \in [-a, a]$, $|p(x)| < 1 + \tau$.*
- *For $x \in [-a, a] \setminus [-\gamma \cdot a, \gamma \cdot a]$, $|p(x) - \mathrm{sign}(x)| < \tau$.*

We will use the following lemma:

Lemma 11 *Let $\tau > 0$. There exist a polynomial p of degree $O\left(\log\left(\frac{1}{\tau}\right)\right)$ such that*

- *For $x \in [-1.5, 1.5]$, $|p(x)| < 1 + \tau$.*
- *For $x \in [-1.5, 1.5] \setminus [-0.5, 0.5]$, $|p(x) - \mathrm{sign}(x)| < \tau$.*

Proof The proof is established by approximating the error function, $\mathrm{erf}(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ by a low degree polynomial. Let $\sigma = 2\sqrt{2\log(\frac{4}{\sqrt{2\pi}\tau})}$. We claim that for every $x > \frac{\sigma}{2}$ we have

$$|\,\mathrm{erf}(x) - 1|, |\,\mathrm{erf}(-x)| \le \frac{\tau}{4}\,. \tag{14}$$

Because $0 \le \mathrm{erf}(x) \le 1$ for all x, and since $\mathrm{erf}(x) = 1 - \mathrm{erf}(-x)$, it is enough to prove that $\mathrm{erf}(x) \ge 1 - \frac{\tau}{4}$. Indeed, we have

$$\begin{aligned} 1 - \mathrm{erf}(x) &= \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-\frac{t^2}{2}} dt \\ &\le \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} t e^{-\frac{t^2}{2}} dt \\ &= \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{t^2}{2}} \Big|_{x}^{\infty} \right] \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ &\le \frac{1}{\sqrt{2\pi}} e^{-\frac{\sigma^2}{2}} = \frac{\tau}{4}\,. \end{aligned}$$

Now, by the Taylor expansion of e^x we have

$$e^{-\frac{x^2}{2}} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{n! 2^n}\,.$$

Integrating element-wise and using the fact that $\mathrm{erf}(0) = \frac{1}{2}$, we have

$$\mathrm{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{n! 2^n (2n+1)}\,.$$

Let r be the $2k$'th Taylor polynomial of erf for $k = \max\{\lceil 2(1.5\sigma)^2 e\rceil, \log_2\left(\frac{4}{\tau}\right)\} = O\left(\log\left(\frac{1}{\tau}\right)\right)$. We have, for $|x| \le 1.5\sigma \le \sqrt{\frac{k}{2e}}$

$$\begin{aligned}
|r(x) - \mathrm{erf}(x)| &\le \frac{2}{\sqrt{\pi}}\sum_{n=k}^{\infty}\frac{|x|^{2n+1}}{n!(2n+1)} \\
&\le \frac{2}{\sqrt{\pi}}\sum_{n=k}^{\infty}\frac{x^{2n}}{n!} \\
&\le \frac{2}{\sqrt{\pi}}\sum_{n=k}^{\infty}\frac{x^{2n}}{\sqrt{2\pi}\left(\frac{n}{e}\right)^n} \\
&\le \frac{\sqrt{2}}{\pi}\sum_{n=k}^{\infty}\left(\frac{x^2 e}{n}\right)^n \\
&\le \frac{\sqrt{2}}{\pi}\sum_{n=k}^{\infty}\left(\frac{1}{2}\right)^n \\
&= \frac{\sqrt{2}}{\pi}\left(\frac{1}{2}\right)^{k-1} \le \left(\frac{1}{2}\right)^k \le \frac{\tau}{4}
\end{aligned}$$

Here, the 4'th inequality follows from the well known fact that $n! \ge \sqrt{2\pi}\left(\frac{n}{e}\right)^n$. Finally, using the last inequality and equation (14), it is not hard to check that the polynomial $p(x) = 2r(\sigma x) - 1$ satisfies the required properties. ■

