Learning an Intersection of a Constant Number of Halfspaces over a Uniform Distribution

Avrim L. Blum * Ravindran Kannan †

Abstract

We present a polynomial-time algorithm to learn an intersection of a constant number of halfspaces in n dimensions, over the uniform distribution on an n-dimensional ball. The algorithm we present in fact can learn an intersection of an arbitrary (polynomial) number of halfspaces over this distribution, if the subspace spanned by the normal vectors to the bounding hyperplanes has constant dimension. This generalizes previous results for this distribution, in particular a result of Baum [1] who showed how to learn an intersection of 2 halfspaces defined by hyperplanes that pass through the origin. (His results in fact held for a variety of symmetric distributions.) Our algorithm uses estimates of second moments to find vectors in a low-dimensional "relevant subspace." We believe that the algorithmic techniques studied here may be useful in other geometric learning applications.

1 Introduction

Algorithms such as Backpropagation have achieved wide success in training artificial neural networks to perform a variety of interesting tasks. However, a mathematical understanding of the conditions under which various training procedures will produce good results efficiently remains elusive.This is especially true in the difficult case of high-dimensional input spaces. The purpose of this work is to increase such understanding. We consider here the classification setting (the goal of learning is to distinguish positive and negative examples of something), and suppose that examples are classified according to some unknown function that can be described by a very simple type of 2-layer n-input neural network. What we show is that if the examples are distributed uniformly in the n-dimensional ball, then statistical techniques of a "principal-component analysis" flavor can be guaranteed to provide useful information which we then use to produce a polynomial-time learning algorithm.

More specifically, we consider the problem of learning a target function over a high n-dimensional space where the set of positive examples can be described as an intersection of a small (constant) number of halfspaces. This sort of function corresponds to what can be represented by a 2-layer neural network with n inputs, a small number of linear-threshold hidden units, and a single output unit that computes the AND function. We show that if examples are distributed uniformly in an n-dimensional ball and classified by such a function, then one can use second moments to find what

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3891. Email: avrim@cs.cmu.edu.
This material is based upon work supported under a National Science Foundation postdoctoral fellowship.
†Carnegie Mellon University and Bellcore. Email: kannan@theory.cs.yale.edu

we call a "nearly-relevant" vector. This is a vector that is approximately in the span of the normal vectors to the hyperplanes bounding the positive region. (The normal vectors to the bounding hyperplanes are the weight vectors for the hidden units in the 2-layer network representing the target function.) We then use the ability to find nearly-relevant vectors in a polynomial-time algorithm to learn in a "probably-approximately-correct prediction" sense (this will be formally defined later but the essence is the algorithm must achieve an arbitrarily low error rate in prediction though it need not find the halfspace equations), when the number of halfspaces is a *constant* and the distribution is as described above. Our algorithm in fact extends to learn an intersection of an arbitrary (polynomial) number of halfspaces, so long as the dimension of the "relevant space"—the space spanned by the normal vectors to the hyperplanes bounding the positive region—is constant. Although our proofs only hold for the specific distribution on examples mentioned, we believe that these results give insight into using such statistical techniques more generally for "reasonably nice" distributions of examples.

1.1 Previous theoretical work

Most theoretical results known on learning when the target function can be represented by an intersection of halfspaces have been negative. For example, in Valiant's distribution-free learning model [11, 6], even an intersection of 2 halfspaces cannot be learned in polynomial time in the worst case unless **RP=NP** [4, 9]. These results are "representation dependent" in that they assume the learner's hypotheses are also intersections of 2 halfspaces. In the representation-independent "prediction" model (the learner's hypothesis may be any polynomial-time prediction algorithm) some intuitive arguments for the difficulty of learning these functions are given by Baum [2]. Long and Warmuth [7] have given negative results for learning the class of convex polytopes given by their vertices in the "prediction" model as well.

A large part of the reason for the negativity of the above theoretical results is the condition in the model that success in learning must hold for *every* distribution on examples. When one relaxes this condition, the theoretical results become less pessimistic. In particular, Baum [1] showed that an intersection of two homogenous halfspaces (the hyperplanes that define them must pass through the origin) is learnable over any distribution $\mathcal{D}$ on examples that is symmetric in the sense that $\mathcal{D}(x) = \mathcal{D}(-x)$ for all x.

1.2 Our results

We consider here the uniform distribution on the n-dimensional unit ball B_n. This is a quite natural specific distribution for this problem. What we present is an algorithm that can learn in a probably-approximately-correct prediction sense, any intersection of a constant number of halfspaces over this distribution. We also do not require that the hyperplanes bounding the positive region pass through the origin.

Our approach is based upon the following high-level idea. Suppose the halfspaces are $w_1 \cdot x \leq a_1$, $w_2 \cdot x \leq a_2$, ..., $w_k \cdot x \leq a_k$ (x and the w_i are vectors, "$\cdot$" means dot-product, and the a_i are scalars). Then, even though examples x are points in n-dimensional space, all that really matters for their classification are their projections onto the $(\leq k)$-dimensional subspace spanned by the vectors $w_1, \ldots, w_k$. This is because the value of $w_i \cdot x$ is entirely determined by that projection. Let $\mathbf{V} = \text{span}(w_1, \ldots, w_k)$. If we could somehow find $\mathbf{V}$, say by finding a set of basis vectors *not*

necessarily the w_i's, we could project all our examples onto this space and reduce the dimensionality of the learning problem to a constant. We then could learn halfspaces in the constant-dimensional space by standard methods. So, the idea is to "focus attention" on the low-dimensional relevant subspace hidden in the high-dimensional input space. Our algorithm will actually be somewhat different from this high-level intuition, but this is the inspiration for our approach.

As mentioned above, our techniques in fact work more generally to learn in polynomial time an intersection of an arbitrary (polynomial) number of halfspaces, so long as the dimension of the "relevant space" $\mathbf{V}$ is constant.

To illustrate how one might hope to find a "relevant direction" (a unit length vector in $\mathbf{V}$) consider the average position of the positive examples $\mu_{\text{pos}} = \mathbf{E}[x : x \text{ is positive}]$. The vector μ_{pos} can be found to $1/poly(n)$ accuracy by simply taking the average of a large enough sample. Now, suppose v is a unit vector orthogonal to space $\mathbf{V}$. Then, $\mu_{\text{pos}} \cdot v = 0$ because the positive region is symmetric about reflection through the normal hyperplane to v. That is, for any positive example $x = x' + x_v$ where x_v is the component of x along direction v, the example $x' - x_v$ is also positive. So, if μ_{pos} is noticeably far from the origin, then a (good approximation to) μ_{pos} is a (good approximation to) a non-zero vector in $\mathbf{V}$.

