Learning Convex Concepts from Gaussian Distributions with PCA

Santosh S. Vempala
School of Computer Science
Georgia Tech
Atlanta, USA
vempala@gatech.edu

***Abstract*—We present a new algorithm for learning a convex set in n-dimensional space given labeled examples drawn from any Gaussian distribution. The complexity of the algorithm is bounded by a fixed polynomial in n times a function of k and ϵ where k is the dimension of the *normal subspace* (the span of normal vectors to supporting hyperplanes of the convex set) and the output is a hypothesis that correctly classifies at least $1-\epsilon$ of the unknown Gaussian distribution. For the important case when the convex set is the intersection of k halfspaces, the complexity is**

$$\textbf{poly}(n, k, 1/\epsilon) + n \cdot \min\, k^{O(\log k/\epsilon^4)}, (k/\epsilon)^{O(k)},$$

improving substantially on the state of the art [Vem04], [KOS08] for Gaussian distributions. The key step of the algorithm is a Singular Value Decomposition after applying a normalization. The proof is based on a monotonicity property of Gaussian space under convex restrictions.

***Index Terms*—High-dimensional learning; convex; PCA; Gaussians; polynomial time;**

I. Introduction

Let a set of points in $\mathbb{R}^n$ be drawn from an unknown distribution with each point labeled *positive* if it belongs to an unknown set K and *negative* otherwise. In this paper, we are interested in the complexity of learning K, i.e., finding a hypothesis that correctly labels almost all of the input distribution, in the case when K is convex and the input distribution is an unknown Gaussian. Nearly matching upper and lower bounds of roughly $2^{\tilde{O}(\sqrt{n})}$ were established for general K in recent work [KOS08] (see also [GR09]). These lower bounds do not apply to many interesting classes of convex sets, e.g., intersections of a polynomial number of halfspaces. It is well-known that a single halfspace can be PAC-learned in polynomial time, and that an intersection of a constant number of halfspaces can be learned from restricted distributions (uniform over a ball, noncencentrated over a ball, Gaussian and logconcave) in polynomial time [Bau90a], [Bau90b], [BK93], [Vem97], [Vem04], [KOS08], [KLT09], [Vem10]. We discuss the history of this problem in detail in Section II. We note that without any assumptions on the input distribution, the complexity of learning an intersection of *two* halfspaces is a major open question.

Here we consider a general setting that captures the intersection of $k < n$ halfspaces, namely we assume that the normal vectors to supporting hyperplanes of the unknown convex set K lie in an (unknown) k-dimensional subspace for some $k < n$. One can view this as a convex set in $\mathbb{R}^k$ lifted orthogonally. To check whether a point x lies in K, one only has check whether the projection of x to the normal subspace lies in K. More precisely, a convex set K in $\mathbb{R}^n$ has a *normal subspace* V defined as the span of all normals to supporting planes of K. It follows that

$$K = \{x \in \mathbb{R}^n \,:\, \pi_V(x) \in K \cap V\}$$

where π_V is the projection to V. For example, if K is defined by an intersection of k halfspaces $a_1 \cdot x \geq b_1, \ldots, a_k \cdot x \geq b_k$, then V is the span of the normals $a_1, \ldots, a_k$. If K is a convex cone, then V is the span of its dual cone. If K is a full-dimensional convex body, then $V = \mathbb{R}^n$.

Our main contribution in this paper is a simple algorithm to approximately recover the normal subspace: apply an affine transformation (that ignores labels), then apply Principal Component Analysis (PCA) to the subset of examples labeled positive and declare the relevant subspace to be the span of the smallest k principal components. Once the normal subspace has been identified, we project examples to it and the complexity of learning the unknown convex set becomes a function of k and not n.

A. Results

We assume that the distribution on examples, $F = (\mu, \Sigma)$, is a Gaussian distribution in $\mathbb{R}^n$. Points are labeled according to a convex set K whose normal subspace has dimension k with the labeling function $\ell : \mathbb{R}^n \to \{-1, 1\}$ defined as:

$$\ell(x) = \begin{cases} 1 \text{ if } x \in K \\ -1 \text{ if } x \notin K. \end{cases}$$

In addition to labeled examples from an unknown Gaussian, the algorithm is given two error parameters $\epsilon, \delta > 0$. The algorithm is required to succeed with probability at least $1-\delta$ in finding a hypothesis that correctly classifies at least $1-\epsilon$ of the unknown input distribution. All our complexity bounds will depend on δ by a factor $\ln(1/\delta)$, so we do not mention this term in the theorems that follow.