Proof (of lemma 10) By rescaling, we can assume w.l.o.g. that $a = 1$. Let $\phi : [-1,1] :\to \mathbb{R}$ be the function

$$\phi(x) = \begin{cases} \frac{1}{\gamma}x & |x| \le \gamma \\ 1 & x \ge \gamma \\ -1 & x \le -\gamma \end{cases}$$

By Jackson's Theorem, there is a polynomial $q : [-1,1] \to \mathbb{R}$ of degree $\le \left\lceil \frac{12}{\gamma} \right\rceil$ with $||q - \phi||_{\infty,[-1,1]} \le \frac{1}{2}$. Also, let r be the polynomial from Lemma 11. It is easy to check that, $p = r \circ q$ satisfies the requirement of the Lemma. ■

A.2. Approximations for short tailed distributions

Lemma 12 *Let $\rho : \mathbb{R} \to \mathbb{R}_+$ a density function such that for some $\gamma, \sigma > 0$ we have*

$$\forall x,\ \rho(x) \le \frac{2}{\sigma} \text{ and } \forall |x| > 2\gamma,\ \rho(x) \le \frac{2}{\sigma}\exp\left(-\frac{x^2}{32\sigma^2}\right)$$

Then, for every $0 < \tau \le \frac{\sigma}{2\gamma}$ there is a polynomial of degree[5] $O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ *such that*

$$\int_{-\infty}^{\infty} |p(x) - \mathrm{sign}(x)|\rho(x)dx \le \tau$$

5. The constant in the big-O notation is universal.

We will use the following fact.

Lemma 13 *Ben-Eliezer et al. (2009) Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree $\le r$ for which $|p(x)| \le b$ in the interval $[-a, a]$. Then, for every $|x| \ge a$ we have $|p(x)| \le b \cdot \left|\frac{2x}{a}\right|^r$.*

Proof (of lemma 12) By lemma 10, there is a polynomial p of degree $O\left(r \log\left(1/\tau\right)\right)$ such that

- For $x \in [-r\tau\sigma, r\tau\sigma]$, $|p(x)| < 2$.
- For $x \in \left[-r\tau\sigma, -\frac{\tau\sigma}{100}\right]$, $|p(x)| < \frac{\tau}{100}$.
- For $x \in \left[\frac{\tau\sigma}{100}, r\tau\sigma\right]$, $|p(x) - 1| < \frac{\tau}{100}$.

We have,

$$\begin{aligned}
\int_{-\infty}^{\infty} |p(x) - \text{sign}(x)|\rho(x)dx &= \int_{|x|<\frac{\tau\sigma}{100}} |p(x) - \text{sign}(x)|\rho(x)dx + \int_{\frac{\tau\sigma}{100}\le|x|\le r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx \\
&\quad + \int_{|x|\ge r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx \\
&\le \int_{|x|<\frac{\tau\sigma}{100}} \frac{6}{\sigma}dx + \int_{\frac{\tau\sigma}{100}\le|x|\le r\tau\sigma} \frac{\tau}{100}\rho(x)dx \\
&\quad + \int_{|x|\ge r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx \\
&\le \frac{\tau}{2} + \int_{|x|\ge r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx
\end{aligned}$$

It remains to bound $\int_{|x|\ge r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx$. We will choose $r \ge \frac{1}{\tau^2}$, and therefore we will have $r\tau\sigma \ge \frac{\sigma}{\tau} \ge 2\gamma$. Hence, by lemma 13 we have

$$\begin{aligned}
\int_{|x|\ge r\tau\sigma} |p(x) - \text{sign}(x)|\rho(x)dx &\le \int_{|x|\ge r\tau\sigma} 3\left(\frac{2x}{r\tau\sigma}\right)^r \frac{2}{\sigma}e^{-\frac{x^2}{32\sigma^2}}dx \\
&\le 12\int_{r\tau\sigma}^{\infty} \left(\frac{2x}{r\tau\sigma}\right)^r \frac{1}{\sigma}e^{-\frac{x^2}{32\sigma^2}}dx \\
&= 12\int_{r\tau}^{\infty} \left(\frac{2y}{r\tau}\right)^r e^{-\frac{y^2}{32}}dy \\
&\le 12\int_{r\tau}^{\infty} \left(\left(\frac{2y}{r\tau}\right)^r e^{-\frac{y^2}{64}}\right) e^{-\frac{y^2}{64}}dy
\end{aligned}$$