While the average μ_{pos} is convenient and easy to examine, unfortunately it might be zero. Because of this, our main object of study will in fact be the second moment. Given a vector v of length 1 and a set of points S, consider $\mathbf{E}_{x \in S}[(v \cdot x)^2]$. This is the average squared distance of points x from the hyperplane $v \cdot x = 0$. Given a set S, the direction v that minimizes $\mathbf{E}_{x \in S}[(v \cdot x)^2]$ can be found in time polynomial in $|S|$ by a standard method of eigenvalue analysis (described in section 3.1). This is superficially like, but different than, the problem of finding a hyperplane of least squared error. What we show (section 3.2) is that if μ_{pos} is sufficiently close to zero, then the direction v that minimizes $\mathbf{E}_{x \in pos}[(v \cdot x)^2]$ (i.e. $\mathbf{E}[(v \cdot x)^2]$ is taken over the uniform distribution on positive examples x) is guaranteed to be a vector in $\mathbf{V}$; moreover, if a few technical conditions are satisfied, then significant deviation from being in $\mathbf{V}$ noticeably increases the second moment. This argument turns out to be quite a bit more difficult than the argument given above for the simple average of the positive points. We believe that this idea of examining directions of minimum second moment should be useful more generally in other learning situations. In fact, our use of minimizing second moments is similar to the statistical technique of "principal component analysis" [10]. In that approach, one looks at directions that *maximize* variance over unlabeled examples in order to gain information about their distribution. In the distribution we consider, all directions have the same variance over the unlabeled examples, but by examining the second moments of the positive (or negative) examples separately, we are able to gain information about the function being learned.

The above two paragraphs describe the basic idea for how we find one (approximation to) a relevant direction. We conjecture, but cannot prove, that the unit vectors $v_1, \ldots, v_k$ where $v_1 = v$ from before and v_i is the unit vector among those orthogonal to $v_1, \ldots, v_{i-1}$ that minimizes $\mathbf{E}_{x \in pos}[(v_i \cdot x)^2]$, will span the relevant space (or an equivalent space to $\mathbf{V}$ if some of the halfspaces are implied by others). If true, and if vectors with a "noticeable" component in an irrelevant direction had noticeably higher second moment, then this would yield a quite simple learning algorithm. What we *can* prove, however, is that the following more complicated and less efficient strategy will succeed. After finding the first relevant direction, our algorithm then continues by looking at "slices" perpendicular to that direction, and employing a recursive approach. The idea here is that sufficiently thin slices can roughly be treated as $(n-1)$-dimensional balls, with a $(k-1)$-dimensional

relevant subspace. In part due to the errors introduced in this approximation, the end hypothesis produced by our algorithm will not be an intersection of halfspaces, but rather a polynomial-time prediction algorithm representing a union of hypotheses for each slice. The running time and number of examples needed by our procedure are doubly exponential in k.

A fairly cleanly-stated theorem regarding the direction of minimum second moment, that applies to more general convex sets and might be useful in other learning applications, is given as Theorem 4 in Section 3.2.

2 Notation, definitions, and preliminaries

For our purposes, an *example* is a point in $\mathbf{R}^n$. A *concept* P is a subset of $\mathbf{R}^n$, and an example x is a positive example of P if $x \in P$ and is a negative example of P otherwise (we use "P" to emphasize that the concept is the set of *positive* examples). A *labeled example* is an example together with its label (positive or negative). Our goal is to learn an unknown target concept from labeled examples.

In this paper, we assume that the target concept is an intersection of ℓ halfspaces in n-dimensional space, where the normal vectors to the bounding hyperplanes of the positive region span a space of a constant dimension k. Let B_n be the n-dimensional ball of radius 1, and let γ_n denote the volume of B_n. We assume that the examples are drawn from the uniform distribution over B_n. Thus, we can define the positive region P to be the intersection of the ℓ halfspaces and B_n. The error of a hypothesis H with respect to the target concept P, for this distribution, is just $\mathrm{vol}((H \Delta P) \cap B_n)/\mathrm{vol}(B_n)$, where $\mathrm{vol}(R)$ denotes the volume of region R. As is standard in computational learning theory, our algorithm is given an error parameter ϵ and a confidence parameter δ and the goal of the algorithm is to produce with probability at least $1-\delta$ a hypothesis with error at most ϵ. We require the hypothesis to be polynomial-time computable (given a point x one can determine whether x is a positive or negative example of the hypothesis in polynomial time) but do not require it to be represented as an intersection of halfspaces.

If c is a scalar, then cB_n is the n-dimensional ball of radius c. It is not hard to see that cB_n has volume $c^n\gamma_n$. If R is some region and A is a (real or vector) function of points x, we write $\mathbf{E}_{x \in R}[A]$ to denote the expectation of A given that x is chosen uniformly from region R. We will sometimes abbreviate this to $\mathbf{E}_R[A]$. If S is a sample of points, then $\mathbf{E}_S[A]$ is the average of A over sample S.

We treat examples both as points and as vectors, and use $||x||$ to denote the *length* of x, which is its (L_2) distance to the origin (its length as a vector). A *direction* is a vector of length 1. A point x is *τ-central* to a region R if a ball of radius τ about x is contained in R. The *mean* or *center of gravity* of a region R is $\mathbf{E}_{x \in R}[x]$.

Suppose $R = \{x \in B_n : w_i \cdot x \leq a_i, \text{ for } i = 1, 2, \ldots, l\}$ and none of the constraints $w_i \cdot x \leq a_i$ is redundant (i.e., dropping any constraint strictly enlarges R). Then, the *relevant subspace of* R, denoted $\mathbf{V}_{\mathrm{rel}}(R)$, is the span of $w_1, \ldots, w_l$. The *irrelevant subspace of* R, denoted $\mathbf{V}_{\mathrm{irrel}}(R)$, is the orthogonal complement of $\mathbf{V}_{\mathrm{rel}}(R)$. I.e., $\mathbf{V}_{\mathrm{irrel}}(R)$ is the collection of all vectors orthogonal to $\mathbf{V}_{\mathrm{rel}}(R)$. Vectors in $\mathbf{V}_{\mathrm{rel}}(R)$ are called *relevant* vectors and vectors in $\mathbf{V}_{\mathrm{irrel}}(R)$ are called *irrelevant* vectors. More generally, if R is some arbitrary region in B_n, we say that a vector v is irrelevant to R if for all pairs of points $x, x' \in B_n$ where $x' = x + cv$ for some scalar c, either both x and x' are in R or both are not. Then, $\mathbf{V}_{\mathrm{irrel}}(R)$ is the span of all irrelevant vectors v for R; again, $\mathbf{V}_{\mathrm{rel}}(R)$ is the orthogonal complement of $\mathbf{V}_{\mathrm{irrel}}(R)$. It is not too hard to see that this definition implies that every vector in $\mathbf{V}_{\mathrm{irrel}}(R)$ is irrelevant, and that this definition is consistent with the previous one. Notice

that non-irrelevant vectors are not necessarily in $\mathbf{V}_{\text{rel}}$. For example, if $R = \{x : x_1 \geq 0, x_2 \geq 0\}$ then any vector with zero first and second components is irrelevant, and $\mathbf{V}_{\text{rel}}(R)$ is the span of the coordinate vectors $\hat{x}_1$ and $\hat{x}_2$. The vector $(1, 1, 1, 0, \ldots, 0)$ is not irrelevant but also not is in $\mathbf{V}_{\text{rel}}(R)$.

