Baum's Algorithm Learns Intersections of Halfspaces with respect to Log-Concave Distributions

Adam R. Klivans[1⋆], Philip M. Long[2], and Alex K. Tang[3]

[1] UT-Austin, klivans@cs.utexas.edu
[2] Google, plong@google.com
[3] UT-Austin, tang@cs.utexas.edu

Abstract. In 1990, E. Baum gave an elegant polynomial-time algorithm for learning the intersection of two origin-centered halfspaces with respect to any symmetric distribution (i.e., any $\mathcal{D}$ such that $\mathcal{D}(E) = \mathcal{D}(-E)$) [3]. Here we prove that his algorithm also succeeds with respect to any mean zero distribution $\mathcal{D}$ with a log-concave density (a broad class of distributions that need not be symmetric). As far as we are aware, prior to this work, it was not known how to efficiently learn any class of intersections of halfspaces with respect to log-concave distributions.
The key to our proof is a "Brunn-Minkowski" inequality for log-concave densities that may be of independent interest.

1 Introduction

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a linear threshold function or *halfspace* if $f(x) = \text{sgn}(w \cdot x)$ for some vector $w \in \mathbb{R}^n$. Algorithms for learning halfspaces from labeled examples are some of the most important tools in machine learning.

While there exist several efficient algorithms for learning halfspaces in a variety of settings, the natural generalization of the problem — learning the intersection of two or more halfspaces (e.g., the concept class of functions of the form $h = f \wedge g$ where f and g are halfspaces) — has remained one of the great challenges in computational learning theory.

In fact, there are no nontrivial algorithms known for the problem of PAC learning the intersection of just two halfspaces with respect to an arbitrary distribution. As such, several researchers have made progress on restricted versions of the problem. Baum provided a simple and elegant algorithm for learning the intersection of two origin-centered halfspaces with respect to any symmetric distribution on $\mathbb{R}^n$ [3]. Blum and Kannan

⋆ Klivans and Tang supported by NSF CAREER Award CCF-643829, an NSF TF Grant CCF-728536, and a Texas Advanced Research Program Award.

[4] and Vempala [16] designed polynomial-time algorithms for learning the intersection of any constant number of halfspaces with respect to the uniform distribution on the unit sphere in $\mathbb{R}^n$. Arriaga and Vempala [2] and Klivans and Servedio [13] designed algorithms for learning a constant number of halfspaces given an assumption that the support of the positive and negative regions in feature space are separated by a margin. The best bounds grow with the margin γ like $(1/\gamma)^{O(\log(1/\gamma))}$.

1.1 Log-Concave Densities

In this paper, we significantly expand the classes of distributions for which we can learn intersections of two halfspaces: we prove that Baum's algorithm succeeds with respect to any mean zero, log-concave probability distribution. We hope that this is a first step towards finding efficient algorithms that can handle intersections of many more halfspaces with respect to a broad class of probability distributions.

A distribution $\mathcal{D}$ is *log-concave* if it has a density f such that $\log f(\cdot)$ is a concave function. Log-concave distributions are a powerful class that capture a range of interesting scenarios: it is known, for example, that the uniform distribution over any convex set is log-concave (if the convex set is centered at the origin, then the corresponding density has mean zero). Hence, Vempala's result mentioned above works for a very special case of log-concave distributions (it is not clear whether his algorithm works for a more general class of distributions). Additionally, interest in log-concave densities among machine learning researchers has been growing of late [10, 7, 1, 9, 14].

There has also been some recent work on learning intersections of halfspaces with respect to the Gaussian distribution on $\mathbb{R}^n$, another special case of a log-concave density. Klivans et al. have shown how to learn (even in the agnostic setting) the intersection of a constant number of halfspaces to any constant error parameter in polynomial-time with respect to any Gaussian distribution on $\mathbb{R}^n$ [12]. Again, it is unclear how to extend their result to log-concave distributions.

1.2 Our approach: Re-analyzing Baum's Algorithm

In this paper, we prove that Baum's algorithm from 1990 succeeds when the underlying probability distribution is not necessarily symmetric, but is log-concave.

