A Random-Sampling-Based Algorithm for Learning Intersections of Halfspaces

SANTOSH S. VEMPALA

Georgia Tech

Abstract. We give an algorithm to learn an intersection of k halfspaces in $\mathbf{R}^n$ whose normals span an l-dimensional subspace. For any input distribution with a *logconcave* density such that the bounding hyperplanes of the k halfspaces pass through its mean, the algorithm (ϵ, δ)-learns with time and sample complexity bounded by

$$\left(\frac{nkl}{\epsilon}\right)^{O(l)} \log \frac{1}{\epsilon\delta}.$$

The hypothesis found is an intersection of $O(k \log(1/\epsilon))$ halfspaces. This improves on Blum and Kannan's algorithm for the uniform distribution over a ball, in the time and sample complexity (previously doubly exponential) and in the generality of the input distribution.

Categories and Subject Descriptors: F.2.0 [**Analysis of Algorithms and Problem Complexity**]: General; G.3 [**Probability and Statistics**]

General Terms: Algorithms

Additional Key Words and Phrases: Intersections of halfspaces, random projection, PAC learning, complexity

ACM Reference Format:

Vempala, S. S. 2010. A random-sampling-based algorithm for learning intersections of halfspaces. J. ACM 57, 6, Article 32, (October 2010), 14 pages.
DOI = 10.1145/1857914.1857916 http://doi.acm.org/10.1145/1857914.1857916

1. *Introduction*

In this article, we study the following fundamental problem in learning theory: given points drawn from a distribution in n-dimensional space, with those in an (unknown) intersection of k halfspaces labeled *positive* and the rest *negative*, the problem is to learn a hypothesis that correctly classifies most of the distribution. For $k = 1$, this corresponds to learning a single halfspace (also called a *perceptron*), one of the oldest problems in machine learning [Minsky and Papert 1969; Rosenblatt

The author acknowledges support from NSF awards AF-0915903 and AF-0910584.

Authors' address: Santosh S. Vempala, School of Computer Science, Georgia Tech, 266 Ferst Drive, Atlanta, GA 30332, e-mail: vempala@gatech.edu.

DOI 10.1145/1857914.1857916 http://doi.acm.org/10.1145/1857914.1857916

1962]. It is equivalent to linear programming and hence can be solved in polynomial time. Other solutions, notably the perceptron algorithm, have also been studied in the literature. The intersection of k halfspaces is a natural generalization of a perceptron that corresponds to a two-level neural network used in many machine learning applications. Moreover, any convex concept can be approximated by an intersection of sufficiently many halfspaces.

The complexity of learning an intersection of halfspaces has been widely studied [Baum 1990a; Blum and Rivest 1992; Long and Warmuth 1994]. The *proper learning* version, that is, learning using an intersection of *exactly* k halfspaces is NP-hard even for $k = 2$ [Blum and Rivest 1992; Megiddo 1996]. This raises the question of whether one can solve the problem of learning an intersection of k halfspaces using more general hypothesis classes, polynomial threshold functions or intersections of more than k halfspaces. For the former, Sherstov [2009, 2010] has recently shown that the intersection of even two halfspaces in $\mathbf{R}^n$ cannot be PAC-learned by a polynomial threshold function of degree n. Learning using more than k halfspaces has the following lower bound: it is NP-hard (under randomized reductions) to learn an intersection of k halfspaces using fewer than $kn^{1-\epsilon}$ halfspaces for any $\epsilon > 0$ [Blum and Rivest 1992; Vempala 2004]. The complexity of the general learning problem, that is, using a richer concept class (e.g., an intersection of poly(n, k) halfspaces or a decision tree of halfspaces) is a major open question.

On the other hand, efficient algorithms are known under some assumptions on the input distribution. Baum [1990b] gave an algorithm for learning an intersection of two homogeneous halfspaces (a halfspace is homogeneous if the hyperplane defining it passes through the origin) over any distribution $\mathcal{D}$ that is origin-symmetric, that is, for any $x \in R^n$, $\mathcal{D}(x) = \mathcal{D}(-x)$ (any point x and its reflection through the origin are equally likely). Baum's algorithm was recently shown to work for logconcave distributions [Klivans et al. 2009]. A few years after Baum's work, Blum and Kannan [1993, 1997] made an important breakthrough. They found a polynomial-time algorithm that works for a constant number of halfspaces for the uniform distribution on the unit ball. Their algorithm does not explicitly find a set of halfspaces; instead it gives a procedure which can be evaluated in polynomial-time (for constant k) on a new example and is guaranteed to be probably approximately correct on the uniform distribution. The running time, the number of examples required and the size of the hypothesis reported by their algorithm are all doubly exponential in k, namely $n^{2^{O(k)}}$.

In this article, we present a randomized algorithm (Section 2) with the following guarantees[1]:

—The running time and number of examples required are (singly) exponential in k.

—The algorithm explicitly finds an intersection of $O(k \log(1/\epsilon)$ halfspaces.

—It works for a general class of distributions including any logconcave distribution.

[1] A preliminary version of this article appeared in FOCS 1997 [Vempala 1997]. While the main algorithm here is essentially the same, this complete version corrects errors and extends the analysis to logconcave input distributions.