If $\mathbf{W}$ is a subspace of $\mathbf{R}^n$ and $x \in \mathbf{R}^n$, let $proj(x, \mathbf{W})$ be the projection of x onto $\mathbf{W}$. Given a set R, we call $proj(x, \mathbf{V}_{\text{rel}}(R))$ the "relevant component" of x and $proj(x, \mathbf{V}_{\text{irrel}}(R))$ the "irrelevant component" of x. Also, we denote the dimension of a space $\mathbf{W}$ by $\dim(\mathbf{W})$. So, $\dim(\mathbf{V}_{\text{rel}}(P)) = k$. Finally, if R is a convex region containing the origin, a "cone" in R is the convex hull of the origin and some patch of the surface of R. An infinitesimally-small cone is the convex hull of the origin and some infinitesimal patch ∂R of the surface of R.

We will make frequent use of the following probabilistic inequalities (Chernoff and Hoeffding bounds). Let $X_1, \ldots, X_m$ be a sequence of m independent $\{0, 1\}$ random variables with $\mathbf{E}[X_i] = p$, and let S be their sum. Then, for $0 \leq \nu \leq 1$, the following inequalities hold:

- $\mathbf{Pr}[S > (1 + \nu)pm] \leq e^{-\nu^2 pm/3}$,
- $\mathbf{Pr}[S < (1 - \nu)pm] \leq e^{-\nu^2 pm/2}$, and
- $\mathbf{Pr}[|\frac{S}{m} - p| > \nu] \leq 2e^{-2m\nu^2}$.

The last inequality holds even if the random variables X_i are identically distributed in $[0, 1]$.

3 Finding one nearly relevant vector

We may assume that $\frac{\epsilon}{2}\gamma_n \leq \text{vol}(P) \leq (1 - \frac{\epsilon}{2})\gamma_n$ since otherwise, by sampling we can notice that "all positive" or "all negative" are sufficiently close hypotheses. We show here how given $0 < \alpha < 1$, with probability at least $1 - \delta$ we can find a direction (unit vector) v such that $||proj(v, \mathbf{V}_{\text{irrel}}(P))|| < \alpha$, in time $poly(n, 1/\alpha, 1/\epsilon, \log(1/\delta))$. The algorithm itself is fairly simple, though the analysis takes a bit of work.

The algorithm is this. We first select a large sample S of positive examples and find the mean μ_S of S. If $||\mu_S|| \geq \epsilon/(16\sqrt{n}(n + 1))$ we return $\mu_S/||\mu_S||$ as our output. Otherwise, we draw a large sample S' of positive examples and compute the direction v that minimizes $\sum_{x \in S'}(v \cdot x)^2$. This computation can be done by finding the eigenvector of least eigenvalue for a matrix defined by the points in S', as described in Section 3.1 below. We then return this vector v as our output. We prove that if S has size $O(\frac{n^4}{\epsilon^2\alpha^2}\log(n/\delta))$ and S' has size $O(\frac{n^7}{\epsilon^6\alpha^4}\log(n/\delta))$, then the unit vector produced will have with high probability a sufficiently small component in the irrelevant space as desired.

The first (easier) part of the argument concerns the goodness of μ_S. Let μ_{pos} denote the mean of P. Since the ball B_n is a bounded region, Hoeffding bounds imply that μ_S and μ_{pos} will likely be close in each coordinate direction, and therefore will be close overall, if S is sufficiently large. In particular, we select S of size $O(\frac{n^4}{\epsilon^2\alpha^2}\log(n/\delta))$ so that with probability $\geq 1 - \delta/2$ the observed mean μ_S is within $\epsilon\alpha/(16\sqrt{n}(n + 1))$ of μ_{pos}. We assume in the following that this holds.

As discussed in the introduction, μ_{pos} lies in $\mathbf{V}_{\text{rel}}(P)$. So, if $||\mu_S||$ is at least $\frac{\epsilon}{16\sqrt{n}(n+1)}$, the vector $\mu_S/||\mu_S||$ is a unit vector with at most an α component in the $\mathbf{V}_{\text{irrel}}(P)$ space as desired and we are done. On the other hand, suppose $||\mu_S|| \leq \frac{\epsilon}{16\sqrt{n}(n+1)}$. In this case, as described above, we draw

a sample S' of positive examples and compute the unit vector v that minimizes $\sum_{x \in S'} (v \cdot x)^2$. We show that the vector found will have the desired properties through a corollary to a main theorem we prove in section 3.2 and a technical lemma. We present and use the lemma and corollary here leaving their proofs to Section 3.2 and the appendix.

Lemma 1 *If $R \subseteq B_n$ is convex and has volume $r\gamma_n$, then the mean μ_R of R is $(\frac{r}{2\sqrt{n}(n+1)})$-central to R.*

Proof: See appendix. ■

Corollary 2 [Corollary to Theorem 4 of Section 3.2] *Let $R \subseteq B_n$ be a convex set with volume $r\gamma_n$, and suppose the origin is τ-central in R. Let v be the unit vector such that $\mathbf{E}_R[(v \cdot x)^2]$ is minimized, and w any other unit vector with $t = ||\text{proj}(w, \mathbf{V}_{\text{irrel}}(R))||$. Then:*

1. *$v \in \mathbf{V}_{\text{rel}}(R)$.*
2. *$\mathbf{E}_R[(w \cdot x)^2] \geq \mathbf{E}_R[(v \cdot x)^2] + \frac{t^2 n \Delta}{k(n+2)}$.*

where $k = \dim(\mathbf{V}_{\text{rel}}(R))$; $\Delta = r(\frac{\tau}{8n^2})$ if $r \leq \frac{1}{2}$, and $\Delta = \frac{(1-r)^2}{r}(\frac{\tau}{8n^2})$ if $r > \frac{1}{2}$.

(The key point is that the second moment in direction w is noticeably larger than the second moment in direction v, so long as t and τ are at least $1/poly$ and r satisfies the conditions given at the start of the section.)