Theorem 1.1: Suppose points in $\mathbb{R}^n$ are drawn from an unknown Gaussian distribution and labeled according to an unknown convex set K with a normal subspace of dimension

k. Then, there is an absolute constant C such that for any $\epsilon > 0$, this concept can be learned to accuracy $1 - \epsilon$ using

$$\frac{C}{\epsilon^6} \cdot n \cdot e^{2k}(k + \ln(1/\epsilon)) \ln^2 n + k^{\tilde{O}(\sqrt{k}/\epsilon^4)}$$

examples and

$$\frac{C}{\epsilon^6} \cdot n \cdot e^{2k}(k + \ln(1/\epsilon)) \ln^2 n + nk^{\tilde{O}(\sqrt{k}/\epsilon^4)}$$

time.

The complexity bounds are considerably smaller when the unknown hypothesis is an intersection of k halfspaces.

Theorem 1.2: Suppose points in $\mathbb{R}^n$ are drawn from an unknown Gaussian distribution in $\mathbb{R}^n$ and labeled according to an intersection of k halfspaces. Then there is an absolute constant C such that for any $\epsilon > 0$, this concept can be learned to accuracy $1 - \epsilon$ using

$$\frac{C}{\epsilon^6} nk^6 \ln^2 n \ln(k/\epsilon) + \min\, k^{O(\ln k/\epsilon^4)}, \left(\frac{k}{\epsilon}\right)^{O(k)}$$

examples and

$$\frac{C}{\epsilon^6} n^3 k^6 \ln^2 n \ln(k/\epsilon) + n \cdot \min\, k^{O(\ln k/\epsilon^4)}, \left(\frac{k}{\epsilon}\right)^{O(k)}$$

time.

The main new theorem underlying both of these is the following guarantee on identifying the normal subspace.

Theorem 1.3: Given examples in $\mathbb{R}^n$ drawn from an unknown Gaussian and labeled according to a convex set K whose normal subspace has dimension k, and any $\epsilon, \delta > 0$, Algorithm Spectral-Subspace (Section III) outputs a subspace V of dimension at most k such that with probability at least $1 - \delta$, the hypothesis

$$\ell(x) = 1 \text{ iff } \pi_V(x) \in K \cap V$$

correctly classifies at least $1 - \epsilon$ of the distribution F. The number of labeled examples required is bounded by

$$S = \frac{C}{\epsilon^6} \cdot ne^{2k}(k + \ln(1/\epsilon)) \ln^2 n \ln(1/\delta)$$

where C is an absolute constant and the time complexity is $O(Sn^2)$. When K is the intersection of k halfspaces, the sample complexity bound improves to

$$s = C \cdot \frac{k^6}{\epsilon^6} \cdot n \ln^2 n \ln(1/\delta) \ln(k/\epsilon)$$

and the time complexity is $O(sn^2)$.

Our analysis relies on a monotonicity property of Gaussian space under convex restrictions (Lemma 4.8): restricting a Gaussian to any convex set decreases the variance in any direction along which the convex set imposes a nontrivial restriction. This property is perhaps a bit surprising since it does not hold for other logconcave distributions, even the uniform distribution over a unit ball — restricting a unit ball in $\mathbb{R}^n$ to a narrow one-dimensional cylinder can make the variance in the direction of the axis of the cylinder increase from roughly $1/\sqrt{n}$ to 1. This monotonicity lemma settles an open question posed by O'Donnell [Goy05].

II. Related work

Learning convex sets is a fundamental topic in algorithms and has been studied from many angles, including its sample complexity and computational complexity. The simplest convex concept, a single halfspace, can be PAC-learned in polynomial time via efficient linear programming algorithms.

The complexity of the next natural question, PAC-learning an intersection of two halfspaces is open. Much progress has been made on learning an intersection of two or more halfspaces under restricted distributions and on learning using restricted hypothesis classes, e.g., polynomial threshold functions [She10].

In 1990, Baum [Bau90b] 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, i.e., for any $x \in R^n$, the density/probability at x is the same as at $-x$. Baum's algorithm was recently shown to work for logconcave distributions [KLT09]. A few years after Baum's work, Blum and Kannan [BK93], [BK97] found a polynomial-time algorithm that works for a constant number of halfspaces for the uniform distribution on the unit ball. 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 1997, we presented an algorithm [Vem97], [Vem04] whose running time and sample complexity were $n^k(k/\epsilon)^{O(k)}$, i.e., singly exponential in k[1]. The algorithm was shown to work for near-uniform distributions on the unit ball with the property that the density does not vary by more than a polynomial factor. Moreover, it explicitly finds an intersection of $O(k \log(1/\epsilon))$ halfspaces. Recently, this algorithm and its analysis were extended to any logconcave distribution in $\mathbb{R}^n$ with the same complexity bounds [Vem10].

