Learning Low-Density Separators

Shai Ben-David and Tyler Lu and Dávid Pál
{shai,ttlu,dpal}@cs.uwaterloo.ca
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada

Miroslava Sotáková
mirka@daimi.au.dk
Dept of Computer Science
University of Aarhus
Denmark

Abstract

We define a novel, basic, unsupervised learning problem - learning the lowest density homogeneous hyperplane separator of an unknown probability distribution. Namely, given a random unlabeled sample generated by some unknown probability distribution, find linear separators that cut that distribution through low-density regions. This task is relevant to several problems in machine learning, such as semi-supervised learning and clustering stability. We investigate the question of existence of a universally consistent algorithm for this problem. We propose two natural learning paradigms and prove that, on input unlabeled random samples generated i.i.d. by any distribution, they are guaranteed to converge to the optimal separator for that distribution. We complement this result by showing that no learning algorithm for our task can achieve uniform learning rates (that are independent of the data generating distribution).

1 Introduction

While the theory of machine learning has achieved extensive understanding of many aspects of supervised learning, our theoretical understanding of unsupervised learning leaves a lot to be desired. In spite of the obvious practical importance of various unsupervised learning tasks, the state of our current knowledge does not provide anything that comes close to the rigorous mathematical performance guarantees that classification prediction theory enjoys.

In this paper we make a small step in that direction by analyzing one specific unsupervised learning task – the detection of low-density linear separators for data distributions over Euclidean spaces.

We consider the scenario in which some unknown probability distribution over $\mathbb{R}^n$ generates a (finite) i.i.d. sample. Taking such a sample as an input we seek to find a homogeneous hyperplane of lowest density that cuts through that distribution. We assume that the underlying data distribution has a continuous density function and define the density of a hyperplane as the integral of that density function over that hyperplane.

Our model can be viewed as a restricted instance of the fundamental issue of inferring information about a probability distribution from the random samples it generates. Tasks of that nature range from the ambitious problem of density estimation [8], through estimation of level sets [4], [13], [1], densest region detection [3], and, of course, clustering. All of these tasks are notoriously difficult with respect to both the sample complexity and the computational complexity aspects (unless one presumes strong restrictions about the nature of the underlying data distribution). Our task seems more modest than these, however, we believe that it is a basic and natural task that is relevant to various practical learning scenarios. We are not aware of any previous work on this problem (from the point of view of statistical machine learning, at least).

One important domain to which the detection of low-density linear data separators is relevant is semi-supervised learning [7]. Semi-supervised learning is motivated by the fact that in many real world classification problems, unlabeled samples are much cheaper and easier to obtain than labeled examples. Consequently, there is great incentive to develop tools by which such unlabeled samples can be utilized to improve the quality of sample based classifiers. Naturally, the utility of unlabeled data to classification depends on assuming some relationship between the unlabeled data distribution and the class membership of data points (see [5] for a rigorous discussion of this point). A common postulate of that type is that the boundary between data classes passes through low-density regions of the data distribution.

The Transductive Support Vector Machines paradigm (TSVM) [9] is an example of an algorithm that implicitly uses such a low density boundary assumption. Roughly speaking, TSVM searches for a hyperplane that has small error on the labeled data and at the same time has wide margin with respect to the unlabeled data sample.

Another area in which low-density boundaries play a significant role is the analysis of clustering stability. Recent work on the analysis of clustering stability found close relationship between the stability of a clustering and the data density along the cluster boundaries – roughly speaking, the lower these densities the more stable the clustering ([6], [12]).

A low-density-cut algorithm for a family $\mathcal{F}$ of probability distributions takes as an input a finite sample generated by some distribution $f \in \mathcal{F}$ and has to output a hyperplane through the origin with low density w.r.t. f. In particular, we consider the family of all distributions over $\mathbb{R}^n$ that have continuous density functions. We investigate two notions of success for low-density-cut algorithms – uniform convergence (over a family of probability distributions) and consistency. For uniform convergence we prove a general negative result, showing that no algorithm can guarantee any fixed convergence rates (in terms of sample sizes). This negative result holds even in the simplest case where the data domain is the one-dimensional unit interval. W

On the positive side, we prove the consistency of two natural algorithmic paradigms; *Soft-Margin* algorithms that choose a margin parameter (depending on the sample size) and output the separator with lowest empirical weight in the margins around it, and *Hard-Margin* algorithms that choose the separator with widest sample-free margins.