Baum's algorithm works roughly as follows. Suppose the unknown target concept C is the intersection of the halfspace H_u defined by $u \cdot x \geq 0$

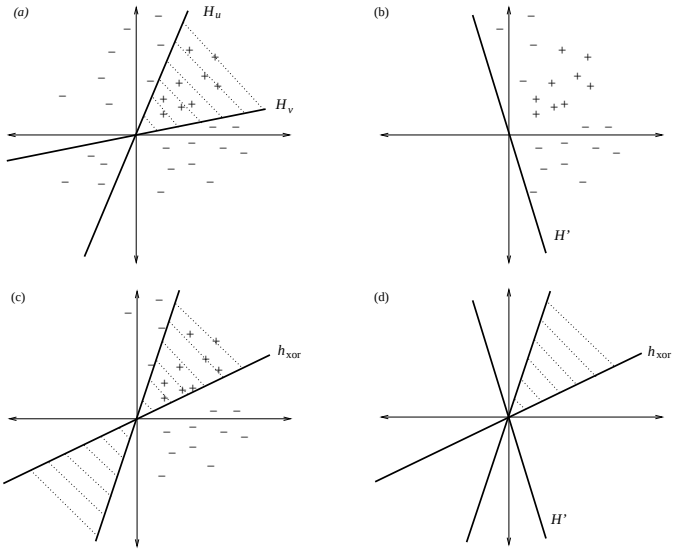

Fig. 1. Baum's algorithm for learning intersections of two halfspaces. (a) The input data, which is labeled using an intersection of two halfspaces. (b) The first step is to find a halfspace containing all the positive examples, and thus, with high probability, almost none of the reflection of the target concept through the origin. (c) The next step is to find a quadratic threshold function consistent with the remaining examples. (d) Finally, Baum's algorithm outputs the intersection of the halfspace found in step b and the classifier found in step c.

and the halfspace H_v defined by $v \cdot x \geq 0$. Note that if $x \in C$ then $(u \cdot x)(v \cdot x) \geq 0$, so that

$$\sum_{ij} u_i v_j x_i x_j \geq 0. \tag{1}$$

If we replace the original features $x_1, \ldots, x_n$ with all products $x_i x_j$ of pairs of features, this becomes a linear inequality. The trouble is that $(u \cdot x)(v \cdot x)$ is also positive if $x \in -C$, i.e., both $u \cdot x \leq 0$ and $v \cdot x \leq 0$. The idea behind Baum's algorithm is to eliminate all the negative examples in $-C$ by identifying a region N in the complement of C (the "negative" region) that, with high probability, includes almost all of $-C$. Then, Baum finds a halfspace in an expanded feature space that is consistent with rest of the examples. (See Figure 1).

To compute N, Baum finds a halfspace H' containing a large set of positive examples in C, and then sets $N = -H'$. Here is where he uses the assumption that the distribution is symmetric: he reasons that

if H' contains a lot of positive examples, then H' contains most of the measure of C, and, since the distribution is symmetric, $-H'$ contains most of the measure of $-C$. Then, if he draws more examples and excludes those in $-H'$, he is unlikely to obtain any examples in $-C$, and for each example x that remains, (1) will hold only if and only if $x \in C$. The output hypothesis classifies an example falling in N negatively, and uses the halfspace in the expanded feature space to classify the remaining examples.

We extend Baum's analysis by showing that, if the distribution is centered and log-concave, then the probability of the region in $-C$ that fails to be excluded by $-H'$ is not much larger than the probability of that part of C that is not covered by H'. Thus, if H' is trained with somewhat more examples, the algorithm can still ensure that $-H'$ fails to cover a small part of $-C$.

Thus, we arrive at the following natural problem from convex geometry: given a cone K whose apex is at the origin in $\mathbb{R}^n$, how does $\Pr(K)$ relate to $\Pr(-K)$ for distributions whose density is log-concave? Were the distribution uniform over a convex set centered at the origin, we could use the Brunn-Minkowski theory to argue that $\Pr(K)$ is always within a factor of n times $\Pr(-K)$ (see the discussion after the proof of Lemma 6). Instead, we are working with a mean zero log-concave distribution, and we do not know of an analog of the Brunn-Minkowski inequality for log-concave densities. Instead, we make use of the fact that the cones we are interested in are very simple and can be described by the intersection of just three halfspaces, and show that $\Pr(K)$ is within a *constant* factor of $\Pr(-K)$. Proving this makes use of tools for analyzing log-concave densities provided by Lovász and Vempala [14].

2 Preliminaries

2.1 VC Theory and sample complexity

We shall assume the reader is familiar with basic notions in computational learning theory such as Valiant's PAC model of learning and VC-dimension (see Kearns & Vazirani for an in-depth treatment [11]).

Theorem 1 ([15, 6]). *Let $\mathcal{C}$ be a class of concepts from the set X to $\{-1, 1\}$ whose VC dimension is d. Let $c \in \mathcal{C}$, and suppose*

$$M(\varepsilon, \delta, d) = O\left(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$$

examples $x_1, \ldots, x_M$ are drawn according to any probability distribution $\mathcal{D}$ over X. Then, with probability at least $1-\delta$, any hypothesis $h \in \mathcal{C}$ that is consistent with c on $x_1, \ldots, x_M$ has error at most ε w.r.t. $\mathcal{D}$.