Klivans, O'Donnell and Servedio [KOS08] gave an algorithm based on approximating an intersection of k halfspaces with a low-degree polynomial threshold function. Their approach has time and sample complexity $n^{O(\log k/\epsilon^4)}$, works for Gaussian input distributions, and outputs a hypothesis that is a polynomial threshold function of degree $O(\log k/\epsilon^4)$. They also showed that for more general convex sets, the complexity of learning from a Gaussian distribution is $2^{\tilde{O}(\sqrt{n}/\epsilon^4)}$ and gave a nearly matching lower bound for the sample complexity [KOS08]. A similar lower bound was also given in [GR09].

III. Algorithm

The number of samples needed by the algorithm below, m, depends on hypothesis class and is given in the theorems providing guarantees for the algorithm. Steps 2 and 3 below are usually referred to together as *PCA*, i.e., computing the Singular Value Decomposition (SVD) after shifting the mean to be the origin.

[1]The conference version [Vem97] claimed a fixed polynomial dependence on n and this was corrected in [Vem04].

Spectral-Subspace

Input: A matrix A whose m rows are points in $\mathbb{R}^n$, their labels, and an integer k.

Output: A subspace V of dimension k.

1) (*Make isotropic*) Apply an affine transformation T so that the resulting set of points $B = AT$ is isotropic, i.e., the rows sum to zero and $(1/m)B^TB = I$.

2) (*Center positive examples*) Let B^+ be the subset of points labeled positive with mean μ^+. Center B^+ at the origin to get C:
$$C = B^+ - \mathbf{1}(\mu^+)^T.$$

3) (*Compute SVD*) Let $v_1, \ldots, v_k$, be the smallest right singular vectors of C. The output subspace is
$$V = T\text{span}\{v_1, \ldots, v_k\}.$$

Once we have an approximation to the relevant subspace, we project samples to it, then learn the concept in the lower-dimensional subspace. One simple way to do this is the following procedure, used as the final step in the algorithm of [Vem97], [Vem10].

1) Cover the unit sphere in $\mathbb{R}^k$ with a set of vectors S such that every unit vector is within angle $\epsilon/4k$ from some vector in S.
2) Then greedily choose a subset of vectors that maximize the number of negatives separated from at least $(1-\epsilon/4)$ of the positives. The halfspaces corresponding to this subset should separate all but at most an $\epsilon/4$ fraction of the negatives.

In $\mathbb{R}^k$, there is a cover with the above property of size $(k/\epsilon)^{O(k)}$. The greedy algorithm is analyzed similar to set cover (see e.g., [Vem10]) and finds a set of at most $O(k\log(1/\epsilon))$ that correctly classify most of the sample. Thus the complexity post-projection is $(k/\epsilon)^{O(k)}$ and the number of samples required is poly$(k, 1/\epsilon)$. We state this guarantee formally in Section IV-A as Theorem 4.5.

The other alternative is to fit a polynomial threshold function as in [KOS08]. For an intersection of k halfspaces, the degree required is $O(\log k/\epsilon^4)$ and thus the complexity is $k^{O(\log k/\epsilon^4)}$.

IV. Analysis

We begin with an overview. The main idea of the algorithm is to identify the normal subspace using PCA. More precisely, if the full distribution is isotropic (the covariance matrix has all unit eigenvalues), we will show that the positive distribution has the same variance as the full distribution along any direction orthogonal to the normal subspace, but for directions in the normal subspace, the variance is smaller; thus finding the smallest k principal components and projecting to their span reduces the dimensionality of the learning problem from n to k.

Why does PCA capture the normal subspace as the span of its smallest principal components? To see this, we first observe that the distribution is a standard Gaussian after the isotropic transformation. It follows immediately from the symmetry of the Gaussian distribution that for directions orthogonal to the normal subspace N, the variance is the same for the positive distribution and the full distribution, since the projection to the normal subspace gives a standard Gaussian. Indeed, it is the same along every line orthogonal to N. The key observation driving the algorithm is that the variance along any direction of a standard Gaussian restricted to a convex body is less than 1, regardless of the shape of the body (Lemma 4.7). Once this is established, the remaining technical challenge is to quantify the decrease in variance and bound the number of samples required to estimate it to sufficient accuracy.

A. Preliminaries

The first lemma below is a well-known fact about Gaussian concentration.

Lemma 4.1: Let X be drawn from a standard Gaussian in $\mathbb{R}^n$. Then, for any $t > 0$,
$$\Pr(|\|X\|^2 - n\| \geq t\sqrt{n}) \leq 2e^{-t^2/8}.$$

The following bound on the number of samples to estimate the first and second moments of a logconcave distribution is based on a lemma from [Rud99], applied in [LV07].