Now, it is possible to choose $r = \Theta\left(\frac{\log(1/\tau)}{\tau^2}\right)$ such that for all $y > r\tau$ we have $\left(\frac{2y}{r\tau}\right)^r \cdot e^{-\frac{y^2}{64}} \le 1$. For such r, the last expression is bounded by $12\int_{\omega\left(\frac{1}{\tau}\right)}^{\infty} e^{-\frac{y^2}{64}}dy = o(\tau)$. ■

A.3. Approximation on a biased strip: proof of lemma 8

In this section we will find a low degree approximation of halfspaces w.r.t. to the distribution from step 3 of our PTAS. Namely, we will prove lemma 8. Let $\rho_{d,\gamma,\theta} : [-1,1] \to \mathbb{R}_+$ be the projection on w^* of the uniform distribution on $T_{d,\gamma}(w)$. By equation (2), it is enough to find τ-approximation of the sign function in ℓ_1, w.r.t. $\rho_{d,\gamma,\theta}$. Namely, it is enough to prove:

Lemma 14 *There is a univariate polynomial p of degree $r = O\left(\frac{\log^2(1/\tau)}{\tau}\right)$ such that*

$$\int_{-1}^{1} |\operatorname{sign}(x) - p(x)| \rho_{d,\gamma,\theta}(x)dx \le \tau \,.$$

Lemma 14 follows immediately from lemma 12 with $\sigma = \frac{\sin(\theta)}{\sqrt{d}}$, the assumptions that $\gamma < \frac{1}{2}$ and $\tau < \frac{\sin(\theta)}{2\gamma\sqrt{d}}$, and the following bound:

Lemma 15

$$\forall z,\ \rho_{d,\gamma,\theta}(z) \le \frac{\sqrt{d}}{\sin(\theta)\sqrt{1-\gamma^2}}$$

$$\forall |z| \ge \gamma,\ \rho_{d,\gamma,\theta}(z) \le \frac{\sqrt{d}}{\sin(\theta)\sqrt{1-\gamma^2}} \exp\left(-(d-1)\frac{(|z|-\gamma)^2}{4\sin^2(\theta)}\right)$$

To prove lemma 15, we will use an explicit formula for $\rho_{d,\gamma,\theta}$. It will be convenient to introduce some notation. Let $\rho_{d,r} : \mathbb{R} \to \mathbb{R}$ be the density function of the random variable that is the inner product of a fixed unit vector in S^{d-1} and a uniform vector in $r \cdot S^{d-1}$. Clearly,

$$\rho_{d,r}(x) = \frac{1}{r} \cdot \rho_{d,1}\left(\frac{x}{r}\right) \tag{15}$$

We will use the following well known inequality

$$\rho_d(x) \le \sqrt{d} \exp\left(-\frac{x^2 d}{4}\right) \tag{16}$$

Lemma 16 *Let A be the probability of $T_{d,\gamma}(w)$ according to the uniform distribution. We have*

$$\rho_{d,\gamma,\theta}(z) = \frac{1}{A} \int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}(u) \cdot \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}(z-u)\, du$$