Lemma 1. *The class of origin-centered halfspaces over $\mathbb{R}^n$ has VC dimension n.*

Lemma 2. *Let $\mathcal{C}$ be a class of concepts from the set X to $\{-1, 1\}$. Let X' be a subset of X, and let $\mathcal{C}'$ be the class of concepts in $\mathcal{C}$ restricted to X'; in other words, let*

$$\mathcal{C}' := \big\{ c|_{X'} \;\big|\; c \in \mathcal{C} \big\}.$$

Then, the VC dimension of $\mathcal{C}'$ is at most the VC dimension of $\mathcal{C}$.

2.2 Log-concave densities

Definition 1 (isotropic, log-concave). *A probability density function f over $\mathbb{R}^n$ is log-concave if $\log f(\cdot)$ is concave. It is isotropic if the covariance matrix of the associated probability distribution is the identity.*

We will use a number of facts that were either stated by Lovász and Vempala, or are easy consequences of their analysis.

Lemma 3 ([14]). *Any halfspace containing the origin has probability at least $1/e$ under a log-concave distribution with mean zero.*

Lemma 4 ([14]). *Suppose f is an isotropic, log-concave probability density function over $\mathbb{R}^n$. Then,*

(a) $f(0) \geq 2^{-7n}$.
(b) $f(0) \leq n(20n)^{n/2}$.
(c) $f(x) \geq 2^{-7n} 2^{-9n\|x\|}$ *whenever* $0 \leq \|x\| \leq 1/9$.
(d) $f(x) \leq 2^{8n} n^{n/2}$ *for every* $x \in \mathbb{R}^n$.
(e) *For every line ℓ through the origin,* $\int_\ell f \leq (n-1)\big(20(n-1)\big)^{(n-1)/2}$.

Proof. Parts a-d are immediate consequences of Theorem 5.14 of [14].

The proof of Part e is like the proof of an analogous lower bound in [14]. Change the basis of $\mathbb{R}^n$ so that ℓ is the x_n-axis, and let h be the marginal over the first $n-1$ variables. Then, by definition,

$$h(x_1, \ldots, x_{n-1}) = \int_\ell f(x_1, \ldots, x_{n-1}, t)\, dt,$$

so that $h(0) = \int_\ell f$. Applying the inequality of Part b gives Part e. $\square$

3 Baum's Algorithm

Let H_u and H_v be the two origin-centered halfspaces whose intersection we are trying to learn. Baum's algorithm for learning $H_u \cap H_v$ is as follows:

1. First, define

$$\begin{aligned} m_1 &:= M(\varepsilon/2, \delta/4, n^2), \\ m_2 &:= M\big(\max\{\delta/(4e\kappa m_1), \varepsilon/2\}, \delta/4, n\big), \text{ and} \\ m_3 &:= \max\{2m_2/\varepsilon, (2/\varepsilon^2)\log(4/\delta)\}, \end{aligned}$$

   where κ is the constant that appears in Lemmas 6 and 7 below.
2. Draw m_3 examples. Let r denote the number of positive examples observed. If $r < m_2$, then output the hypothesis that labels every point as negative. Otherwise, continue to the next step.
3. Use linear programming to find an origin-centered halfspace H' that contains all r positive examples.
4. Draw examples until we find a set S of m_1 examples in H'. (Discard examples in $-H'$.) Then, use linear programming to find a weight vector $w \in \mathbb{R}^{n \times n}$ such that the hypothesis $h_{\text{xor}} : \mathbb{R}^n \to \{-1, 1\}$ given by

$$h_{\text{xor}}(x) := \operatorname{sgn}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} x_i x_j\right)$$

   is consistent with all examples in S.
5. Output the hypothesis $h : \mathbb{R}^n \to \{-1, 1\}$ given by

$$h(x) := \begin{cases} h_{\text{xor}}(x) & \text{if } x \in H', \\ -1 & \text{otherwise.} \end{cases}$$