Lemma 4.2: Let $X_1, \ldots, X_m$ be random samples from an isotropic logconcave distribution in $\mathbb{R}^n$ so that the sample mean is $\hat{\mu}$ and the sample covariance matrix is $\hat{\Sigma}$. For any $1 > \gamma, \delta > 0$, if
$$m > C \cdot \frac{1}{\gamma^2} \cdot n\log^2(n/\delta),$$
then with probability at least $1-\delta$, we have $\|\mu\| \leq \gamma$ and $\|\Sigma - I\|_2 \leq \gamma$.

The following lemma appears to be a folklore fact about one-dimensional logconcave functions. We only use it for the special case of a Gaussian density function restricted to an interval. Here and in the rest of the paper, for a one-dimensional random variable X with density f, we write $\mathbb{E}_f(X) = \mathbb{E}(f)$ and $\text{Var}_f(X) = \text{Var}(f)$.

Lemma 4.3: Let $f : \mathbb{R} \to \mathbb{R}_+$ be a logconcave density function. For $r \in \mathbb{R}$, let f_r be density obtained by restricting the support of f to $x \leq r$, i.e.,
$$f_r(x) = \begin{cases} \frac{f(x)}{\int_{-\infty}^{r} f(x)\,dx} \text{ if } x \leq r \\ 0 \text{ otherwise.} \end{cases}$$
Then, $\text{Var}(f_r) \leq \text{Var}(f)$.

Proof: It suffices to show that
$$\frac{d\text{Var}(f_r)}{dr} = ((r - \mathbb{E}(f_r))^2 - \text{Var}(f_r))f_r \geq 0.$$

For $r = 0$, this says that for a random variable X from a logconcave density with nonnegative support,

$$\mathbb{E}(X^2) \leq 2\mathbb{E}(X)^2,$$

which follows from Lemma 5.3(c) in [LV07]). ∎

We will use the following one-dimensional variant of the localization lemma of Lovász and Simonovits [LS93], [KLS95].

Lemma 4.4: Let f_1, f_2, f_3, f_4 be nonnegative continuous functions defined on an interval $[a, b]$ and let $\alpha, \beta > 0$. Then the following are equivalent:

1) (a) For every logconcave function F on $\mathbb{R}$,

$$\int_a^b F(t) f_1(t)\, dt = \int_a^b F(t) f_2(t)\, dt$$

and

$$\int_a^b F(t) f_3(t)\, dt \leq \int_a^b F(t) f_4(t)\, dt;$$

2) (b) For every subinterval $[a', b'] \subseteq [a, b]$ and every real γ,

$$\int_{a'}^{b'} e^{\gamma t} f_1(t)\, dt = \int_{a'}^{b'} e^{\gamma t} f_2(t)\, dt$$

and

$$\int_{a'}^{b'} e^{\gamma t} f_3(t)\, dt \leq \int_{a'}^{b'} e^{\gamma t} f_4(t)\, dt.$$

We conclude this section with known guarantees learning an intersection of halfspaces from Gaussians. The first is a simple greedy algorithm applied to a net of candidate halfspaces (described briefly in Section III).

Theorem 4.5: [Vem10] An intersection of k halfspaces in $\mathbb{R}^k$ can be PAC-learned in time $(k/\epsilon)^{O(k)}$ using $O((k^2/\epsilon))$ labeled examples from a Gaussian distribution, as a hypothesis that is an intersection of $O(k \log(1/\epsilon))$ halfspaces. This guarantee holds even if an arbitrary $\epsilon/2$ fraction of the examples are wrongly labeled.

The second algorithm uses polynomial threshold functions and has an agnostic learning guarantee, i.e., it finds a hypothesis of classification error at most ϵ more than the best possible error using an intersection of k halfspaces. There is no need to assume a bound on the classification error of the best intersection of halfspaces.

Theorem 4.6: [KOS08] An intersection of k halfspaces in $\mathbb{R}^k$ can be agnostically learned from a Gaussian input distribution with time and sample complexity $k^{O(\log k/\epsilon^4)}$ as a polynomial threshold function of degree $O(\log k/\epsilon^4)$. Any convex set in $\mathbb{R}^k$ can be agnostically learned from a Gaussian input distribution with time and sample complexity $k^{O(\sqrt{k}/\epsilon^4)}$, as a polynomial threshold function of degree $O(\sqrt{k}/\epsilon^4)$.

B. A monotonicity property of Gaussians

The next lemma is a monotonicity property of Gaussians and plays an important role in the proof.