Proof of Corollary 2: See Section 3.2. ■

Now, since $||\mu_S|| \leq \frac{\epsilon}{16\sqrt{n}(n+1)}$, and μ_S is within $\frac{\epsilon\alpha}{16\sqrt{n}(n+1)}$ of μ_{pos}, we have that $||\mu_{\text{pos}}|| \leq \frac{\epsilon}{8\sqrt{n}(n+1)}$. By Lemma 1 and our conditions on vol(P), this means: the origin is τ-central in P for $\tau \geq \frac{(\epsilon/2)}{2\sqrt{n}(n+1)} - \frac{\epsilon}{8\sqrt{n}(n+1)} = \frac{\epsilon}{8\sqrt{n}(n+1)}$. So, Corollary 2 implies that if $||proj(w, \mathbf{V}_{\text{irrel}}(R))|| \geq \alpha$ then,

$$\begin{aligned} \mathbf{E}_P[(w \cdot x)^2] &\geq \mathbf{E}_P[(v \cdot x)^2] + \\ &\quad \left[\frac{\alpha^2 n}{k(n+2)}\right]\left((\frac{\epsilon}{2})^2 \frac{\epsilon}{8\sqrt{n}(n+1)}\right)\left(\frac{1}{8n^2}\right) \\ &= \mathbf{E}_P[(v \cdot x)^2] + f, \\ &\quad \text{for } f = \frac{\alpha^2 \epsilon^3}{256 k n^{3/2}(n+1)(n+2)}. \end{aligned}$$

We now choose S' sufficiently large so that with probability $1 - \delta/2$, every unit vector w has the observed $\mathbf{E}_{S'}[(w \cdot x)^2]$ within $f/3$ of the true $\mathbf{E}_P[(w \cdot x)^2]$. Note that since $\mathbf{E}[(w \cdot x)^2] = \mathbf{E}[\sum_{i,j} w_i w_j x_i x_j] = \sum_{i,j} w_i w_j \mathbf{E}[x_i x_j]$, it is enough to have each observed $\mathbf{E}[x_i x_j]$ close to its true value, where x_i and x_j are the components of x in the ith and jth coordinates. Therefore, by Hoeffding bounds, this will hold with high probability if S' has size $O((n^7/\epsilon^6\alpha^4)\log(n^2/\delta))$.

So, if v is the vector that minimizes $\mathbf{E}_P[(v \cdot x)^2]$ and w is a vector with $||proj(w, \mathbf{V}_{\text{irrel}}(R))|| \geq \alpha$, then

$$\begin{aligned} \mathbf{E}_{S'}[(w \cdot x)^2] &\geq \mathbf{E}_P[(w \cdot x)^2] - f/3 \\ &\geq \mathbf{E}_P[(v \cdot x)^2] + 2f/3 \\ &\geq \mathbf{E}_{S'}[(v \cdot x)^2] + f/3. \end{aligned}$$

Thus, no such direction will be found by minimizing the observed second moment, which is what we wanted. So, we have:

Theorem 3 *There is an algorithm that given $0 < \delta, \epsilon, \alpha < 1$, in time poly$(n, \frac{1}{\alpha}, \frac{1}{\epsilon}, \log(1/\delta))$ finds a direction v such that if $\frac{\epsilon}{2}\gamma_n \leq \text{vol}(P) \leq (\frac{1-\epsilon}{2})\gamma_n$, then with probability $\geq 1-\delta$, $||\text{proj}(v, \mathbf{V}_{\text{irrel}}(P))|| < \alpha$.*

3.1 How to minimize second moment

Given a finite set of points S, we want to find a non-zero vector v that minimizes

$$\sum_{x \in S} \frac{(v \cdot x)^2}{||v||^2}.$$

A standard method for doing this is as follows [10]. Let $\mathbf{A}$ be a matrix with the points in S as row vectors, and let $\mathbf{B} = \mathbf{A}^T\mathbf{A}$. So, equivalently, we want the column vector $\mathbf{v}$ such that $\mathbf{v}^T\mathbf{B}\mathbf{v}/(\mathbf{v}^T\mathbf{v})$ is minimized. This is just the eigenvector of $\mathbf{B}$ having the least eigenvalue, for the following reason.

By definition, $\mathbf{B}$ is a symmetric matrix, and so it is a standard linear-algebra fact that $\mathbf{B}$ has an orthonormal basis of eigenvectors. Also, by definition of $\mathbf{B}$, all the eigenvalues of $\mathbf{B}$ are real and non-negative. Say $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form an orthonormal basis of eigenvectors for $\mathbf{B}$, and that $\mathbf{B}\mathbf{v}_i = \lambda_i \mathbf{v}_i$. So, for any column vector $\mathbf{v}$, we can write $\mathbf{v} = \sum_i c_i \mathbf{v}_i$ for some scalars c_i. Thus,

$$\frac{\mathbf{v}^T\mathbf{B}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \frac{(\sum_i c_i \mathbf{v}_i^T)(\sum_i c_i \lambda_i \mathbf{v}_i)}{\sum_i c_i^2} = \frac{\sum_i \lambda_i c_i^2}{\sum_i c_i^2}.$$

This is clearly minimized when $c_i = 1$ for λ_i having the least eigenvalue, and $c_{i'} = 0$ for $i' \neq i$.

Thus, all we must do is find the eigenvector of least eigenvalue for an n by n matrix, and this can be done to as many bits of precision as desired in polynomial time by standard techniques.

3.2 Why the minimum second moment direction is useful

In this section we show why the direction of minimum second moment lies in the relevant space. The high level idea is that the planes "constrict" the positive region in relevant directions but not in irrelevant ones. The theorems we show, at least in rough form, are fairly intuitive for low dimensional spaces, but seem to require quite a bit of work to prove in n dimensions.

In the following theorems and lemmas, $R \subseteq B_n$ is a convex region which one can think of as the positive region P. Given an example x, we define $\text{rel}(x) = \textit{proj}(x, \mathbf{V}_{\text{rel}}(R))$ and $\text{irrel}(x) = \textit{proj}(x, \mathbf{V}_{\text{irrel}}(R))$. We begin by stating our main theorem.

Theorem 4 *Let $R \subseteq B_n$ be a convex set and suppose that the origin is τ-central in R. Let v be the unit vector in $\mathbf{V}_{\text{rel}}$ such that $\mathbf{E}_R[(v \cdot x)^2]$ is minimized. Then, for any unit vector $w \in \mathbf{V}_{\text{irrel}}(R)$, we have:*

$$\mathbf{E}_R\left[(w \cdot x)^2\right] \geq \mathbf{E}_R\left[(v \cdot x)^2\right] + \frac{n\Delta}{k(n+2)}$$

where Δ is as defined in Corollary 2 (that is, $\Delta = r(\frac{\tau}{8n^2})$ if $r \leq \frac{1}{2}$, and $\Delta = \frac{(1-r)^2}{r}(\frac{\tau}{8n^2})$ if $r > \frac{1}{2}$) and $k = \dim(\mathbf{V}_{\text{rel}}(R))$.

We prove this theorem through a sequence of lemmas, the main one being Lemma 7. We first prove Corollary 2 given the theorem, then state and prove the lemmas, and then finally prove the theorem.