Outline of proof. In Theorem 2, we prove that Baum's algorithm learns $H_u \cap H_v$ in the PAC model, when the distribution on $\mathbb{R}^n$ is log-concave and has mean zero. Here we give an informal explanation of the proof. In step 3, the algorithm finds a halfspace H' that contains all but a small fraction of the positive examples. In other words, $\Pr\big(H_u \cap H_v \cap (-H')\big)$ is small. This implies that points in $-H'$ have a small chance of being positive, so we can just classify them as negative. To classify points in H', the algorithm learns a hypothesis h_{xor} in step 4. We must show that h_{xor} is a good hypothesis for points in H'. Under a log-concave distribution with mean zero, for *any* intersection of three halfspaces, its probability

mass is at most a constant times the probability of its reflection about the origin; this is proved in Lemma 7. In particular,

$$\Pr\big((-H_u) \cap (-H_v) \cap H'\big) \leq \kappa \Pr\big(H_u \cap H_v \cap (-H')\big) \tag{2}$$

for some constant $\kappa > 0$. Therefore, since $\Pr\big(H_u \cap H_v \cap (-H')\big)$ is small, we can conclude that $\Pr\big((-H_u) \cap (-H_v) \cap H'\big)$ is also small. This implies that, with high probability, points in H' will *not* lie in $(-H_u) \cap (-H_v)$; thus, they must lie in $H_u \cap H_v$, $H_u \cap (-H_v)$, or $(-H_u) \cap H_v$. Such points are classified according to the symmetric difference $H_u \triangle H_v$ restricted to H'. (Strictly speaking, the points are classified according to the negation of the concept $H_u \triangle H_v$ restricted to H'; that is, we need to invert the labels so that positive examples are classified as negative and negative examples are classified as positive.) By Lemmas 1 and 2, together with the fact that h_{xor} can be interpreted as a halfspace over $\mathbb{R}^{n^2}$, the class of such concepts has VC dimension at most n^2. Hence, we can use the VC Theorem to conclude that the hypothesis h_{xor} has low error on points in H'.

Now, we describe the strategy for proving (2). In Lemma 7, we prove that $\Pr(-R) \leq \kappa \Pr(R)$, where R is the intersection of any three origin-centered halfspaces. This inequality holds when the probability distribution is log-concave and has mean zero. First, we prove in Lemma 6 that the inequality holds for the special case when the log-concave distribution not only has mean zero, but is also isotropic. Then, we use Lemma 6 to prove Lemma 7. We consider Lemma 7 to be a Brunn-Minkowski-type inequality for log-concave distributions (see the discussion after the proof of Lemma 6).

To prove Lemma 6, we will exploit the fact that, if R is defined by an intersection of three halfspaces, the probability of R is the same as the probability of R with respect to the marginal distribution over examples projected onto the subspace of $\mathbb{R}^n$ spanned by the normal vectors of the halfspaces bounding R — this is true, roughly speaking, because the dot products with those normal vectors are all that is needed to determine membership in R, and those dot products are not affected if we project onto the subspace spanned by those normal vectors. The same holds, of course, for $-R$.

Once we have projected onto a 3-dimensional subspace, we perform the analysis by proving upper and lower bounds on the probabilities of R and $-R$, and showing that they are within a constant factor of one another. We analyze the probability of R (respectively $-R$) by decomposing it into layers that are varying distances r from the origin. To analyze each

layer, we will use upper and lower bounds on the density of points at a distance r. Since the sizes (even the shapes) of the regions at distance r are the same for R and $-R$, if the densities are close, then the probabilities must be close.

Lemma 5 provides the upper bound on the density in terms of the distance (the lower bound in Lemma 4c suffices for our purposes). We only need the bound in the case $n = 3$, but we go ahead and prove a bound for all n. Kalai, Klivans, Mansour, and Servedio prove a one-dimensional version in Lemma 6 of [9]. We adapt their proof to the n-dimensional case.

Lemma 5. *Let $f : \mathbb{R}^n \to \mathbb{R}^+$ be an isotropic, log-concave probability density function. Then, $f(x) \leq \beta_1 e^{-\beta_2\|x\|}$ for all $x \in \mathbb{R}^n$, where $\beta_1 := 2^{8n}n^{n/2}e$ and $\beta_2 := \frac{2^{-7n}}{2(n-1)(20(n-1))^{(n-1)/2}}$.*

Proof. We first observe that if $\|x\| \leq 1/\beta_2$, then $\beta_1 e^{-\beta_2\|x\|} \geq \beta_1 e^{-1} = 2^{8n}n^{n/2}$. By Lemma 4d, $f(x) \leq \beta_1 e^{-\beta_2\|x\|}$ if $\|x\| \leq 1/\beta_2$. Now, assume there exists a point $v \in \mathbb{R}^n$ such that $\|v\| > 1/\beta_2$ and $f(v) > \beta_1 e^{-\beta_2\|v\|}$. We shall show that this assumption leads to a contradiction. Let $[0, v]$ denote the line segment between the origin 0 and v. Every point $x \in [0, v]$ can be written as a convex combination of 0 and v as follows: $x = \big(1 - \|x\|/\|v\|\big)0 + \big(\|x\|/\|v\|\big)v$. Therefore, the log-concavity of f implies that

$$f(x) \geq f(0)^{1-\|x\|/\|v\|} f(v)^{\|x\|/\|v\|}.$$

We assumed that $f(v) > \beta_1 e^{-\beta_2\|v\|}$. So Lemma 4a implies

$$f(x) > \big(2^{-7n}\big)^{1-\|x\|/\|v\|} \beta_1^{\|x\|/\|v\|} e^{-\beta_2\|x\|}.$$