Lemma 4.7: Let $f : \mathbb{R} \to \mathbb{R}_+$ be a logconcave function such that

$$\int_{\mathbb{R}} x e^{-(x-\mu)^2/2} f(x)\, dx = 0.$$

Then,

$$\frac{\int_{\mathbb{R}} x^2 e^{-(x-\mu)^2/2} f(x)\, dx}{\int_{\mathbb{R}} e^{-(x-\mu)^2/2} f(x)\, dx} \leq 1$$

with equality holding only if f is constant everywhere in $\mathbb{R}$. Moreover, if the support of f is $[a, b]$ with $|a| > b > 0$, then

$$\frac{\int_a^b x^2 e^{-(x-\mu)^2/2} f(x)\, dx}{\int_a^b e^{-(x-\mu)^2/2} f(x)\, dx} < 1 - \frac{1}{2\pi} e^{-b^2}.$$

We note that this lemma is closely related to the Brascamp-Lieb inequality [BL76]. One version of this inequality (see e.g., [BL00]) says

$$\mathrm{Var}_\mu(f) \leq \int_{\mathbb{R}^n} \|\nabla f\|^2\, d\mu$$

where μ is the standard Gaussian distribution in $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function (locally Lipschitz). Here if we take f to be the inner product with a fixed unit vector $u \in \mathbb{R}^n$, we get the variance along u on the LHS and 1 on the RHS. However, we are unable to use this inequality directly for two reasons. First, convex restrictions are not smooth functions, and second, we need quantitative bounds on how much smaller the LHS is compared to 1.

Proof: Using Lemma 4.4, it suffices to prove the conclusion for $f(x) = ce^{\gamma x}$ for any $\gamma \in \mathbb{R}, c > 0$. Then the equation can be rewritten as:

$$\begin{aligned}
&\int_a^b x e^{-(x-\mu)^2/2} f(x)\, dx \\
&= c \int_a^b x e^{-\frac{(x-\mu)^2}{2} + \gamma x}\, dx \\
&= c e^{\mu\gamma + \gamma^2/2} \int_a^b x e^{-(x-\mu-\gamma)^2/2}\, dx = 0
\end{aligned}$$

and the LHS of the inequality is

$$\frac{\int_a^b x^2 e^{-(x-\mu-\gamma)^2/2}\, dx}{\int_a^b e^{-(x-\mu-\gamma)^2/2}\, dx}.$$

In words, we have a Gaussian density restricted to $[a, b]$ so that the mean is zero and we wish to bound its second moment, i.e., we wish to bound its variance. Since $|a| \geq b$, it follows that $\gamma + \mu \geq 0$. Define $g_{a,b}$ to be the density proportional to $e^{-(x-\mu-\gamma)^2}$ restricted to $[a, b]$. (Recall that the restriction of a logconcave function to any interval is logconcave). We now apply Lemma 4.3 twice, first to observe that

$$\mathrm{Var}(g_{a,b}) \leq \mathrm{Var}(g_{-\infty,b})$$

and then to observe that the variance is a decreasing function of $\mu + \gamma$ (for $\mu + \gamma \geq 0$) since

$$\mathrm{Var}(g_{-\infty,b-\epsilon}) \leq \mathrm{Var}(f_{-\infty,b})$$

for any $\epsilon > 0$. Thus, it suffices to bound the variance of $g_{-\infty,b}$ when $\gamma + \mu = 0$, i.e.,

$$\begin{aligned}
&\frac{\int_{-\infty}^{b} x^2 e^{-x^2/2}\,dx}{\int_{-\infty}^{b} e^{-x^2/2}\,dx} - \left(\frac{\int_{-\infty}^{b} x e^{-x^2/2}\,dx}{\int_{-\infty}^{b} e^{-x^2/2}\,dx}\right)^2 \\
&\leq 1 - \left(\frac{\int_{-\infty}^{-b} x e^{-x^2/2}\,dx}{\int_{-\infty}^{b} e^{-x^2/2}\,dx}\right)^2 \\
&\leq 1 - \left(\frac{\int_{-\infty}^{-b} e^{-x^2/2}\,dx}{\int_{-\infty}^{-b} e^{-x^2/2}\,dx} \cdot b\right)^2 \\
&\leq 1 - \frac{e^{-b^2}}{2\pi}.
\end{aligned}$$

Here we used a tail bound on the standard Gaussian: the integral of the standard Gaussian density between b and ∞ is at most $e^{-b^2/2}/\sqrt{2\pi}b$. ∎

Lemma 4.7 readily leads to the following more succinct statement.

Lemma 4.8: Let g be the standard Gaussian density function in $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}_+$ be any logconcave function. Define the function h to be the density proportional to their product, i.e.,

$$h(x) = \frac{f(x)g(x)}{\int_{\mathbb{R}^n} f(x)g(x)\,dx}.$$

Then, for any unit vector $u \in \mathbb{R}^n$,

$$\mathrm{Var}_h(u \cdot x) \leq 1 - \frac{e^{-b^2}}{2\pi}$$

where the support of f along u is $[a_0, a_1]$ and $b = \min\{|a_0|, |a_1|\}$.