The algorithm is inspired by the following observation: the true complexity of the problem is determined not by the dimension n or the number of halfspaces k, but by the dimension l of the subspace spanned by the normal vectors to their defining hyperplanes.

We assume that the halfspaces are homogeneous, that is, the hyperplanes defining them pass through the origin. Let the halfspaces be $w_1 \cdot x \geq 0, w_2 \cdot x \geq 0, \ldots, w_k \cdot x \geq 0$. The intersection of these halfspaces is the positive region P:

$$P = \{x \mid w_i \cdot x \geq 0 \text{ for } i = 1, \ldots, k\}.$$

Our goal will be to find a set of normal vectors, so that the intersection of the halfspaces they define will be close to P. For this, we consider the set of all normal vectors that define hyperplanes through the origin that *do not intersect* the positive region P. Formally, it is,

$$P^* = \{v \in \mathbf{R}^n \mid v \cdot x \leq 0 \, \forall x \in P\},$$

and is called the dual cone or *polar* cone of P [Grötschel et al. 1988]. The polar can be defined for any subset of $\mathbf{R}^n$. For a closed cone C, it satisfies $(C^*)^* = C$, that is, the polar of the polar is the original set (see, e.g., Grötschel et al. [1988]).

In our setting, it follows that P^* is the cone at the origin formed by the vectors $w_1, \ldots, w_k$, i.e.,

$$P^* = \left\{ v = -\sum_{j=1}^{k} \alpha_j w_j \mid \forall j,\ \alpha_j \geq 0 \right\}.$$

If we could identify the k-dimensional span of P^*, then we could use an algorithm with running time exponential in the dimension, by trying all "distinct" halfspaces in this low-dimensional space. Identifying this subspace is the major ingredient of the algorithm.

The first step of the algorithm is computing an approximation to P^* (in $\mathbf{R}^n$) from a large sample of points in P. The approximation is the polar C^* of the conical hull C of the sample. To bound the sample complexity of approximating the polar, before computing it, we apply an affine transformation to the sample in order to make it isotropic, that is, to center it at the origin and make its covariance matrix the identity. For a logconcave distribution in $\mathbf{R}^n$, this can be achieved using $\tilde{O}(n)$ sample points. After the isotropic transformation, the number of points needed to ensure that C^* is close to P^* grows exponentially in l (and not exponentially in n). This is the key observation of the algorithm, proved formally in Theorem 4. The proof is based on some simple properties of logconcave distributions.

With C^* in hand, we apply a procedure based on random projection to identify an l-dimensional subspace V close to the span of P^*. Then, we project C^* (which is n-dimensional) to this relevant subspace. Let the projection be $\pi_V(C^*)$. The next step is to choose vectors from V to guarantee that for each w_i there is at least one vector in the sample close to it in angle. We do this by simply considering all points of a sufficiently fine grid in $V \cap B_l$, where B_l is the unit ball in $\mathbf{R}^l$. Finally, we prune the set of candidate vectors using a greedy heuristic.

In the next section, we fix notation and state some useful facts. Then we describe the algorithm (called Polar k-Planes) in detail and proceed with its analysis. We conclude this section with a statement of the main theorem.

THEOREM 1. *Suppose we are given examples in $\mathbf{R}^n$ drawn from an unknown logconcave distribution labeled using an intersection of k halfspaces whose normals span an l-dimensional subspace and whose bounding hyperplanes pass through the mean of the input distribution. Then, for any $\epsilon, \delta > 0$, with probability at least $1 - \delta$, Algorithm Polar k-Planes outputs a set of $O(k \log(1/\epsilon))$ halfspaces whose intersection correctly classifies at least $1 - \epsilon$ of the input distribution. The time and sample complexity of the algorithm are bounded by*

$$n^{l+4} \left(\frac{Ck^2 l}{\epsilon^2} \right)^l \log^3(n/\epsilon\delta)$$

where $C > 1$ is an absolute constant.

We note that if $l = k$, then the complexity is

$$n^{k+4} \left(\frac{Ck^3}{\epsilon^2} \right)^k \log^3(n/\epsilon\delta),$$

dominated by the term n^k.

Klivans et al. [2008] have given an algorithm for learning an intersection of k halfspaces assuming a Gaussian distribution on examples. Their approach is to use a polynomial threshold function and has complexity $n^{O(\log k/\epsilon^4)}$ to learn a hypothesis of similar complexity. For a Gaussian distribution (a special case of logconcave distributions), the dependence of their algorithm on k is much better (and on ϵ is worse). It is not clear if their approach extends to logconcave distributions as their analysis appears to rely crucially on properties of Gaussian distributions.

1.1. PRELIMINARIES.

We use the standard PAC model where examples are points in $\mathbf{R}^n$ and are drawn from an unknown distribution $\mathcal{D}$ and presented along with their labels. For parameters $\epsilon, \delta > 0$, with probability at least $1 - \delta$ the algorithm has to find a hypothesis that has error at most ϵ on $\mathcal{D}$, that is, it correctly classifies a new unlabeled example drawn from $\mathcal{D}$ with probability at least $1 - \epsilon$.