Because $2^{-7n} \leq 1$ and $1 - \|x\|/\|v\| \leq 1$, we know that $\big(2^{-7n}\big)^{1-\|x\|/\|v\|} \geq 2^{-7n}$. Because $\beta_1 \geq 1$, we know that $\beta_1^{\|x\|/\|v\|} \geq 1$. We can therefore conclude that $f(x) > 2^{-7n}e^{-\beta_2\|x\|}$. Integrating over the line ℓ through 0 and v, we get

$$\int_\ell f \geq \int_{[0,v]} f > \int_0^{\|v\|} 2^{-7n} e^{-\beta_2 r}\, dr = \frac{2^{-7n}}{\beta_2}\left(1 - e^{-\beta_2\|v\|}\right).$$

We assumed that $\|v\| > 1/\beta_2$, so $1 - e^{-\beta_2\|v\|} > 1 - e^{-1}$. Thus,

$$\int_\ell f > \frac{2^{-7n}}{\beta_2}\big(1 - e^{-1}\big) = 2\big(1 - e^{-1}\big)(n-1)\big(20(n-1)\big)^{(n-1)/2}.$$

Since $2\big(1 - e^{-1}\big) > 1$, we conclude that $\int_\ell f > (n-1)\big(20(n-1)\big)^{(n-1)/2}$, but this contradicts Lemma 4e. □

Now we are ready for Lemma 6, which handles the isotropic case.

Lemma 6. *Let R be the intersection of three origin-centered halfspaces in $\mathbb{R}^n$. Assume that the points in $\mathbb{R}^n$ are distributed according to an isotropic, log-concave probability distribution. Then, $\Pr(-R) \leq \kappa \Pr(R)$ for some constant $\kappa > 0$.*

Proof. Let u_1, u_2, and u_3 be normals to the hyperplanes that bound the region R. Then,

$$R = \{x \in \mathbb{R}^n \mid u_1 \cdot x \geq 0 \text{ and } u_2 \cdot x \geq 0 \text{ and } u_3 \cdot x \geq 0\}.$$

Let U be the linear span of u_1, u_2, and u_3. Choose an orthonormal basis (e_1, e_2, e_3) for U and extend it to an orthonormal basis $(e_1, e_2, e_3, \ldots, e_n)$ for all of $\mathbb{R}^n$. Write the components of the vectors x, u_1, u_2, and u_3 in terms of this basis:

$$\begin{aligned} x &= (x_1, x_2, x_3, x_4, \ldots, x_n), \\ u_1 &= (u_{1,1}, u_{1,2}, u_{1,3}, 0, \ldots, 0), \\ u_2 &= (u_{2,1}, u_{2,2}, u_{2,3}, 0, \ldots, 0), \\ u_3 &= (u_{3,1}, u_{3,2}, u_{3,3}, 0, \ldots, 0). \end{aligned}$$

Let $\text{proj}_U(x)$ denote the projection of x onto U; that is, let $\text{proj}_U(x) := (x_1, x_2, x_3)$. Likewise, let $\text{proj}_U(R)$ denote the projection of R onto U; that is, let $\text{proj}_U(R) := \{\text{proj}_U(x) \mid x \in R\}$. Observe that

$$\begin{aligned} x \in R &\Leftrightarrow u_{j,1}x_1 + u_{j,2}x_2 + u_{j,3}x_3 \geq 0 \text{ for all } j \in \{1,2,3\} \\ &\Leftrightarrow \text{proj}_U(x) \in \text{proj}_U(R). \end{aligned} \tag{3}$$

Let f denote the probability density function of the isotropic, log-concave probability distribution on $\mathbb{R}^n$. Let g be the marginal probability density function with respect to (x_1, x_2, x_3); that is, define

$$g(x_1, x_2, x_3) := \int_{\mathbb{R}^{n-3}} \cdots \int f(x_1, x_2, x_3, x_4, \ldots, x_n)\, dx_4 \cdots dx_n.$$