The paper is organized as follows: Section 2 provides the formal definition of our learning task as well as the success criteria that we investigate. In Section 3 we present two natural learning paradigms for the problem over the real line and prove their universal consistency over a rich class of probability distributions. Section 4 extends these results to show the learnability of lowest-density homogeneous linear cuts for probability distributions over R^d for arbitrary dimension, d. In Section 5 we show that the previous universal consistency results cannot be improved to obtain *uniform* learning rates (by any finite-sample based algorithm). We conclude the paper with a discussion of directions for further research.

2 Preliminaries

We consider probability distributions over $\mathbb{R}^d$. For concreteness, let the domain of the distribution be the d-dimensional unit ball.

A *linear cut learning algorithm* is an algorithm that takes as input a finite set of domain points, a sample $S \subseteq \mathbb{R}^d$, and outputs a homogenous hyperplane, $L(S)$ (determined by a weight vector, $\mathbf{w} \in \mathbb{R}^d$, such that $||\mathbf{w}||_2 = 1$).

We investigate algorithms that aim to detect hyperplanes with low density with respect to the sample-generating probability distribution.

Let $f : \mathbb{R}^d \to \mathbb{R}_0^+$ be a d-dimensional density function. We assume that f is continuous. For any homogeneous hyperplane $h(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \ : \ \mathbf{w}^T\mathbf{x} = 0\}$ defined by a unit weight vector $\mathbf{w} \in \mathbb{R}^d$, we consider the $(d-1)$-dimensional integral of the density over h,

$$\overline{f}(\mathbf{w}) := \int_{h(\mathbf{w})} f(\mathbf{x}) \, \mathrm{d}x \ .$$

Note that $\mathbf{w} \mapsto \overline{f}(\mathbf{w})$ is a continuous mapping defined on the $(d-1)$-sphere $\mathcal{S}^{d-1} = \{\mathbf{w} \in \mathbb{R}^d \ : \ \|\mathbf{w}\|_2 = 1\}$. Note that, for any such weight vector $\mathbf{w}$, $\overline{f}(\mathbf{w}) = \overline{f}(-\mathbf{w})$. For the 1-dimensional case, these hyperplanes are replaced by points, $\mathbf{x}$ on the real line, and $\overline{f}(\mathbf{x}) = f(\mathbf{x})$ – the density at the point $\mathbf{x}$.

Definition 1. *A* linear cut learning algorithm *is a function that maps samples to homogeneous hyperplanes. Namely,*

$$L : \bigcup_{m=1}^{\infty} (\mathbb{R}^d)^m \to \mathcal{S}^{d-1}.$$

When $d = 1$, we require that

$$L : \bigcup_{m=1}^{\infty} \mathbb{R}^m \to [0, 1].$$

(The intention is that L finds the lowest density linear separator of the sample generating distribution.)

Definition 2. *Let μ be a probability distribution and f its density function. For a weight vector* $\mathbf{w}$ *we define the half-spaces $h^+(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \ : \ \mathbf{w}^T\mathbf{x} \geq 0\}$ and $h^-(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \ : \ \mathbf{w}^T\mathbf{x} \leq 0\}$. For any weight vectors $\mathbf{w}$ and $\mathbf{w}'$,*

1. $D_E(\mathbf{w}, \mathbf{w}') = 1 - |\mathbf{w}^{\mathbf{T}}\mathbf{w}'|$
2. $D_\mu(\mathbf{w}, \mathbf{w}') =$ $\min\{\mu(h^+(\mathbf{w})\Delta h^+(\mathbf{w}')), \mu(h^-(\mathbf{w})\Delta h^+(\mathbf{w}'))\}$

3. $D_f(\mathbf{w}, \mathbf{w}') = |\overline{f}(\mathbf{w}') - \overline{f}(\mathbf{w})|$

We shall mostly consider the distance measure D_E in $\mathbb{R}^d$, for $d > 1$ and $D_E(x, y) = |x - y|$ for $x, y \in \mathbb{R}$. In theses cases we omit any explicit reference to D. All of our results hold as well when D is taken to be the probability mass of the symmetric difference between $L(S)$ and $\mathbf{w}^*$ and when D is taken to be $D(\mathbf{w}, \mathbf{w}') = |\overline{f}(\mathbf{w}) - \overline{f}(\mathbf{w}')|$.