C. PCA works

As a warm-up, to convey the main idea, we prove the following theorem which assumes access to the full distribution (not just a sample).

Theorem 4.9: Let $F = (\mu, \Sigma)$ be a Gaussian distribution in $\mathbb{R}^n$ and F^+ be the restriction of F to a convex set K with normal subspace N of dimension k. Suppose the smallest k principal components of F^+ are $v_1, \ldots, v_k$. Then,

$$N = \mathrm{span}\{v_1, \ldots, v_k\}.$$

Proof: The first step of our algorithm effectively makes the input distribution a standard Gaussian. To simplify notation, we assume this is the original distribution. By the symmetry of the Gaussian density along directions orthogonal to N, the variances of F and F^+ are equal along any direction orthogonal to N. So we project both distributions to N, to get F_k and F_k^+. It is clear that F_k^+ is the restriction of F_k to K. Moreover, the mean μ^+ of F^+ lies in N. Let $P = K \cap N$ denote the positive region projected to N. We will now bound the variance of the Gaussian restricted to P along any direction in N. Let v be such a direction. Project the distribution to v.

For convenience, we can assume that $v = e_1$. Define the function $f : \mathbb{R} \to \mathbb{R}_+$ as:

$$f(z) = \int_{x \in \mathbb{R}^k : x_1 = z} h((x_2, \ldots, x_k))\chi^P(x)\,dx$$

where χ^P is the indicator function of P and h is the density of a $(k-1)$-dimensional Gaussian. The integrand above is a product of logconcave functions and hence also logconcave; the function f is logconcave since it is a projection of a logconcave function. The integral of the density over the cross-section of P orthogonal to v at z is proportional to $e^{-z^2/2}f(z)$.

Then, since the mean of the projected distribution is $\mu_v = \mu^+ \cdot v$, we have

$$\int_a^b z e^{-(z-\mu_v)^2/2} f(z - \mu_v)\,dz = 0$$

where $[a, b]$ is the support of P along v after shifting by μ_v. Then,

$$\frac{\int_a^b z^2 e^{-(z-\mu_v)^2/2} f(z-\mu_v)\,dz}{\int_a^b e^{-(z-\mu_v)^2/2} f(z-\mu_v)\,dz} < 1$$

using Lemma 4.7. Thus, the variance along a vector v is less than 1 if it lies in N and equal to 1 if it is orthogonal to N. ∎

The main remaining difficulty is to handle the error introduced by sampling. To do this, we derive bounds on the singular value gap and on the error in estimating singular values introduced by sampling to conclude that the subspace identified must indeed be close to the normal subspace N.

Proof: (of Theorem 1.3.) We assume w.l.o.g. that the full distribution is isotropic. Let N be the normal subspace of the convex set K used to label examples. Let the projection of the Gaussian restricted to K to the subspace N induce the density f_N. The projection of the full distribution F to N is a standard Gaussian. The support of f_N is a convex set $K \cap N$ and its mean is μ^+ (which lies in N). For convenience in the rest of this proof, we translate μ^+ to be origin. Define $M(u)$ be the second moment (equal to the variance since we shifted the mean to the origin) along a unit vector u according to the density f_K. Consider any direction $u \in N$ along which there is a tangent plane to K at distance at most r from the origin. Then, by Lemma 4.7,

$$M(u) = \mathbb{E}_{f_N}((u \cdot x)^2) < 1 - \frac{1}{8}e^{-r^2}.$$

Given this bound, the rest of the proof is conceptually straightforward. The vectors found by PCA are the k orthogonal vectors with smallest second moments. So, we aim to show that sampling preserves the gap between moments of vectors in the relevant subspace and those orthogonal to it. One complication is that although the relevant subspace has dimension k, it could effectively be lower-dimensional due to mild restrictions (boundary far from the origin) in some directions. The variance in such directions could be indistinguishably close to 1. To overcome this, we note that all that matters

is that we recover the subspace in which the moments are distinctly smaller than 1. This is because the concept is preserved (i.e., correctly classifies the input distribution) even upon projection to this possibly smaller subspace. We now make this argument formal.

Let $a(r) = (1/8)e^{-r^2}$. Let N_1 be the subspace of N in which for every $u \in N_1$, $M(u) \leq 1 - a(r)/2$. We will now choose our sample size so that the smallest k singular vectors span a subspace close to N_1. We first argue that any subspace close to N_1 (for a suitable choice of r) will be a good subspace, i.e., will preserve most of the concept upon projection.