Proof of Corollary 2. Let v be the direction in $\mathbf{V}_{\text{rel}}(R)$ such that $\mathbf{E}_R[(v \cdot x)^2]$ is minimized (over directions in $\mathbf{V}_{\text{rel}}(R)$). We need only prove claim (2) for this v as claim (1) follows. For direction w, let $w = w' + w''$ where $w' \in \mathbf{V}_{\text{rel}}(R)$ and $w'' \in \mathbf{V}_{\text{irrel}}(R)$. So, $\mathbf{E}_R[(w \cdot x)^2] = \mathbf{E}_R[(w' \cdot x + w'' \cdot x)^2] = \mathbf{E}_R[(w' \cdot x)^2] + \mathbf{E}_R[(w'' \cdot x)^2] + 2\mathbf{E}_R[(w' \cdot x)(w'' \cdot x)]$. The crossterm above equals zero because by definition of $\mathbf{V}_{\text{irrel}}(R)$, for each value of c we have $\mathbf{E}_{x \in R, w' \cdot x = c}[w'' \cdot x] = 0$. So, by Theorem 4 (and recalling that $||w''||^2 = t^2$ and $||w'||^2 = 1 - t^2$) we have:

$$\begin{aligned} \mathbf{E}_R[(w \cdot x)^2] &\geq (1-t^2)\mathbf{E}_R[(v \cdot x)^2] + \\ &\quad t^2 \left[\mathbf{E}_R[(v \cdot x)^2] + \tfrac{n\Delta}{k(n+2)}\right] \\ &\geq \mathbf{E}_R[(v \cdot x)^2] + \tfrac{t^2 n\Delta}{k(n+2)}. \quad \blacksquare \end{aligned}$$

Lemma 5 $\mathbf{E}_{B_n}[||x||^2] = \frac{n}{n+2}$. *Note, by symmetry, this implies that for any v of length 1,* $\mathbf{E}_{x \in B_n}[(v \cdot x)^2] = \frac{1}{n+2}$.

Proof. Straightforward calculation. $\mathbf{E}_{B_n}[||x||^2] = \dfrac{1}{\gamma_n} \displaystyle\int_{r=0}^{r=1} r^2(\gamma_n n r^{n-1})dr = \dfrac{n}{n+2}$. $\blacksquare$

Lemma 6 *If R is a convex region in B_n that contains the origin, then*

$$\mathbf{E}_{x \in R}[||\text{irrel}(x)||^2] \geq \mathbf{E}_{x \in B_n}[||\text{irrel}(x)||^2].$$

Proof. For a point $y \in B_n \cap \mathbf{V}_{\text{rel}}(R)$, let $\text{rel}^{-1}(y)$ be the collection of all points $x \in B_n$ such that $\text{rel}(x) = y$. That is, $\text{rel}^{-1}(y)$ is a "fiber" of all points with y as their component in space $\mathbf{V}_{\text{rel}}(R)$. Notice that for any y, $\mathbf{E}_{x \in \text{rel}^{-1}(y)}[||\text{irrel}(x)||^2]$ is just a function of the distance of y from the origin, and moreover the expectation decreases with increasing $||y||$. In fact, when $||y|| = 1$, the quantity is zero.

Now, let dC be a cone in $B_n \cap \mathbf{V}_{\text{rel}}(R)$ of infinitesimal solid angle $d\theta$ and let $d\tilde{C}$ be the set $\text{rel}^{-1}(dC)$ (extending rel^{-1} in the obvious way to sets). By symmetry, $\mathbf{E}_{d\tilde{C}}[||\text{irrel}(x)||^2] = \mathbf{E}_{B_n}[||\text{irrel}(x)||^2]$; this is because cone dC has the same proportion of points at each distance from the origin as does the entire set $B_n \cap \mathbf{V}_{\text{rel}}(R)$.

Because R is convex and contains the origin, $dC \cap R$ is simply cone dC with all points greater than some distance from the origin removed. So, by our observation above, $\mathbf{E}_{\text{rel}^{-1}(dC \cap R)}[||\text{irrel}(x)||^2] \geq \mathbf{E}_{\text{rel}^{-1}(dC)}[||\text{irrel}(x)||^2]$. By definition of $\mathbf{V}_{\text{rel}}(R)$, we know $\text{rel}^{-1}(dC \cap R) = d\tilde{C} \cap R$, so $\mathbf{E}_{(d\tilde{C} \cap R)}[||\text{irrel}(x)||^2] \geq \mathbf{E}_{d\tilde{C}}[||\text{irrel}(x)||^2]$. Since cone dC was arbitrary, this holds for all such infinitesimal cones and we can integrate the expectation over them to get $\mathbf{E}_R[||\text{irrel}(x)||^2] \geq \mathbf{E}_{B_n}[||\text{irrel}(x)||^2]$. $\blacksquare$

The next (and main) lemma shows that as long as there is a reasonable fraction of both positive and negative examples and the positive region contains a small ball about the origin, the average squared length of a positive example is less by some noticeable amount than the average squared length of a point in B_n (and therefore is less by some noticeable amount that the average squared length of a negative example).

Lemma 7 *Let $R \subseteq B_n$ be a convex set and suppose that the origin is τ-central in R. Let $r = \mathrm{vol}(R)/\mathrm{vol}(B_n)$. Then,*

$$\mathbf{E}_R\left[\,||x||^2\,\right] \;\leq\; \mathbf{E}_{B_n}\left[\,||x||^2\,\right](1-\Delta)$$

where $\Delta = r(\frac{\tau}{8n^2})$ if $r \leq \frac{1}{2}$, and $\Delta = \frac{(1-r)^2}{r}(\frac{\tau}{8n^2})$ if $r > \frac{1}{2}$.

Proof (of Lemma 7). The basic idea of the proof is as follows. Imagine breaking up B_n into a union of infinitesimally small cones. Each cone has the same value of $\mathbf{E}[||x||^2]$. For a cone dC, say that the *length* of the intersection $dC \cap R$ is the maximum $||x||$ over $x \in dC \cap R$. What we show is that a noticeable volume of R is contained in cones dC such that $dC \cap R$ has length at most $1-\alpha$ for some α noticeably greater than zero. This in turn gives us the result we want. Now, to the specifics.

Let α be some small quantity greater than 0 to be determined later. We begin by showing that a substantial part of the surface of R lies inside $(1-\alpha)B_n$. To do so, we first provide bounds on the volume of $R \cap (1-\alpha)B_n$. We know $\mathrm{vol}(R) = r\gamma_n$. Also note that $\mathrm{vol}(B_n - (1-\alpha)B_n) = (1-(1-\alpha)^n)\gamma_n$. So,

$$\begin{aligned}
\mathrm{vol}(R \cap (1-\alpha)B_n) &\geq r\gamma_n - (1-(1-\alpha)^n)\gamma_n \\
&= \gamma_n(r + (1-\alpha)^n - 1) \\
&\geq \gamma_n(r - \alpha n). \qquad (1)
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{vol}((1-\alpha)B_n - R) &\geq (1-\alpha)^n\gamma_n - r\gamma_n \\
&\geq \gamma_n(1 - \alpha n - r). \qquad (2)
\end{aligned}$$

We now use the following isoperimetric inequality due to Lovász and Simonovits [8].

Fact 8 (Theorem 2.1 of [8]) *Let T be a convex set in $\mathbf{R}^n$ partitioned into two sets S and $T-S$ by an $(n-1)$-dimensional surface with surface area ($(n-1)$-dimensional volume) a. Then,*

$$a \;\geq\; \frac{1}{diam(T)}\min[\mathrm{vol}(S),\ \mathrm{vol}(T-S)]$$

where $diam(T)$ is the maximum distance between two points in T.