Definition 3. *Let $\mathcal{F}$ denote a family of probability distributions over $\mathbb{R}^d$. We assume that all members of $\mathcal{F}$ have density functions, and identify a distribution with its density function. Let D denote a distance function over hyperplanes. For a linear cut learning algorithm, L, as above,*

1. *We say that L, is* consistent *for $\mathcal{F}$ w.r.t a distance measure D, if, for any probability distribution f in $\mathcal{F}$, if f attains a unique minimum density hyperplane then*

$$\forall \epsilon > 0 \quad \lim_{m \to \infty} \Pr_{S \sim f^m} [D(L(S), \mathbf{w}^*) \geq \epsilon] = 0. \quad (1)$$

   where $\mathbf{w}^*$ is the minimum density hyperplane for f.

2. *We say that L is* uniformly convergent *for $\mathcal{F}$ (w.r.t a distance measure, D), if, for every $\epsilon, \delta > 0$, there exists a $m(\epsilon, \delta)$ such that for any probability distribution $f \in \mathcal{F}$, if f has a unique minimizer $\mathbf{w}^*$ then, for all $m \geq m(\epsilon, \delta)$ we have*

$$\Pr_{S \sim f^m} [D(L(S), \mathbf{w}^*) \geq \epsilon] \leq \delta. \quad (2)$$

3 The One Dimensional Problem

Let $\mathcal{F}_1$ be the family of all probability distributions over the unit interval $[0, 1]$ that have continuous density function. We consider two natural algorithms for lowest density cut over this family. The first is a simple bucketing algorithm. We explain it in detail and show its consistency in section 3.1. The second algorithm is the *hard-margin* algorithm which outputs the midpoint of the largest gap between two consecutive points the sample. In section 3.2 we show *hard-margin* algorithm is consistent and in section 3.1 that the bucketing algorithm is consistent. In section 5 we show there are no algorithms that are uniformly convergent for $\mathcal{F}_1$.

3.1 The Bucketing Algorithm

The algorithm is parameterized by a function $k : \mathbb{N} \to \mathbb{N}$. For a sample of size m, the algorithm splits the interval $[0, 1]$ into $k(m)$ equal length subintervals (*buckets*). Given an input sample S, it counts the number of sample points lying in each bucket and outputs the mid-point of the bucket with fewest sample points. In case of ties, it picks the rightmost bucket. We denote this algorithm by B_k. As it turns out, there exists a choice of $k(m)$ which makes the algorithm B_k consistent for $\mathcal{F}_1$.

Theorem 4. *If the number of buckets $k(m) = o(\sqrt{m})$ and $k(m) \to \infty$ as $m \to \infty$, then the bucketing algorithm B_k is consistent for $\mathcal{F}_1$.*

Proof. Fix $f \in \mathcal{F}_1$, assume f has a unique minimizer x^*. Fix $\epsilon, \delta > 0$. Let $U = (x^* - \epsilon/2, x^* + \epsilon/2)$ be an neighbourhood of the unique minimizer x^*. The set $[0, 1] \setminus U$ is compact and hence there exists $\alpha := \min f([0, 1] \setminus U)$. Since x^* is the unique minimizer of f, $\alpha > f(x^*)$ and hence $\eta := \alpha - f(x^*)$ is positive. Thus, we can pick a neighbourhood V of x^*, $V \subset U$, such that for all $x \in V$, $f(x) < \alpha - \eta/2$.