Then, it follows from (3) that

$$\begin{aligned} \Pr(R) &= \int_R \cdots \int f(x_1, x_2, x_3, x_4, \ldots, x_n)\, dx_1 \cdots dx_n \\ &= \iiint_{\text{proj}_U(R)} g(x_1, x_2, x_3)\, dx_1\, dx_2\, dx_3. \end{aligned}$$

Note that g is isotropic and log-concave, because the marginals of an isotropic, log-concave probability density function are isotropic and log-concave (see [14, Theorem 5.1, Lemma 5.2]). Thus, we can use Lemma 4c and Lemma 5 to bound g. The bounds don't depend on the dimension n, because g is a probability density function over $\mathbb{R}^3$. For brevity of notation, let $y := (x_1, x_2, x_3)$. By Lemma 4c, there exist constants κ_1 and κ_2 such that

$$g(y) \geq \kappa_1 e^{-\kappa_2 \|y\|} \quad \text{for } \|y\| \leq 1/9. \tag{4}$$

And by Lemma 5, there exist constants κ_3 and κ_4 such that

$$g(y) \leq \kappa_3 e^{-\kappa_4 \|y\|} \quad \text{for all } y \in \mathbb{R}^3. \tag{5}$$

Let $R' := \text{proj}_U(R) \cap B(0, 1/9)$, where $B(0, 1/9)$ denotes the origin-centered ball of radius $1/9$ in $\mathbb{R}^3$. Use (4) and (5) to derive the following lower and upper bounds:

$$\begin{aligned} \iiint_{R'} \kappa_1 e^{-\kappa_2 \|y\|} \, dy_1 \, dy_2 \, dy_3 &\leq \iiint_{\text{proj}_U(R)} g(x_1, x_2, x_3) \, dx_1 \, dx_2 \, dx_3 \\ &\leq \iiint_{\text{proj}_U(R)} \kappa_3 e^{-\kappa_4 \|y\|} \, dy_1 \, dy_2 \, dy_3. \end{aligned} \tag{6}$$

Recall that

$$\Pr(R) = \iiint_{\text{proj}_U(R)} g(x_1, x_2, x_3) \, dx_1 \, dx_2 \, dx_3.$$

Now, we transform the integrals in the lower and upper bounds in (6) to spherical coordinates. The transformation to spherical coordinates is given by $r := \sqrt{y_1^2 + y_2^2 + y_3^2}$, $\varphi := \arctan\left(\frac{y_2}{y_1}\right)$, $\vartheta := \arccos\left(\frac{y_3}{\sqrt{y_1^2+y_2^2+y_3^2}}\right)$. The determinant of the Jacobian of the above transformation is known to be $r^2 \sin\vartheta$ [5]. Thus (see [5]), inequality (6) becomes

$$\iiint_{R'} \kappa_1 r^2 e^{-\kappa_2 r} \sin\vartheta \, dr \, d\varphi \, d\vartheta \leq \Pr(R) \leq \iiint_{\text{proj}_U(R)} \kappa_3 r^2 e^{-\kappa_4 r} \sin\vartheta dr \, d\varphi \, d\vartheta.$$

Let A denote the surface area of the intersection of $\text{proj}_U(R)$ with the unit sphere S^2; that is, let

$$A := \iint_{\text{proj}_U(R) \cap S^2} \sin\vartheta \, d\varphi \, d\vartheta.$$

Then, it follows that

$$A \int_0^{1/9} \kappa_1 r^2 e^{-\kappa_2 r}\, dr \leq \Pr(R) \leq A \int_0^{\infty} \kappa_3 r^2 e^{-\kappa_4 r} dr.$$

If we let

$$\kappa_5 := \int_0^{1/9} \kappa_1 r^2 e^{-\kappa_2 r}\, dr \quad \text{and} \quad \kappa_6 := \int_0^{\infty} \kappa_3 r^2 e^{-\kappa_4 r}\, dr,$$

then $\kappa_5 A \leq \Pr(R) \leq \kappa_6 A$. By symmetry, $\kappa_5 A \leq \Pr(-R) \leq \kappa_6 A$. Therefore, it follows that $\Pr(-R) \leq (\kappa_6/\kappa_5) \Pr(R)$. □

If the distribution were uniform over a convex set K whose centroid is at the origin, then the proof of Lemma 6 could be modified to show that the probabilities of R and $-R$ are within a factor of n without requiring that R is the intersection of three halfspaces; we would only need that R is a cone (closed under positive rescaling). This could be done by observing that the probability of R is proportional to the average distance of a ray contained in R to the boundary of K. Then we could apply the Brunn-Minkowski inequality (see [8, Lemma 29]) which states that for any direction x, the distance from the origin to the boundary of K in the direction of x is within a factor n of the distance to the boundary of K in the direction $-x$.