Consider a ball in N of radius $r = \sqrt{k} + 2\sqrt{\log(1/\epsilon)}$ centered at the origin. By the concentration of a Gaussian (Lemma 4.1), the total measure of the projection of F outside this ball is at most ϵ^2. Let N_1 be spanned by orthogonal vector $v_1, \ldots, v_l$ such that each of their variances is less than $1 - \frac{1}{16}e^{-r^2}$, and the variance of any vector orthogonal to N_1 is higher than this. Then in the subspace orthogonal to their span, we can assume that the support of f contains the ball of radius r, and so this orthogonal subspace can be ignored. Therefore, the intersection of K with the subspace V spanned by $v_1, \ldots, v_l$ correctly classifies at least $1 - \epsilon^2$ of F projected to V.

When K is an intersection of k halfspaces, it suffices to set $r = 2\sqrt{\log(k/\epsilon)}$. A tangent plane at greater distance than r cuts off at most ϵ^2/k of the underlying Gaussian and therefore an intersection of k halfspaces, each at this distance from the origin, retains at least $1 - \epsilon^2$ measure.

Finally, we show that the subspace found by the algorithm contains a subspace close to N. The algorithm uses an i.i.d. sample of vectors $x_1, \ldots, x_m$ from the positive distribution. The estimate of the variance in any direction u is

$$\tilde{M}(u) = u^T \left(\frac{1}{m} \sum_{i=1}^{m} x_i x_i^T \right) u.$$

We choose m large enough to guarantee that for every direction $u \in \mathbb{R}^n$,

$$(1 - \gamma)M(u) \leq \tilde{M}(u) \leq (1 + \gamma)M(u)$$

with probability at least $1 - \delta/4$. Using Lemma 4.2, the number of samples to estimate the variance to within relative error γ in every direction is $O(n \log^2 n \log(1/\delta)/\gamma^2)$. We will choose γ at the end.

Thus, for unit vectors u orthogonal to N, the estimated moment is $\tilde{M}(u) \geq 1 - \gamma$, while for $u \in N_1$,

$$\tilde{M}(u) \leq (1 + \gamma)(1 - \frac{a(r)}{2}).$$

For general u, we write $u = \sqrt{\alpha}u_1 + \sqrt{(1-\alpha)}u_2$ where $u_1 \in N_1$, $u_2 \in N \cap N_1^{\perp}$ and $\alpha \in [0, 1]$. Then,

$$\tilde{M}(u) \geq (1 - \gamma)\left(\alpha M(u_1) + (1 - \alpha)M(u_2)\right).$$

If u is chosen as a small singular vector, then it must have second moment smaller than the second moment along u_1 as estimated by sampling, i.e.,

$$\left((1 - \gamma)\left(\alpha M(u_1) + (1 - \alpha)\right)\right) \leq (1 + \gamma)M(u_1)$$

which implies that

$$\alpha \geq 1 - \frac{2\gamma}{(1 - \gamma)} \frac{M(u_1)}{(1 - M(u_1))}$$

We now use the fact that $M(u_1) \leq 1 - (a(r)/2)$ and set $\gamma = \epsilon a(r)/36k$, to get $\alpha > 1 - (\epsilon/16k)$, i.e., the vector u is very close to N_1; its projection to N_1 has squared length at least $1 - (\epsilon/16k)$. Substituting this value of γ above gives the complexity bound.

Let V be the subspace found by the algorithm and V_1 be the subspace within it closest to N_1 of dimension l. First, by the definition of N_1, by projecting K to N_1, we have a hypothesis that correctly classifies $1 - \epsilon^2$ of the input distribution. Next noting that V_1 is close to N_1 in that every vector in V_1 has a large projection to N_1, we conclude that the projection of K to V_1 correctly labels at least $1 - \epsilon$ of the distribution. This is because the induced cylinders differ only near their boundary or far from the origin and the measure of their symmetric difference is at most $\epsilon/2$.

■

D. Proof of Theorems 1.1 and 1.2

Theorem 1.3 tells us that the algorithm identifies a subspace that preserves the unknown hypothesis approximately under projection to the subspace, i.e., we can assume that in the subspace V identified by the algorithm, there exists a concept from the same concept class that has error at most $\epsilon/2$ on the projected distribution. Applying Theorems 4.5 and 4.6 then give us the guarantees.

V. DISCUSSION

Learning low-dimensional concepts is a natural and fundamental framework in learning theory (e.g., learning from relevant variables or learning juntas [Lit87], [Blu94], [MOS03]). Among the most attractive instantiations is the open problem of PAC-learning an intersection of two halfspaces from an arbitrary distribution.