The assumptions on growth of $k(m)$ imply that there exists m_0 such that for all $m \geq m_0$

$$1/k(m) < |V|/2 \quad (3)$$

$$2\sqrt{\frac{\ln(1/\delta)}{m}} < \frac{\eta}{2k(m)} \quad (4)$$

Fix any $m \geq m_0$. Divide $[0, 1]$ into $k(m)$ buckets each of length $1/k(m)$. For any bucket I, $I \cap U = \emptyset$,

$$\mu(I) \geq \frac{\alpha}{k(m)} . \quad (5)$$

Since $1/k(m) < |V|/2$ there exists a bucket J such that $J \subseteq V$. Furthermore,

$$\mu(J) \leq \frac{\alpha - \eta/2}{k(m)} . \quad (6)$$

For a bucket I, we denote by $|I \cap S|$ the number of sample points in the bucket I. From the well known Vapnik-Chervonenkis bounds [2], we have that with probability at least $1 - \delta$ over i.i.d. draws of sample S of size m, for any bucket I,

$$\left| \frac{|I \cap S|}{m} - \mu(I) \right| \leq \sqrt{\frac{\ln(1/\delta)}{m}} . \quad (7)$$

Fix any sample S satisfying the inequality (7) . For

any bucket I, $I \cap U = \emptyset$,

$$\begin{aligned}
\frac{|J \cap S|}{m} &\leq \mu(J) + \sqrt{\frac{\ln(1/\delta)}{m}} && \text{by (7)} \\
&\leq \frac{\alpha - \eta/2}{k(m)} + \sqrt{\frac{\ln(1/\delta)}{m}} && \text{by (6)} \\
&< \frac{\alpha}{k(m)} - 2\sqrt{\frac{\ln(1/\delta)}{m}} + \sqrt{\frac{\ln(1/\delta)}{m}} && \text{by (4)} \\
&\leq \mu(I) - \sqrt{\frac{\ln(1/\delta)}{m}} && \text{by (5)} \\
&\leq \frac{|I \cap S|}{m} && \text{by (7)}
\end{aligned}$$

Since $|J \cap S| > |I \cap S|$, the algorithm B_k must not output the mid-point of any bucket I for which $I \cap U = \emptyset$. Henceforth, the algorithm's output, $B_k(S)$, is the mid-point of an bucket I which intersects U. Thus the estimate $B_k(S)$ differs from x^* by at most the sum of the radius of the neighbourhood U and the radius of the bucket. Since the length of a bucket is $1/k < |V|/2$ and $V \subset U$, the sum of the radii is

$$|U|/2 + |V|/4 < \frac{3}{4}|U| < \epsilon \, .$$

Combining all the above, we have that for any $\epsilon, \delta > 0$ there exists m_0 such that for any $m \geq m_0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, $|B_k(S) - x^*| < \epsilon$. This is the same as saying that B_k is consistent for f. □

Note that in the above proof we cannot replace the condition $k(m) = o(\sqrt{m})$ with $k(m) = O(\sqrt{m})$ since Vapnik-Chervonenkis bounds do not allow us to detect $O(1/\sqrt{m})$-difference between probability masses of two buckets.

The following theorems shows that if there are too many buckets the bucketing algorithm is not consistent anymore.

Theorem 5. *If the number of buckets $k(m) = \omega(m/\log m)$, then B_k is not consistent for $\mathcal{F}_1$.*

To prove the theorem we need a proposition of the following lemma dealing with the classical coupon collector problem.

Lemma 6 (The Coupon Collector Problem [11])**.** *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then for any constant $c \in \mathbb{R}$, and $m = n \ln n + cn$,*

$$\lim_{n \to \infty} \Pr[X > m] = 1 - e^{-e^{-c}} \, .$$

Proof of Theorem 5. Consider the following density f on $[0, 1]$,

$$f(x) = \begin{cases} (4 - 16x)/3 & \text{if } x \in [0, \frac{1}{4}] \\ (16x - 4)/3 & \text{if } x \in (\frac{1}{4}, \frac{1}{2}) \\ 4/3 & \text{if } x \in [\frac{1}{2}, 1] \end{cases}$$

which attains unique minimum at $x^* = 1/4$.

From the assumption on the growth of $k(m)$ for all sufficiently large m, $k(m) > 4$ and $k(m) > 8m/\ln m$. Consider all buckets lying in the interval $[\frac{1}{2}, 1]$ and denote them by $b_1, b_2, \ldots, b_n$. Since the bucket size is less than $1/4$, they cover the interval $[\frac{3}{4}, 1]$. Hence their length total length is at least $1/4$ and hence there are

$$n \geq k(m)/4 > 2m/\ln m$$

such buckets.