For a labeled set of examples S, we let S^+ denote the subset labeled positive and S^- be the subset labeled negative.

We assume that the label of an example is determined by the intersection of k halfspaces in $\mathbf{R}^n$, such that the normal vectors to the hyperplanes bounding these halfspaces span a subspace of dimension l. We assume that the hyperplanes bounding the halfspaces pass through the origin. Each point $x \in R^n$ is labeled *positive* or *negative* in accordance with the following rule:

$$\ell(x) = \begin{cases} + & \textit{if} \quad Wx \geq 0 \\ - & \quad \text{otherwise.} \end{cases}$$

Here W is a real matrix of rank l with k rows and n columns. Each row w_i is the normal to a halfspace $w_i \cdot x \geq 0$. Formally, the *positive region* P is:

$$P = \{x \in \mathbf{R}^n \ : \ Wx \geq 0\}.$$

We recall an important fact about sample complexity.

FACT 1. *The VC-dimension of the intersection of k homogenous halfspaces in $\mathbf{R}^n$ is nk.*

For a linear subspace V of $\mathbf{R}^n$, we denote orthogonal projection to V by π_V. The unit ball in $\mathbf{R}^n$ is denoted by B_n and signifies the set of all points within unit Euclidean distance from the origin.

Let P^* denote the dual cone of the cone formed by the positive region.

$$P^* = \{v \in \mathbf{R}^n \mid v \cdot x \leq 0 \, \forall x \in P\} = \left\{ v = -\sum_{i=1}^{l} \alpha_i w_i, \mid \forall i, \, \alpha_i \geq 0 \right\}.$$

If P, Q are two cones, then $P^* \subseteq Q^*$ iff $Q \subseteq P$.

For two convex cones K, K' in $\mathbf{R}^k$ we say that K is ϵ-*enclosed* by K' if $K \subseteq K'$, and for every point $x \in K'$ there is some point $y \in K$ such that the angle between the vectors x and y is at most ϵ. Conversely, we say that K' ϵ-*encloses* K.

The *projection width* of a convex body K along a direction (unit vector) v is the length of the 1-dimensional projection of K onto v:

$$\mathsf{width}(K, v) = \max_{x \in K} x \cdot v - \min_{x \in K} x \cdot v.$$

We assume that the input distribution $\mathcal{D}$ has a logconcave density function centered at the origin. This class includes the uniform distribution over any convex body and also any Gaussian. Logconcave densities have several useful properties.

THEOREM 2 [DINGHAS 1957; LEINDLER 1972; PREKOPA 1973a, 1973b]. *The product and convolution of two logconcave functions are logconcave. Any marginal of a logconcave density is logconcave.*

A distribution F is said to be *isotropic* if a random variable X drawn from F satisfies

$$\mathsf{E}(X) = 0 \text{ and } \mathsf{E}(XX^T) = I.$$

The second condition is equivalent to saying that for any unit vector $u \in \mathbf{R}^n$, $\mathsf{E}((u^T x)^2) = 1$. We say that a distribution is near-istropic or θ-isotropic if for any u,

$$\frac{1}{\theta} \leq \mathsf{E}((u^T x)^2) \leq \theta.$$

Using a theorem of Rudelson [1999], and a moment bound in Lovász and Vempala [2007], it is known that the number of samples required to compute an affine transformation that puts any logconcave distribution F ϵ-close to isotropic position is bounded. The following theorem is Corollary A.2 from Kalai and Vempala [2006].

THEOREM 3 [KALAI AND VEMPALA 2006]. *There exists a constant C such that for $m > Cn \log^2 n \log^3(1/\delta)$ independent and identically distributed samples from a logconcave distribution F in $\mathbf{R}^n$, for $\delta < 1/n$, an isotropic transformation of the samples when applied to F puts it in 2-isotropic position with probability at least $1 - \delta$.*

We will also use the following properties of isotropic logconcave functions from Lemmas 5.5 and 5.14 of Lovász and Vempala [2007].

LEMMA 1 [LOVÁSZ AND VEMPALA 2007]. *For any isotropic logconcave density function* $f : \mathbf{R}^n \rightarrow \mathbf{R}_+$,

(a) *If* $n = 1$, *then* $\max f \leq 1$.
(b) $f(0) \geq 2^{-7n}$.
(c) *For any* $v \in \mathbf{R}^n$ *with* $\|v\| \leq 1/9$, $f(v) \geq 2^{-9n\|v\|} f(0)$.