We apply this inequality with $T = (1-\alpha)B_n$, $S = R \cap (1-\alpha)B_n$, and the separating surface being $\partial R \cap T$ (the surface of R inside T), to get (using equations (1) and (2)):

$$\begin{aligned}
&\mathrm{vol}_{n-1}(\partial R \cap (1-\alpha)B_n) \;\geq \\
&\quad \tfrac{1}{2}\gamma_n \min\left[\, r - \alpha n,\ 1 - \alpha n - r \,\right]. \qquad (3)
\end{aligned}$$

We also need the following.

Claim 1 *The integrated volume over all infinitesimal cones of R that are contained in $(1-\alpha)B_n$ is at least $\frac{\tau}{n}\mathrm{vol}_{n-1}(\partial R \cap (1-\alpha)B_n)$.*

That is, the volume of the set of points $x \in R$ such that the ray from the origin through x hits the surface of R while still inside $(1-\alpha)B_n$ is at least the above quantity.

Proof of claim. Let ρ be an infinitesimal patch of ∂R inside $(1-\alpha)B_n$, with surface area dS. Let x be a point on ρ and let θ be the angle between the tangent hyperplane H to ∂R and the hyperplane normal to x (treating x as a vector). Then, $dS\cos\theta$ is the surface area of ρ projected onto the hyperplane normal to x. So, the volume of the cone from the origin to ρ is $(dS\cos\theta)\frac{||x||}{n}$.

We were given (in the statement of the lemma) that R contains a ball of radius τ about the origin. Also, R is convex which implies that all points on the far side of hyperplane H from the origin are not in R. So, $||x||\cos\theta$, which is the closest distance of H to the origin, is at least τ. Now combining this with the result of the last paragraph we get that the volume of the cone from the origin to ρ is at least $dS\frac{\tau}{n}$. Integrating over all infinitesimal patches ρ yields the claim. ■

To finish the proof of lemma 7, we consider two cases.

Case 1: $r \leq 1/2$, so the "min" in inequality (3) is $r - \alpha n$.

For this case, we choose $\alpha = \frac{r}{2n}$. So, $\mathrm{vol}_{n-1}(\partial R \cap (1-\alpha)B_n) \geq \frac{1}{2}\gamma_n(r/2) = r\gamma_n/4$. Applying Claim 1 we have that the volume of cones of R ending within $(1-\alpha)B_n$ is at least $(\frac{\tau}{4n})r\gamma_n$. In other words, at least a $\frac{\tau}{4n}$ fraction of R is inside such cones.

Now, each cone C that ends within $(1-\alpha)B_n$ has length at most $(1-\alpha)$ which means that $\mathbf{E}_C[\,||x||^2\,] \leq (1-\alpha)^2\mathbf{E}_{B_n}[||x||^2]$. Each cone C that doesn't end within $(1-\alpha)B_n$ has $\mathbf{E}_C[\,||x||^2\,] \leq \mathbf{E}_{B_n}[||x||^2]$. So, integrating over all cones, we have:

$$\begin{aligned}
\mathbf{E}_R[||x||^2] &\leq \frac{\tau}{4n}(1-\alpha)^2\mathbf{E}_{B_n}[||x||^2] + (1-\tfrac{\tau}{4n})\,\mathbf{E}_{B_n}[||x||^2] \\
&\leq \mathbf{E}_{B_n}[||x||^2]\left(1 - \tfrac{\tau\alpha}{2n} + \tfrac{\tau\alpha^2}{4n}\right) \\
&\leq \mathbf{E}_{B_n}[||x||^2]\,(1 - \tfrac{\tau\alpha}{4n}) \\
&\leq \mathbf{E}_{B_n}[||x||^2]\left(1 - \tfrac{\tau r}{8n^2}\right) \square
\end{aligned}$$

Case 2: $r \geq 1/2$. Here we set $\alpha = \frac{1-r}{2n}$. The proof is similar to Case 1 and proceeds as follows.

Using equation (3) we have $\mathrm{vol}_{n-1}(\partial R \cap (1-\alpha)B_n) \geq (1-r)\gamma_n/4$. Applying Claim 1 we get that at least a fraction $f = \frac{(1-r)}{r}(\frac{\tau}{4n})$ of R is inside cones of R ending within $(1-\alpha)B_n$. So, we get:

$$\begin{aligned}
\mathbf{E}_R[||x||^2] &\leq f(1-\alpha)^2\mathbf{E}_{B_n}[||x||^2] + (1-f)\mathbf{E}_{B_n}[||x||^2] \\
&\leq \mathbf{E}_{B_n}[||x||^2]\left(1 - 2f\alpha + f\alpha^2\right) \\
&\leq \mathbf{E}_{B_n}[||x||^2]\,(1 - f\alpha) \\
&\leq \mathbf{E}_{B_n}[||x||^2]\left(1 - \tfrac{(1-r)^2\tau}{8n^2 r}\right) \square
\end{aligned}$$

So we have proved Lemma 7. ■

Proof of Theorem 4. First, lemmas 5 and 6 imply that $\mathbf{E}_R[(w\cdot x)^2] \geq \frac{1}{n+2}$ for the following reason. By symmetry (all directions $w \in \mathbf{V}_{\mathrm{irrel}}(R)$ have the same value of $\mathbf{E}_R[(w\cdot x)^2]$) and the fact that $\mathbf{V}_{\mathrm{irrel}}(R)$ is an $(n-k)$-dimensional space, we have $\mathbf{E}_R[(w\cdot x)^2] = \frac{1}{n-k}\mathbf{E}_R[||\mathrm{irrel}(x)||^2]$. This

quantity by Lemma 6 is at least $\frac{1}{n-k}\mathbf{E}_{B_n}[||\text{irrel}(x)||^2]$, which (by symmetry and lemma 5) equals $\frac{1}{n+2}$.

Now, using $||x||^2 = ||\text{rel}(x)||^2 + ||\text{irrel}(x)||^2$, we can apply lemmas 5 and 7 to get $\mathbf{E}_R[||\text{rel}(x)||^2] + \mathbf{E}_R[||\text{irrel}(x)||^2] \leq \mathbf{E}_{B_n}[||\text{rel}(x)||^2] + \mathbf{E}_{B_n}[||\text{irrel}(x)||^2] - \Delta\frac{n}{n+2}$. This implies

$$\begin{aligned}\mathbf{E}_R[||\text{rel}(x)||^2] &\leq \mathbf{E}_{B_n}[||\text{rel}(x)||^2] - \tfrac{n\Delta}{n+2} \qquad \text{(Lemma 6)}\\ &= \tfrac{k}{n+2} - \tfrac{n\Delta}{n+2}.\\ &\quad \text{(by observation in Lemma 5)}\end{aligned}$$

Now, by definition of v (in particular the fact that $\mathbf{E}_R[(v\cdot x)^2] \leq \frac{1}{k}\mathbf{E}_R[||\text{rel}(x)||^2]$), we have $\mathbf{E}_R[(v\cdot x)^2] \leq \frac{1}{n+2} - \frac{n\Delta}{k(n+2)}$. The theorem follows. ■

4 The algorithm

This section describes the learning algorithm. We assume the target concept P is an intersection of ℓ halfspaces in n dimensions, such that $\mathbf{V}_{\text{rel}}(P)$ has dimension k. We want to learn with error ϵ and failure probability at most δ.