Proof Let x be a uniform vector in the strip $T_{d,\gamma}(w)$, and let $y = \langle w^*, x\rangle$. We note that $\rho_{d,\gamma,\theta}$ is the density of y. We write

$$x = \alpha \cdot w + z$$

where $\langle w, z\rangle = 0$. For $(w^*)^\perp = w^* - \langle w^*, w\rangle w$ we have,

$$\begin{aligned} y = \langle w^*, x\rangle &= \alpha \cdot \langle w^*, w\rangle + \langle w^*, z\rangle \\ &= \alpha \cdot \cos(\theta) + \langle (w^*)^\perp, z\rangle \end{aligned}$$

We note that the density function of the distribution of $\alpha \cdot \cos(\theta)$ is given by

$$\tau(u) = \begin{cases} \frac{1}{A}\rho_{d,\cos(\theta)}(u) & |u| \le \gamma \cdot \cos(\theta) \\ 0 & |u| > \gamma \cdot \cos(\theta) \end{cases}$$

Now, given α, z is a uniform vector of norm $\sqrt{1-\alpha^2}$ in the orthogonal complement of w, and $(w^*)^\perp$ is a vector of norm $\sin(\theta)$ in that space. It follows that the density function of $\langle (w^*)^\perp, z\rangle$ given that $\alpha \cdot \cos(\theta) = u$ is $\rho_{d-1,\sin(\theta)\cdot\sqrt{1-\frac{u^2}{\cos^2(\theta)}}} = \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}$. It therefore follows that

$$\rho_{d,\gamma,\theta}(z) = \frac{1}{A}\int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}\left(u\right) \cdot \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}\left(z-u\right) du$$

■

We are now ready to prove lemma 15.

Proof (of lemma 15) Let A be the probability of the strip $T_{d,\gamma}(w)$ according to the uniform distribution on the sphere. We have, using equations (15) and (16),

$$\begin{aligned}
\rho_{d,\gamma,\theta}(z) &= \frac{1}{A}\int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}\left(u\right) \cdot \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}\left(z-u\right) du \\
&\le \frac{1}{A}\int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}\left(u\right) \cdot \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}\left(0\right) du \\
&\le \frac{\sqrt{d-1}}{\sin(\theta)\sqrt{1-\gamma^2}}
\end{aligned}$$

Similarly, for $|z| > \gamma$,

$$\begin{aligned}
\rho_{d,\gamma,\theta}(z) &= \frac{1}{A}\int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}\left(u\right) \cdot \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}\left(z-u\right) du \\
&\le \frac{1}{A}\int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}\left(u\right) \cdot \rho_{d-1,\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}\left(|z|-\gamma\right) du \\
&\le \frac{1}{A\sin(\theta)\sqrt{1-\gamma^2}}\int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}\left(u\right) \cdot \rho_{d-1,1}\left(\frac{|z|-\gamma}{\sqrt{\sin^2(\theta)-\tan^2(\theta)u^2}}\right) du \\
&\le \frac{1}{A\sin(\theta)\sqrt{1-\gamma^2}}\int_{-\gamma\cos(\theta)}^{\gamma\cos(\theta)} \rho_{d,\cos(\theta)}\left(u\right) \cdot \rho_{d-1,1}\left(\frac{|z|-\gamma}{\sin(\theta)}\right) du \\
&= \frac{1}{\sin(\theta)\sqrt{1-\gamma^2}}\rho_{d-1,1}\left(\frac{|z|-\gamma}{\sin(\theta)}\right) \\
&\le \frac{\sqrt{d}}{\sin(\theta)\sqrt{1-\gamma^2}}\exp\left(-(d-1)\frac{(|z|-\gamma)^2}{4\sin^2(\theta)}\right)
\end{aligned}$$

■

Proof (of lemma 7) By equation (2), in is enough to show that the there is a univariate polynomial p of degree $r = O\left(\frac{\log^2(1/\tau)}{\tau^2}\right)$ such that

$$\int_{-1}^{1} |p(x) - \text{sign}(x)| \rho_{d,1}(x) dx \le \tau \ .$$

This, however, follows immediately from lemma 12 and equation (16). ■