2. *The Algorithm*

The parameters ϵ_i, m, m_1 and N in the description below will be specified later.

```
Polar k-Planes

Input: Access to labeled points in $\mathbf{R}^n$, dimension $l$, error parameters $\epsilon, \delta$.
Output: A set of halfspaces.

(1) Scale. Let $S$ be a set of $m$ samples. Apply an isotropic transformation, i.e., compute
        $z = \mathsf{E}(x),\ A = \mathsf{E}((x - z)(x - z)^T)$
    over $S$ and apply $A^{-1/2}(x - z)$ to each sample point in $S$.
(2) Approximate the dual cone. Let $C$ be the conical hull of the positive examples and
    $C^*$ be the polar cone of $C$:
        $C^* = \{y \in R^n \mid \forall x \in C,\ y \cdot x \leq 0\}$.
(3) Identify the irrelevant subspace. Find directions of small width as follows: pick a
    sample $X$ of $N$ uniform random unit vectors and let
        $x_1 = \operatorname{argmin}_{x \in X} \mathsf{width}(C^* \cap B_n, x)$.
    Similarly, for $i = 2, \ldots, n - l$, using a fresh sample in each iteration, compute
        $x_i = \operatorname{argmin}_{x \in X, x \perp x_1, \ldots, x_{i-1}} \mathsf{width}(C^* \cap B_n, x)$.
(4) Project. Let $V$ be the subspace orthogonal to $\{x^1, \ldots, x^{n-l}\}$; $\pi_V(C^*)$ is the projection
    of $C^*$ to $V$.
(5) Cover the projection. Let $S_1$ be a new set of $m_1$ labeled examples. Choose a set of
    vectors $U$ from $\epsilon_3 \mathbf{Z}^l \cap B_l$ as follows: for $i = 1, 2, \ldots$,
    (a) let $u_i$ be the vector such that $u_i \cdot x \leq 0$ separates the largest number of remaining
        negative examples from at least $1 - \epsilon_4$ of $S_1^+$.
    (b) Discard the negative examples that are separated in this way by $u_i$.
    (c) Stop when the number of remaining negatives is less than an $\epsilon/2$ fraction of $S_1^-$.
(6) Output the set of vectors $U$. Given an unlabeled point $x$, it is labeled positive if
    $(u_i \cdot M)x \leq 0$ for all $u_i \in U$, and negative otherwise.
```

In Step 3 of the algorithm, we need to compute the width of the set $C^* \cap B_n$ along a direction x. To do this, we note that a separation oracle for $C^* \cap B_n$ can be constructed easily from its definition:

$$C^* = \{y \in \mathbf{R}^n \ :\ \forall x \in C,\ x \cdot y \leq 0\}.$$

We observe that since C is a convex cone generated by a sample of points, it suffices to check the constraints for each point in the sample. Given a separation oracle for $C^* \cap B_n$, the width computation consists of two linear optimization problems, both of which can be solved via the Ellipsoid algorithm using a polynomial number of calls to the separation oracle and polynomial additional complexity [Grötschel et al. 1988]. The overall complexity of each linear optimization is $O(mn^3 \log(n/\epsilon))$ for a sample of m points in $\mathbf{R}^n$.

3. *Analysis*

We begin the analysis with a bound on the number of samples from P required to approximate its dual. Then we bound N, the number of random projection trials to find each "irrelevant" direction. This is followed by the guarantees for the covering step.

3.1. Approximating the Dual.

The bound on the number of examples required by the algorithm is dominated by the first step. Let S be the set of examples. Then $|S|$ should be large enough to guarantee that P^* is ϵ_1-enclosed by the dual to the cone formed by S.

Theorem 4. *Let F be an isotropic distribution with a logconcave density function f. Let P be an intersection of k homogenous halfspaces whose normals span an l-dimensional subspace; assume P contains a rotational cone of angle ϵ_0. Let S be a random sample of positive examples from F and C be the cone at the origin formed by S. Let C^* be the dual cone of C. For any $0 \le \epsilon_1 \le \epsilon_0$ and $0 \le \delta \le 1/2$, if*

$$|S| \ge 2^{12(l+1)} l^{l/2} \epsilon_1^{-l} \left(nk \log \frac{l}{\epsilon_1} + \log \frac{1}{\delta} \right),$$

then with probability at least $1 - \delta$, the cone P^* is ϵ_1-enclosed by C^*.

Proof. Let $Q \subseteq P$ be the minimal cone ϵ_1-enclosed by P. The cone Q is nonempty by our assumption of a cone of angle $\epsilon_0 \ge \epsilon_1$ in P. Let Q^* be the polar cone of Q. It follows that P^* is ϵ_1-enclosed by Q^*. To see that Q^* is a convex cone, we note that it can be viewed as a union of rotational cones of angle ϵ_1 centered at each point of P^*.

We will show that P^* is ϵ_1-enclosed by C^* by showing that the C^* is contained in Q^*. For this, it is sufficient to show that, with high probability, for each point h on the boundary of Q^*, there is a supporting plane H of C^* which separates h from P^*.

Since h is on the boundary of Q^*, by the definition of the dual cone, $h \cdot x = 0$ gives a supporting hyperplane of Q such that Q lies in the halfspace $h \cdot x \ge 0$. Consider the convex region

$$P' = P \cap \{x \,:\, h \cdot x \le 0\}$$

and take $y \in P'$. Now $y \cdot h \le 0$. On the other hand since the dual of P^* is P and $y \in P$, for any $w \in P^*$, we have $y \cdot w \ge 0$. Thus, $y \cdot x = 0$ is a hyperplane that separates h from P^*.