With section 3 on hand, the recursive algorithm is as follows: We first use the procedure of section 3 to find, with probability at least $1-\delta/4$, a unit length vector u whose irrelevant component is at most ϵ_1 (ϵ_1 will be specified later). The algorithm then considers slices perpendicular to u. Each slice is of the form

$$\mathsf{slice}(m) \quad = \quad \{x \in B_n : m\epsilon_1 \leq u\cdot x \leq (m+1)\epsilon_1\}$$

where m is an integer and ϵ_1 is the thickness of each slice. So, there are $2/\epsilon_1$ slices total. Call a slice "big" if its volume is at least $\frac{1}{12}\epsilon_1\epsilon\gamma_n$. We will recursively PAC-learn each big slice to error at most $\epsilon/6$ and probability of failure at most $\frac{\delta\epsilon}{48n}$. (We later show that this is a "$k-1$" problem.) We just classify each small slice as negative. The final hypothesis produced will in essence be a depth-k "linear-threshold decision tree", where a decision feeds an example into the hypothesis for the appropriate slice. Before giving a more detailed description of the algorithm, we prove some technical facts needed for the analysis of the algorithm.

Suppose the (unknown) positive region is

$$P \quad = \quad \{x \in B_n : a_i\cdot x \leq a_{io} \text{ for } i = 1,2,\ldots,\ell\}$$

where the a_i's are unit length vectors that span the k dimensional relevant space $\mathbf{V}_{\text{rel}}(P)$. Let u_1 be the relevant component of u. Each $a_i = \lambda_i u_1 + b_i$ for some real number λ_i and vector $b_i \in \mathbf{V}_{\text{rel}}(P)$ orthogonal to u_1. Let $c_i = \lambda_i u + b_i$ and let

$$P_1 = \{x \in B_n : c_i\cdot x \leq a_{io} \text{ for } i = 1,2,\ldots,\ell\}.$$

For each integer m, let $P_1(m)$ be the intersection of P_1 with the "outer" bounding hyperpplane of $\mathsf{slice}(m)$. Specifically, for $m \geq 0$, $P_1(m)$ is the intersection of P_1 with the hyperplane $\{x : u\cdot x = (m+1)\epsilon_1\}$, and for $m \leq -1$, $P_1(m)$ is the intersection of P_1 with $\{x : u\cdot x = m\epsilon_1\}$. Note that $P_1(m)$

is an $n-1$ dimensional convex set with a "relevant space" of dimension $k-1$, since $b_1, b_2, \ldots, b_l$ span the $k-1$ dimensional space of vectors in $\mathbf{V}_{\rm rel}(P)$ orthogonal to u_1. We want to assume that all examples that land in $\mathsf{slice}(m)$ are labeled according to $P_1(m)$ so we can recurse on a relevant space of one less dimension, but this will introduce some errors which need to be analyzed; we first prove some facts which will be useful for this analysis.

For $x \in B_n$, define $f(x)$ to be the projection of x onto the smaller bounding plane of its slice. That is, if $x = \lambda u + y$, with y perpendicular to u, then $f(x) = \lceil \frac{\lambda}{\epsilon_1} \rceil \epsilon_1 u + y$ if λ is nonnegative, and $f(x) = \lfloor \frac{\lambda}{\epsilon_1} \rfloor \epsilon_1 u + y$ if λ is negative. We call an $x \in B_n$ "good" (or say $x \in$GOOD) if it satisfies *all* of the following:

(i) x belongs to both P and P_1 or x belongs to neither.

(ii) $f(x)$ belongs to both P and P_1 or it belongs to neither.

(iii) $f(x)$ belongs to B_n.

(iv) Either both x and $f(x)$ belong to P or neither does.

In particular, for an example x from GOOD $\cap$ $\mathsf{slice}(m)$, its label according to P is the same as the label of $f(x)$ according to $P_1(m)$.

The algorithm will take each labeled example x in a big slice $\mathsf{slice}(m)$, attach the same label to $f(x)$ (if $f(x)$ is in B_n) and treat these as labeled examples for $P_1(m)$. Say that $x \in B_n$ is "bad" (or $x \in$BAD) if $x \notin$GOOD. We then have the following.

Lemma 9 *The volume of the bad set is at most* $12\ell\gamma_{n-1}\epsilon_1 + \gamma_n n\epsilon_1 \leq 13\ell n\gamma_n\epsilon_1$.

Proof: We first observe that $||a_i - c_i|| \leq 2\epsilon_1$ since we may assume $|\lambda_i| < 2$. So for $x \in B_n$, we have $a_i x - 2\epsilon_1 \leq c_i x \leq a_i x + 2\epsilon_1$ and so the set of x that violates (i) is contained in $\cup_{i=1}^{l}\{x \in B_n : a_{io} - 2\epsilon_1 \leq a_i x \leq a_{io} + 2\epsilon_1\}$; this has volume at most $4\ell\epsilon_1\gamma_{n-1}$, since $||a_i|| = 1$ for all i. If x violates (ii), then $f(x)$ belongs to $\cup_{i=1}^{\ell}\{x \in B_n : a_{io} - 2\epsilon_1 \leq a_i x \leq a_{io} + 2\epsilon_1\}$ by the above. But $||x - f(x)|| \leq \epsilon_1$, so we have that x belongs to the set $\cup_{i=1}^{\ell}\{x \in B_n : a_{io} - 2\epsilon_1 - \epsilon_1 \leq a_i x \leq a_{io} + 2\epsilon_1 + \epsilon_1\}$ whose volume is at most $6\ell\epsilon_1\gamma_{n-1}$. If $f(x)$ does not belong to B_n, then x belongs to $\{x : 1 - \epsilon_1 \leq ||x|| \leq 1\}$, a set of volume $\leq \gamma_n n\epsilon_1$. If x violates (iv), then note that x belongs to the set $\cup_{i=1}^{\ell}\{x \in B_n : a_{io} - \epsilon_1 \leq a_i x \leq a_{io} + \epsilon_1\}$ whose volume is at most $2\ell\epsilon_1\gamma_{n-1}$. ■

We will define a slice to be "bad" if the volume of the bad set intersected with the slice is at least

$$\epsilon_3 \quad \times \text{ (the volume of the slice)}$$

where ϵ_3 will be specified later. Otherwise, the slice is "good."

We use $E(k, n, \epsilon, \delta)$ to denote the number of examples the algorithm will need, where k, n, ϵ, and δ stand for the usual quantities. Let us assume by induction that we have already computed $E(k-1, n-1, \epsilon/6, \frac{\delta\epsilon}{48n})$, and abbreviate this by $N(k-1)$. We are now ready to describe the

algorithm.

THE ALGORITHM

1. Find with probability at least $1 - \delta/4$, a vector u with irrelevant component at most ϵ_1 as described above.[1]

2. Declare all small slices to be all negative.

3. Pick $\frac{24}{\epsilon_1 \epsilon}\left[N(k-1) + 4\log(\frac{8}{\delta\epsilon_1})\right]$ labeled examples from B_n.