In Lemma 6, we assumed that the distribution is isotropic. The next lemma shows that this assumption can be removed (provided that the mean of the distribution is still zero). A key insight is that, under a linear transformation, the image of the intersection of three halfspaces is another intersection of three halfspaces. To prove the lemma, we use a particular linear transformation that brings the distribution into isotropic position. Then, we apply Lemma 6 to the transformed distribution and the image of the three-halfspace intersection.

Lemma 7. *Let R be the intersection of three origin-centered halfspaces in $\mathbb{R}^n$. Assume that the points in $\mathbb{R}^n$ are distributed according to a log-concave probability distribution with mean zero. Then, $\Pr(-R) \leq \kappa \Pr(R)$, where κ is the same constant that appears in Lemma 6.*

Proof. Let X be a random variable in $\mathbb{R}^n$ with a mean-zero, log-concave probability distribution. Let V denote the convariance matrix of X. Let W be a matrix square root of the inverse of V; that is, $W^2 = V^{-1}$. Then, the random variable $Y := WX$ is log-concave and isotropic. (Technically, if the rank of the convariance matrix V is less than n, then V would not

be invertible. But, in that case, the probability distribution degenerates into a probability distribution over a lower-dimensional subspace. We just repeat the analysis on this subspace.) Let $W(R)$ and $W(-R)$ respectively denote the images of R and $-R$ under W. Notice that $W(-R) = -W(R)$. Also, notice that $X \in R \Leftrightarrow Y \in W(R)$ and that $X \in -R \Leftrightarrow Y \in W(-R) = -W(R)$. Let u_1, u_2, and u_3 be normals to the hyperplanes that bound R. Then,

$$\begin{aligned} W(R) &= \big\{ Wx \;\big|\; x \in \mathbb{R}^n \text{ and } u_j^T x \geq 0 \text{ for all } j \in \{1,2,3\} \big\} \\ &= \big\{ y \in \mathbb{R}^n \;\big|\; u_j^T W^{-1} y \geq 0 \text{ for all } j \in \{1,2,3\} \big\} \\ &= \big\{ y \in \mathbb{R}^n \;\big|\; \big((W^{-1})^T u_j\big)^T y \geq 0 \text{ for all } j \in \{1,2,3\} \big\}. \end{aligned}$$

Therefore, $W(R)$ is the intersection of three origin-centered halfspaces, so we can apply Lemma 6 to obtain

$$\Pr(X \in -R) = \Pr\big(Y \in -W(R)\big) \leq \kappa \Pr\big(Y \in W(R)\big) = \kappa \Pr(X \in R).$$

□

Finally, we analyze Baum's algorithm using the probability bound given in Lemma 7.

Theorem 2. *In the PAC model, Baum's algorithm learns the intersection of two origin-centered halfspaces with respect to any mean zero, log-concave probability distribution in polynomial time.*

Proof. If the probability p of observing a positive example is less than ε, then the hypothesis that labels every example as negative has error less than ε; so the algorithm behaves correctly if it draws fewer than m_2 positive examples in this case. If $p \geq \varepsilon$, then by the Hoeffding bound,

$$\Pr(r < m_2) \leq \Pr\left(\frac{r}{m_3} < \frac{\varepsilon}{2}\right) \leq \Pr\left(\frac{r}{m_3} < p - \frac{\varepsilon}{2}\right) \leq e^{-m_3\varepsilon^2/2} \leq \delta/4.$$

Thus, if $p \geq \varepsilon$, then the probability of failing to draw at least m_2 positive examples is at most $\delta/4$. For the rest of this proof, we shall assume that the algorithm succeeds in drawing at least m_2 positive examples.

Observe that the hypothesis output by the algorithm has error

$$\begin{aligned} \operatorname{err}(h) = {} & \Pr\big(-H'\big) \Pr\big(H_u \cap H_v \mid -H'\big) \\ & + \Pr\big(H'\big) \Pr\big(h_{\text{xor}}(x) \neq c(x) \;\big|\; x \in H'\big), \end{aligned} \tag{7}$$

where $c : \mathbb{R}^n \to \{-1, 1\}$ denotes the concept corresponding to $H_u \cap H_v$. First, we give a bound for

$$\begin{aligned}\Pr\big(-H'\big)\Pr\big(H_u \cap H_v \mid -H'\big) &= \Pr\big(H_u \cap H_v \cap (-H')\big) \\ &= \Pr(H_u \cap H_v)\Pr\big(-H' \mid H_u \cap H_v\big).\end{aligned}$$