We bound the probability of choosing such a witness y in our sample S. This is the measure $\mu_f(P')$. Since P' is an intersection of $k + 1$ halfspaces, we can project to their span V to get a density function g and distribution G. Note that the dimension of V is at most $l + 1$. From Theorem 2, we have that g is also logconcave and isotropic. We also have that

$$\mu_g(\pi_V(P')) = \mu_f(P').$$

To bound this measure, we will use two properties: (a) a lower bound on the fraction of the unit sphere taken up by $\pi_V(P')$ and (b) a lower bound on the measure of any line (through the origin).

For the first, we observe that for any unit vector z in Q, all the vectors within angle ϵ_1 of z are contained in P. Suppose not, say there exists a vector z' on the boundary of P whose angle with z is smaller than ϵ_1. Then there is a supporting plane of P at z' with normal $h(z')$, so that upon moving $h(z')$ by an angle less than ϵ_1, the plane intersects Q, that is, $h(z')$ lies outside Q^*. This contradicts the maximality of Q^* which asserts that for for any point in P^*, all points within angle ϵ_1 of it are contained in Q^*. Next, note that $h \cdot x = 0$ is a supporting plane of Q, and let q be a point in Q for which $h \cdot q = 0$. Then, by the previous argument, the cone of all points within angle ϵ_1 of q is contained in P. Further, a half-cone of this angle is contained in P'. Thus, the fraction of the unit sphere B_{l+1} taken up by $\pi_V(P')$ is at least the fraction taken up by a half-cone of angular radius ϵ_1.

From Lemma 1(b) and (c), we get that the density function of an isotropic logconcave distribution in $\mathbf{R}^{l+1}$ satisfies

(i) $f(0) \geq 2^{-7(l+1)}$
(ii) For any v with $|v| \leq 1/9$, $f(v) \geq 2^{-9|v|(l+1)} f(0)$.

Using these two properties, we have,

$$\begin{aligned}\mu_f(P') = \mu_g(\pi_V(P')) &\geq \mathsf{Vol}(P' \cap (1/9)B_{l+1})2^{-8(l+1)} \\ &\geq \frac{1}{2}\left(\frac{2\epsilon_1}{\pi} \cdot \frac{1}{9}\right)^l \left(\frac{e\pi}{l+1}\right)^{(l+1)/2} 2^{-8(l+1)} \\ &\geq 2^{-12l}\epsilon_1^l l^{-l/2} = \mu \text{ (say)} .\end{aligned}$$

This bounds the probability that any single point h on the boundary of Q^* is cut off. To prove it with high probability for every point in the boundary of Q^*, we will use the VC-dimension of the intersection of $k+1$ halfspaces. We would like to show that with high probability, for a sufficiently large sample of points, *every* intersection of $k+1$ halfspaces that has probability mass at least μ will contain at least one point from the sample. For this, consider the hypothesis class of all intersections of $k+1$ halfspaces. Then, by the VC theorem (see, e.g., Theorem A.6 of Vempala [2004]), and using Fact 1, if we consider a sample of size

$$\frac{8}{\mu}\left(n(k+1)\log\frac{48}{\mu} + \log\frac{2}{\delta}\right),$$

the probability that any consistent hypothesis (i.e., one that labels the sample correctly) has error more than μ is less than δ. In other words, every set of $l+1$ halfspaces whose intersection has probability mass (according to F) at least μ will see at least one example in the sample. Using $\mu = 2^{-10l}\epsilon_1^l l^{-l/2}$, we get that the bound in the theorem. □

3.2. IDENTIFYING THE RELEVANT SUBSPACE. In this section, we analyze the procedure to approximately identify the irrelevant subspace and hence the subspace spanned by P^* .

Our goal is to find a direction x_1 such that the projection width of P^* along x_1 is small. Along any direction orthogonal to the span of P^*, the projection width of C^* is at most ϵ_1. We generate random vectors so that one of them is nearly orthogonal to the span of P^*. The minimum projection width of C^* among vectors in such a sample would then be about ϵ_1 or smaller.

In the next lemma, we estimate the probability that x_1 is nearly orthogonal to P^*.

LEMMA 2. *Let V be an l-dimensional subspace of $\mathbf{R}^n$ and $x \in \mathbf{R}^n$ be a random vector with standard Gaussian coordinates.*

(1) *For any $t \leq 1$,*

$$\mathsf{P}(\|\pi_V(x)\| \leq t) \geq \frac{1}{2}\left(\frac{t}{\sqrt{l}}\right)^l.$$

(2) *For $0 \leq \alpha \leq 1$, with probability $1-\delta$, a sample of size $4(\frac{\sqrt{l}}{\alpha})^l \ln \frac{1}{\delta}$ random unit vectors contains a vector x_1 satisfying*

$$\|\pi_V(x_1)\| \leq \frac{\alpha}{\sqrt{n-l}}.$$

PROOF. For the first part, we observe that $y = \pi_V(x)$ is an l-dimensional random vector with coordinates drawn independently from $N(0,1)$. The desired probability is

$$\begin{aligned}
\mathsf{P}\left(\sum_{i=1}^{l} y_i^2 \leq t^2\right) &= \frac{\int_0^t e^{-r^2/2} r^{l-1}\, dr}{\int_0^\infty e^{-r^2/2} r^{l-1}\, dr} \\
&\geq \frac{1}{2}\frac{\int_0^t e^{-r^2/2} r^{l-1}\, dr}{\int_0^{2\sqrt{l}} e^{-r^2/2} r^{l-1}\, dr} \\
&\geq \frac{1}{2} e^{(l-1)/2} \frac{\mathsf{Vol}(tB_l)}{\mathsf{Vol}(2\sqrt{l}B_l)} \\
&\geq \frac{1}{2}\left(\frac{t}{\sqrt{l}}\right)^l.
\end{aligned}$$