Here we made progress on learning under Gaussian distributions. We can learn arbitrary convex k-dimensional concepts with complexity growing only as a function of k and a fixed polynomial in n. For intersections of halfspaces, for any fixed error bound $\epsilon > 0$, the complexity is poly$(n)k^{O(\log k)}$ and so one can learn an intersection of up to $2^{O(\sqrt{\log n})}$ halfspaces in polynomial time. Previously it was known how to learn an intersection of a constant number of halfspaces from a Gaussian distribution in polynomial time.

The main ingredient of our algorithm, PCA, is fast, highly developed numerically, and widely used in practice, but has very few general guarantees. This might not be surprising given its simplicity and generality; however, it turns out to be provably effective for this problem.

We note that our algorithm does not apply to learning from other simple distributions, e.g., uniform in a ball. As remarked earlier, the variance can go up or down in a particular direction when a ball is restricted to a convex subset and so

the directions identified by PCA could lie far from the normal subspace.

Perhaps the two most directly related open problems are (a) to find a fully polynomial algorithm for learning an intersection of halfspaces from Gaussian input distributions (i.e., polynomial in k as well) and (b) to extend to other distributions, e.g., is there a polynomial-time algorithm to learn a polytope given uniform random points from it?

VI. Acknowledgments

We are grateful to the Adam Kalaivans duo and to an anonymous reviewer for helpful comments; to Emmanuel Milman for suggesting the proof of Lemma 4.3; and to the NSF for awards AF-0915903 and AF-0910584.

References

[Bau90a] Eric B. Baum, *On learning a union of half spaces*, J. Complexity **6** (1990), no. 1, 67–101.

[Bau90b] ———, *Polynomial time algorithms for learning neural nets*, COLT, 1990, pp. 258–272.

[BK93] Avrim Blum and Ravi Kannan, *Learning an intersection of k halfspaces over a uniform distribution*, FOCS, 1993, pp. 312–320.

[BK97] Avrim Blum and Ravindran Kannan, *Learning an intersection of a constant number of halfspaces over a uniform distribution*, J. Comput. Syst. Sci. **54** (1997), no. 2, 371–380.

[BL76] H. J. Brascamp and E. H. Lieb, *On extensions of the brunn-minkowski and prekopa-liendler theorems, including inequalities for log-concave functions, and with application to the diffusion equationslogarithmic concave measures and functions*, J. Functional Anal. **22** (1976), 366–389.

[BL00] S. G. Bobkov and M. Ledoux, *From brunn-minkowski to brascamp-lieb and to logarithmic sobolev inequalities*, Geom. Funct. Anal **10** (2000), 1028–1052.

[Blu94] Avrim L. Blum, *Relevant examples and relevant features: Thoughts from computational learning theory*, In AAAI Fall Symposium on Relevance, 1994.

[Goy05] Navin Goyal, *Open problems*, AIM workshop on Algorithmic Convex Geometry. http://www.aimath.org/WWN/convexgeometry/convexgeometry.pdf (2005).

[GR09] Navin Goyal and Luis Rademacher, *Learning convex bodies is hard*, COLT, 2009.

[KLS95] Ravi Kannan, László Lovász, and Miklós Simonovits, *Isoperimetric problems for convex bodies and a localization lemama*, Discrete & Computational Geometry **13** (1995), 541–559.

[KLT09] Adam R. Klivans, Philip M. Long, and Alex K. Tang, *Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions*, APPROX-RANDOM, 2009, pp. 588–600.

[KOS08] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio, *Learning geometric concepts via gaussian surface area*, FOCS, 2008, pp. 541–550.

[Lit87] Nick Littlestone, *Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm (extended abstract)*, FOCS, 1987, pp. 68–77.

[LS93] László Lovász and Miklós Simonovits, *Random walks in a convex body and an improved volume algorithm*, Random Struct. Algorithms **4** (1993), no. 4, 359–412.

[LV07] László Lovász and Santosh Vempala, *The geometry of logconcave functions and sampling algorithms*, Random Struct. Algorithms **30** (2007), no. 3, 307–358.

[MOS03] Elchanan Mossel, Ryan O'Donnell, and Rocco A. Servedio, *Learning juntas*, STOC, 2003, pp. 206–212.

[Rud99] Mark Rudelson, *Random vectors in the isotropic position*, Journal of Functional Analysis **164** (1999), no. 1, 60 – 72.

[She10] Alexander A. Sherstov, *Optimal bounds for sign-representing the intersection of two halfspaces by polynomials*, STOC, 2010.

[Vem97] Santosh Vempala, *A random sampling based algorithm for learning the intersection of half-spaces*, FOCS, 1997, pp. 508–513.

[Vem04] Santosh S. Vempala, *The random projection method*, AMS, 2004.

[Vem10] Santosh Vempala, *A random sampling based algorithm for learning the intersection of half-spaces*, JACM, to appear (2010).