Notice that $\Pr(-H' \mid H_u \cap H_v)$ is the error of the hypothesis corresponding to H' over the distribution conditioned on $H_u \cap H_v$. But the VC Theorem works for any distribution, so, since H' contains every one of $M\big(\max\{\delta/(4e\kappa m_1), \varepsilon/2\}, \delta/4, n\big)$ random positive examples, it follows from Lemma 1 that, with probability at least $1 - \delta/4$,

$$\Pr\big(-H' \mid H_u \cap H_v\big) \leq \max\left\{\frac{\delta}{4e\kappa m_1}, \frac{\varepsilon}{2}\right\}.$$

Since $\Pr(H_u \cap H_v) \leq 1$, it follows that

$$\Pr\big(H_u \cap H_v \cap (-H')\big) \leq \max\left\{\frac{\delta}{4e\kappa m_1}, \frac{\varepsilon}{2}\right\}.$$

Therefore, the left term in (7) is at most $\varepsilon/2$. All that remains is to bound the right term.

From Lemma 7, it follows that

$$\Pr\big((-H_u) \cap (-H_v) \cap H'\big) \leq \kappa \Pr\big(H_u \cap H_v \cap (-H')\big) \leq \frac{\delta}{4em_1}.$$

By Lemma 3, $\Pr(H') \geq 1/e$. Therefore,

$$\Pr\big((-H_u) \cap (-H_v) \;\big|\; H'\big) = \frac{\Pr\big((-H_u) \cap (-H_v) \cap H'\big)}{\Pr(H')} \leq \frac{\delta}{4m_1}.$$

Thus, each of the m_1 points in S has probability at most $\delta/4m_1$ of being in $(-H_u) \cap (-H_v)$, so with probability at least $1 - \delta/4$, none of the m_1 points are in $(-H_u) \cap (-H_v)$. Thus, each point in $x \in S$ lies in $H_u \cap H_v$, $H_u \cap (-H_v)$, or $(-H_u) \cap (H_v)$; if $x \in H_u \cap H_v$, then x is labeled as positive; if $x \in H_u \cap (-H_v)$ or $x \in (-H_u) \cap H_v$, then x is labeled as negative. In other words, the points in S are classified according to the negation of $H_u \triangle H_v$ restricted to the halfspace H'. Thus, the linear program executed in Step 4 successfully finds a classifier h_{xor} consistent with the examples in S. By Lemma 1 and Lemma 2, the class of symmetric differences of origin-centered halfspaces restricted to H' has VC dimension at most n^2. Therefore, the VC Theorem implies that, with probability at least $1 - \delta/4$,

$$\Pr\big(h_{\text{xor}}(x) \neq c(x) \;\big|\; x \in H'\big) \leq \frac{\varepsilon}{2}.$$

Since $\Pr(H') \leq 1$, the right term in (7) is at most $\varepsilon/2$. Adding up the probabilities of the four ways in which the algorithm can fail, we conclude that the probability that $\mathrm{err}(h) > \varepsilon$ is at most $4(\delta/4) = \delta$. □

References

1. D. Achlioptas and F. McSherry. On spectral learning with mixtures of distributions. *COLT*, 2005.
2. R. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 616–623, 1999.
3. E. Baum. A polynomial time algorithm that learns two hidden unit nets. *Neural Computation*, 2(4):510–522, 1990.
4. A. Blum and R. Kannan. Learning an intersection of a constant number of halfspaces under a uniform distribution. *Journal of Computer and System Sciences*, 54(2):371–380, 1997.
5. E. K. Blum and S. V. Lototsky. *Mathematics of Physics and Engineering*. World Scientific, 2006.
6. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.
7. C. Caramanis and S. Mannor. An inequality for nearly log-concave distributions with applications to learning. *IEEE Transactions on Information Theory*, 53(3):1043–1057, 2007.
8. J. D. Dunagan. *A geometric theory of outliers and perturbation*. PhD thesis, MIT, 2002.
9. A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2005.
10. R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*, pages 444–457, 2005.
11. M. Kearns and U. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, 1994.
12. A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, 2008.
13. A. Klivans and R. Servedio. Learning intersections of halfspaces with a margin. In *Proceedings of the 17th Annual Conference on Learning Theory,*, pages 348–362, 2004.
14. L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.
15. V. Vapnik. *Estimations of dependences based on statistical data*. Springer, 1982.
16. S. Vempala. A random sampling based algorithm for learning the intersection of halfspaces. In *Proc. 38th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 508–513, 1997.