For the second part, we consider random Gaussian vectors and separately bound $\|y\|$ from above and $\|x-y\|$ from below to obtain the desired conclusion for random unit vectors. The upper bound on $\|y\|$ comes from the previous part. To bound $\|x-y\|$ from below, we observe that it is a the length of an $(n-l)$-dimensional Gaussian. Therefore,

$$\mathsf{E}(\|x-y\|^2) = n-l$$

and

$$\mathsf{P}(\|x-y\|^2 \leq \frac{1}{2}\sqrt{n-l}) \leq \frac{1}{2}.$$

Thus, with probability at least

$$\frac{1}{4}\left(\frac{t}{\sqrt{l}}\right)^l,$$

a random unit vector x has $\|\pi_V(x)\| \leq 2t/\sqrt{n-l}$, which implies the conclusion of the lemma with $t = \alpha/2$. □

LEMMA 3. *Assume that P^* is $\frac{\epsilon_2}{4\sqrt{n-l}}$-enclosed by C^* for some $0 \leq \epsilon_2 \leq \pi/4$. Let each iteration of the procedure to identify the irrelevant subspace uses*

$$N \geq 4\left(\frac{4\sqrt{l(n-l)}}{\epsilon_2}\right)^l \ln\frac{2n}{\delta}$$

random unit vectors. Let W be the relevant subspace identified. Then, with probability at least $1-\delta/2$, any unit vector u in P^* has a projection $\pi_W(u)$ such that

$$u \cdot \pi_W(u) \geq 1 - \frac{\epsilon_2^2}{4}.$$

PROOF. From Lemma 2, with $\delta/2n$ in place of δ, there will be a vector x^1 in a random sample of

$$N = 4\left(\frac{\sqrt{l}}{\alpha}\right)^l \ln\frac{2n}{\delta}$$

unit vectors such that

$$\pi_V(x_1) \leq \frac{\alpha}{\sqrt{n-l}}$$

where V is the span of P^*. We will set $\alpha = \epsilon_2/4\sqrt{\ln(n-l)}$ at the end.

Next, using the fact that P^* is $\frac{\epsilon_2}{4\sqrt{n-l}}$-enclosed by C^*, for any vector unit vector $u \in C^*$,

$$\begin{aligned} |u \cdot x_1| &\leq |u \cdot \pi_V(x_1)| + |u \cdot \pi_{V^\perp}(x_1)| \\ &\leq \|\pi_V(x_1)\| + \frac{\epsilon_2}{4\sqrt{n-l}} \\ &\leq \frac{\alpha}{\sqrt{n-l}} + \frac{\epsilon_2}{4\sqrt{n-l}}. \end{aligned}$$

Thus, the algorithm chooses a vector x_1 along which the projection width of $C^* \cap B_n$ is at most the above quantity.

We can view the second iteration of the algorithm as first projecting C^* and P^* to the subspace orthogonal to x_1 and then sampling from unit vectors in that subspace. The projected C^* continues to $\epsilon_2/4\sqrt{n-l}$-enclose P^* since the angle between two points cannot increase by projection. We now apply the previous argument in $\mathbf{R}^{n-1}$. Using Lemma 2 $n-l$ times, there exist vectors $x_1, \ldots, x_{n-l}$ in the samples examined by the algorithm, such that for any $1 \leq j \leq n-l$,

$$\|\pi_V(x_j)\| \leq \frac{\alpha}{\sqrt{n+1-l-j}} + \frac{\epsilon_2}{4\sqrt{n-l}}.$$

Now consider any unit vector $u \in C^*$ and its projection $\pi_W(u)$, orthogonal to the span of $x_1, \ldots, x_{n-l}$.

$$u = \pi_W(u) + \sum_{j=1}^{n-l} (u \cdot x_j) x_j.$$

Hence,

$$\begin{aligned}
\|u - \pi_W(u)\|^2 &= \sum_{j=1}^{n-l} (u \cdot x_j)^2 \\
&\leq \sum_{j=1}^{n-l} 2\frac{\alpha^2}{(n+1-l-j)} + 2\frac{\epsilon_2^2}{16(n-l)} \\
&\leq 2\alpha^2 \ln(n-l) + \frac{\epsilon_2^2}{8}.
\end{aligned}$$

We set $\alpha = \epsilon_2/4\sqrt{\ln(n-l)}$ to get a bound of $\epsilon_2^2/4$ above. Thus,

$$u \cdot \pi_W(u) = \|\pi_W(u)\|^2 \geq 1 - \frac{\epsilon_2^2}{4}. \quad \square$$

3.3. PROOF OF THEOREM 1. We are now ready to prove the main theorem. The first step of the algorithm puts the sample in isotropic position and by Theorem 3, with high probability the distribution is 2-isotropic. For the rest of this proof, we will assume the distribution after the first step is isotropic (the reader can verify that the proof readily extends to near-isotropy).