4. If any big slice has less than $N(k-1)$ examples in it, halt and declare failure.

5. Otherwise, take the first $N(k-1)$ examples that land in each big slice, attach the same label to $f(x)$ (as x) and recursively learn the $n-1$ dimensional problem with $k-1$ relevant directions with error at most $\epsilon/6$ and failure probability at most $\delta\epsilon/(48n)$.

   The overall concept learned is: if x lies in a big slice, then it is positive iff $f(x)$ is positive according to the recursively learned concept.

Lemma 10 *In the above algorithm, the probability we halt in step (4) is at most $\delta/4$.*

Proof: In a given big slice, the expected number of examples we get is at least: $2[N(k-1) + 4\log(\frac{8}{\epsilon_1\delta})]$. The probability that we get less that half that many is at most (by Chernoff bounds) $e^{-2[N(k-1)+4\log(8/\epsilon_1\delta)]/8} \leq \epsilon_1\delta/8$. So, the chance there is *any* such slice is at most $\delta/4$. ■

We are now ready to specify ϵ_3. We choose: $\epsilon_3 = \left(\frac{\epsilon}{24n}\right)\frac{1}{N(k-1)}$.

We call a good slice "corrupt" if at least one (among the first $N(k-1)$) example in the slice lands in the bad region. [Note that we do not know which slices are corrupt. So corruptness is used only in the analysis, not by the algorithm.]

The expected number of examples that land in the bad region of a good slice is at most ϵ_3 times the number of examples in the slice, so is at most $\frac{\epsilon}{24n}$; thus by Markov inequality, the probability that some example lands in the bad region is at most $\frac{\epsilon}{24n}$. Now noting that different good slices being corrupted are independent events, we apply Chernoff bounds to get that the probability there are more than $(\frac{\epsilon}{6n\epsilon_1})$ corrupt slices is at most $e^{-\epsilon/(36n\epsilon_1)}$ which is at most $\delta/4$ for sufficiently small ϵ_1. Since the volume of any slice is at most $\gamma_{n-1}\epsilon_1 \leq n\gamma_n\epsilon_1$, we have:

Lemma 11 *The probability that the total volume of corrupt slices exceeds $\gamma_n\epsilon/6$ is at most $\delta/4$.*

In the uncorrupted good slices, there is still a probability $\frac{\delta\epsilon}{48n}$ that the recursively-called algorithm will fail. So, the expected number of such slices (which we will call "failed slices") is at most $\frac{\delta\epsilon}{24\epsilon_1 n}$. By Markov's inequality, the probability there are more than $\frac{\epsilon}{6\epsilon_1 n}$ failed slices is at most $\delta/4$. Again, since the volume of a slice is at most $n\gamma_n\epsilon_1$ we have:

[1] As mentioned in section 3, we first sample to check whether "all positive" or "all negative" are sufficiently close hypotheses before running the procedure. $\delta/4$ is the combined failure probability.

Lemma 12 *The probability that the total volume of failed slices exceeds $\gamma_n \epsilon/6$ is at most $\delta/4$.*

We have not specified the choice of ϵ_1 yet; now we do so. It is chosen so that $13\ell n\epsilon_1/\epsilon_3 \leq \epsilon/6$.

Confidence analysis: We want that the first vector u has irrelevant component at most ϵ_1, that all big slices get sufficient examples, and that lemmas 11 and 12 above hold. Each fails with probability at most $\delta/4$ so our total failure probability is at most δ.

Error analysis: In the worst case, our hypothesis is completely erroneous in the small, bad, corrupt, and failed slices. In addition, there are errors in the remaining slices from the recursive calls. The total error volume is therefore at most as follows:

- Small Slices : Total Volume $\leq \epsilon\gamma_n/6$.
- Bad Slices : Total volume $\leq 13\ell n\gamma_n\epsilon_1/\epsilon_3 \ \leq \epsilon\gamma_n/6$.
- Corrupt Slices : Total Volume $\leq \epsilon\gamma_n/6$.
- Failed Slices : Total volume $\leq \epsilon\gamma_n/6$.
- Remaining slices : Error $\leq \epsilon/6$ of volume. So total error volume is at most $\epsilon\gamma_n/6$.

Thus, the total error of the hypothesis produced is at most $5\epsilon/6 < \epsilon$.

Number of examples needed: To find the first vector u to error less than ϵ_1 requires $O(\epsilon_1^{-4} \cdot poly(n, \ell, \frac{1}{\epsilon}, \log\frac{1}{\delta}))$ examples. This quantity is $O(N(k-1)^4 \cdot poly(n, \ell, \frac{1}{\epsilon}, \log\frac{1}{\delta}))$. The additional number of examples needed in step (2) of the algorithm is $O(N(k-1)^2 \cdot poly(n, \ell, \frac{1}{\epsilon}, \log\frac{1}{\delta}))$. Thus, it is clear that since k is a constant, this will be polynomial in the desired parameters, although it is doubly-exponential in k.

References

[1] E. B. Baum. Polynomial time algorithms for learning neural nets. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 258–272. Morgan Kaufmann, 1990.

[2] E. B. Baum. On learning a union of half spaces. *Journal of Complexity*, 6(1):67–101, March 1990.

[3] M. Berger. *Geometry*. Springer-Verlag, 1987.

[4] A. Blum and R. Rivest. Training a 3-node neural network is NP-Complete. *Neural Networks*, 5:117–127, 1992.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[6] M. Kearns, M. Li, L. Pitt, and L. G. Valiant. On the learnability of Boolean formulae. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pages 285–295, May 1987.

[7] P. M. Long and M. K. Warmuth. Composite geometric concepts and polynomial predictability. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 273–287. Morgan Kaufmann, 1990.

[8] L. Lovász and M. Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, volume I, pages 346–354, St. Louis, October 1990.

[9] N. Megiddo. On the complexity of polyhedral separability. Technical Report RJ 5252, IBM Almaden Research Center, August 1986.

[10] D. F. Morrison. *Multivariate Statistical Methods*, 3rd Ed. McGraw-Hill, 1990.

[11] L. G. Valiant. A theory of the learnable. *CACM* 27(11):1134-1142, 1984.

Appendix

Proof of Lemma 1. Let y be the closest point to μ_R on the boundary of R. Let $\tau = ||y - \mu_R||$. Since R is convex and y is the closest point, the there are no points in R on the opposite side of the hyperplane H through y perpendicular to the vector $y - \mu_R$. It is a fact that in general if a point z is not in a convex region $R \subseteq \mathbf{R}^n$, then neither is the point $\mu_R + n(\mu_R - z)$ [3]. Thus, all points in R lie between the hyperplane H and the hyperplane parallel to H through $\mu_R + n(\mu_R - y)$. These hyperplanes are separated by distance $(n+1)\tau$, so the volume of R is at most $(n+1)\tau\gamma_{n-1} \leq (n+1)\tau(2\gamma_n\sqrt{n})$. Setting this to $r\gamma_n$ yields the desired inequality for τ. ■