Next we claim that we can assume that the positive region P contains a rotational cone of radius $\epsilon_0 = \epsilon/l$. If not, the measure of P is at most ϵ, and therefore labeling all of space as negative achieves error at most ϵ. To see the bound on the measure of P, we note that if P does not contain a rotational cone of radius ϵ_0, then moving each bounding hyperplane by ϵ_0 to get a smaller halfspace results in a set of halfspaces with an empty intersection. Now, for each halfspace, we can view the distribution projected along the normal to the halfspace as a one-dimensional isotropic logconcave distribution. Using Lemma 1(a), moving the hyperplane in this manner changes the mass of the halfspace by at most ϵ_0, and so the total mass of P is at most $\epsilon_0 l$.

Applying Theorem 4 and Lemma 3, with probability at least $1-\delta/2$, the algorithm identifies an l-dimensional subspace V' such that for any vector $u \in C^*$,

$$u \cdot \pi_{V'}(u) \geq 1 - \frac{\epsilon_2^2}{4}.$$

In particular this holds for the unknown normal vectors $w_1, \ldots w_k$. Let their projections be w_i'.

The algorithm next considers vectors from the set $\epsilon_3 \mathbf{Z}^l \cap B_l$ in the subspace V'. Let $z_1, \ldots, z_k$ be the vectors in this set closest in angle to $w_1, w_2, \ldots, w_k$. Then, we have that for any i, the angle between z_i and w_i is at most $\epsilon_2 + \epsilon_3\sqrt{l}$. We use this to bound the measure of the symmetric difference of the halfspaces $w_i \cdot x \geq 0$ and $z_i \cdot x \geq 0$. Since the distribution on examples, F, is isotropic and logconcave, its projection to the span of w_i and z_i is a 2-dimensional isotropic logconcave function. Thus, the measure of any line through the origin is the value at zero of the marginal orthogonal to the line, and it is bounded by 1 using Lemma 1(a). Thus, the measure of the symmetric difference of the two halfspaces is at most $\epsilon_2 + \epsilon_3\sqrt{l}$. And so the intersection of the k halfspaces $z_i \cdot x \geq 0$ misclassifies at most $(\epsilon_2 + \epsilon_3\sqrt{l})k$ fraction of F. We set

$$\epsilon_2 = \frac{\epsilon}{8k} \text{and } \epsilon_3 = \frac{\epsilon}{8k\sqrt{l}}$$

to get the error to be at most $\epsilon/4$. From Lemma 3, this also fixes

$$\epsilon_1 = \frac{\epsilon_2}{4\sqrt{n-l}} = \frac{\epsilon}{32k\sqrt{n-l}}.$$

From these parameter values, we get

$$N = 4\left(\frac{32k\sqrt{l(n-l)}}{\epsilon}\right)^l \ln(2n/\delta)$$

from Lemma 3 and

$$m = \left(2^{20}\frac{k\sqrt{l(n-l)}}{\epsilon}\right)^l (nk\log(n/\epsilon) + \log(1/\delta))$$

from Theorem 4 as sufficiently large values. The algorithms makes $O(N)$ calls to a linear optimizer over the polar. As discussed earlier, the complexity of each optimization problem is $O(mn^3\log(n/\epsilon))$ giving an overall running time of $O(Nmn^3\log(n/\epsilon))$ up to the projection step. This matches the bound claimed in the theorem and is the dominant term in the complexity of the algorithm.

It remains to show that the greedy procedure used by the algorithm to choose from vectors in $\epsilon_3\mathbf{Z}^l \cap B_l$ finds vectors that achieve comparable error. For any grid vector u considered by the algorithm, it needs to accurately estimate the fraction of positive examples in the halfspace and fraction of negative examples not in the halfspace.

Let S_1 be a fresh sample of examples used for this estimation. From standard VC-dimension bounds, if

$$m_1 = |S_1| \geq \frac{128}{\epsilon}\left(nl\log\frac{48}{\epsilon} + \log\frac{2}{\delta}\right),$$

then every for any halfspace $u \cdot x \geq 0$, the estimate of the measure of the halfspace given by this sample is within $\epsilon/8$ of the true measure.

Our procedure to pick halfspaces is the following: consider grid vectors u which have almost all positive examples in the halfspace $H_u : \{x : u \cdot x \geq 0\}$, that is, $|H_u \cap S| - |H_u \cap S_1^+| \leq \epsilon_4 m_1 = \epsilon m_1/4$, and among these pick the one that has the maximum number of negatives on the other side, that is, $|H_{-u} \cap S^-|$ is maximized. Then, the negative examples in H_{-u} are discarded and the procedure is repeated.

This can be viewed as a set cover problem. The elements are all the negative examples. The sets are vectors u that have almost all positive examples in the halfspace H_u. The vectors $z_1, \ldots, z_k$ give a solution that covers all but an $\epsilon/4$ fraction of the elements. Our algorithm is a greedy algorithm. We use this standard guarantee on the greedy algorithm for maximum coverage [Hochbaum and Pathria

1998]: a greedily chosen collection of k sets covers at least a $1 - (1/e)$ fraction of the elements covered by the optimal collection, and an $k \log r$ size greedy collection covers at least a $1-(1/r)$ fraction of the elements covered by an optimal collection.

LEMMA 4. *Suppose there exist k halfspaces whose intersection correctly classifies* $1 - (\epsilon/4)$ *of the distribution. Then, the intersection of a greedily chosen set of* $2k \log r$ *halfspaces will correctly classify at least* $(1 - (1/r))(1 - (\epsilon/4))$ *fraction of the distribution.*

Setting r to be $\frac{2}{\epsilon}$ gives us a set of $2k \log(2/\epsilon)$ planes that correctly classify $1 - \epsilon$ of the distribution with probability $1 - \delta$. The probability is only over the previous steps of the algorithm guaranteeing the existence of good vectors $z_1, \ldots, z_k$; the final greedy procedure is deterministic.

4. *Discussion*

We have presented an algorithm to learn an intersection of k homogenous halfspaces whose normals span an l-dimensional subspace given labeled examples from any logconcave distribution. The key ingredients of the algorithm are approximating the dual of the intersection and identifying the span of the normals by random projection. This approach seems suitable for any convex low-dimensional concept. Open problems include further improving the dependence on k and extending beyond logconcave distributions.

REFERENCES

BAUM, E. B. 1990a. On learning a union of half spaces. *J. Complexity 6*, 1, 67–101.

BAUM, E. B. 1990b. Polynomial time algorithms for learning neural nets. In *COLT*. 258–272.

BLUM, A., AND KANNAN, R. 1993. Learning an intersection of k halfspaces over a uniform distribution. In *Proceedings of the Conference on Foundations of Computer Science*. 312–320.

BLUM, A., AND KANNAN, R. 1997. Learning an intersection of a constant number of halfspaces over a uniform distribution. *J. Comput. Syst. Sci. 54*, 2, 371–380.

BLUM, A., AND RIVEST, R. L. 1992. Training a 3-node neural network is np-complete. *Neural Networks 5*, 1, 117–127.

DINGHAS, A. 1957. Uber eine klasse superadditiver mengenfunktionale vonbrunn-minkowski-lusternikschem typus. *Math. Zeitschr. 68*, 111–125.

GRÖTSCHEL, M., LOVÁSZ, L., AND SCHRIJVER, A. 1988. *Geometric Algorithms and Combinatorial Optimization*. Springer.

HOCHBAUM, D. S., AND PATHRIA, A. 1998. Analysis of the greedy approach in problems of maximum k-coverage. *Naval Res. Quart. 45*, 615–627.

KALAI, A. T., AND VEMPALA, S. 2006. Simulated annealing for convex optimization. *Math. Oper. Res. 31*, 2, 253–266.

KLIVANS, A. R., LONG, P. M., AND TANG, A. K. 2009. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *APPROX-RANDOM*. 588–600.

KLIVANS, A. R., O'DONNELL, R., AND SERVEDIO, R. A. 2008. Learning geometric concepts via gaussian surface area. In *Proceedings of the Conference on Foundations of Computer Science*. 541–550.

LEINDLER, L. 1972. On a certain converse of Hölder's inequality ii. *Acta Sci. Math. Szeged 33*, 217–223.

LONG, P. M., AND WARMUTH, M. K. 1994. Composite geometric concepts and polynomial predictability. *Inf. Comput. 113*, 2, 230–252.

LOVÁSZ, L., AND VEMPALA, S. 2007. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms 30*, 3, 307–358.

MEGIDDO, N. 1996. On the complexity of polyhedral separability. Tech. Rep. RJ 5252, *IBM Almaden Research Center*.

MINSKY, M., AND PAPERT, S. 1969. *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

PREKOPA, A. 1973a. Logarithmic concave measures and functions. *Acta Sci. Math. Szeged 34*, 335–343.

PREKOPA, A. 1973b. On logarithmic concave measures with applications to stochastic programming. *Acta Sci. Math. Szeged 32*, 301–316.

ROSENBLATT, F. 1962. *Principles of Neurodynamics*. Spartan Books, Washington, DC.

RUDELSON, M. 1999. Random vectors in the isotropic position. *J. Funct. Anal. 164*, 1, 60 – 72.

SHERSTOV, A. A. 2009. The intersection of two halfspaces has high threshold degree. In *Proceedings of the Conference on Foundations of Computer Science*. 343–362.

SHERSTOV, A. A. 2010. Optimal bounds for sign-representing the intersection of two halfspaces by polynomials. In *Proceedings of the Symposium on Theorey of Computing*. 523–532.

VEMPALA, S. 1997. A random sampling based algorithm for learning the intersection of half-spaces. In *Proceedings of the Conference on Foundations of Computer Science*. 508–513.

VEMPALA, S. S. 2004. *The Random Projection Method*. AMS.

RECEIVED AUGUST 2003; REVISED APRIL 2010; ACCEPTED APRIL 2010