We will show that for m large enough, with probability at least $1/2$, at least one of the buckets $b_1, b_2, \ldots, b_n$ receives no sample point. Since probability masses of $b_1, b_2, \ldots, b_n$ are the same, we can think of these buckets as coupon types we are collecting and the sample points as coupons. By Lemma 6, it suffices to verify, that the number of trials, m, is at most, say, $\frac{2}{3} n \ln n$. Indeed, we have for large enough m

$$\begin{aligned}
\frac{2}{3} n \ln n \geq \frac{2}{3} \frac{2m}{\ln m} \ln\left(\frac{2m}{\ln m}\right) = \\
\frac{4}{3} \frac{m}{\ln m} (\ln m + \ln 2 - \ln \ln m) \geq m \, .
\end{aligned}$$

Now, Lemma 6 implies that for sufficiently large m, with probability at least $1/2$, at least one of the buckets $b_1, b_2, \ldots, b_n$ contains no sample point.

If there are empty buckets in $[\frac{1}{2}, 1]$, the algorithm outputs a point in $[\frac{1}{2}, 1]$. Since this happens with probability at least $1/2$ and since $x^* = 1/4$, the algorithm cannot be consistent. □

When the number of buckets $k(m)$ is asymptotically somewhere in between $\sqrt{m}$ and $m/\ln m$, the bucketing algorithm switches from being consistent to failing consistency. It remains an open question to determine where exactly the transition occurs.

3.2 The Hard-Margin Algorithm

Let the *hard-margin* algorithm be the function that outputs the mid-point of the largest interval between the adjacent sample points. More formally, given a sample S of size m, the algorithm sorts the sample $S \cup \{0, 1\}$ so that $x_0 = 0 \leq x_1 \leq x_2 \leq \cdots \leq x_m \leq 1 = x_{m+1}$ and outputs the midpoint $(x_i + x_{i+1})/2$ where the index i, $0 \leq i \leq m$, is such that the gap $[x_i, x_{i+1}]$ is the largest.

Henceforth, the notion *largest gap* refers to the length of the largest interval between the adjacent points of a sample.

Theorem 7. *The hard-margin algorithm is consistent for the family $\mathcal{F}_1$.*

To prove the theorem we need the following property of the distribution of the largest gap between two adjacent elements of m points forming an i.i.d. sample from the uniform distribution on $[0,1]$. The following statement follows as a corollary of Lévy's work [10]. However, we will present a direct and much simpler proof.

Lemma 8. *Let L_m be the random variable denoting the largest gap between adjacent points of an i.i.d. sample of size m from the uniform distribution on $[0,1]$. For any $\epsilon > 0$*

$$\lim_{m\to\infty} \Pr\left[L_m \in \left((1-\epsilon)\frac{\ln m}{m}, (1+\epsilon)\frac{\ln m}{m}\right)\right] = 1.$$

Proof of Lemma. Consider the uniform distribution over the unit circle. Suppose we draw an i.i.d. sample of size m from this distribution. Let K_m denote the size of the largest gap between two adjacent samples. It is not hard so see that the distribution of K_m is the same as that of L_{m-1}. Furthermore, since $\frac{\ln(m)/m}{\ln(m+1)/(m+1)} \to 1$, we can thus prove the lemma with L_m replaced by K_m.

Fix $\epsilon > 0$. First, let us show that for m sufficiently large K_m is with probability $1 - o(1)$ above the lower bound $(1-\epsilon)\frac{\ln m}{m}$. We split the unit circle $b = \frac{m(1-\epsilon)}{\ln m}$ buckets, each of length $(1-\epsilon)\frac{\ln m}{m}$. It follows from Lemma 6, that for any constant $\zeta > 0$ and an i.i.d. sample of $(1-\zeta)b\ln b$ points at least one bucket is empty with probability $1 - o(1)$. We show that for some ζ, $m \le (1-\zeta)b\ln b$. The expression on the right side can be rewritten as

$$\begin{aligned}(1-\zeta)b\ln b &= \frac{(1-\zeta)(1+\delta)m}{\ln m}\ln\left(\frac{(1-\zeta)(1+\delta)m}{\ln m}\right)\\ &\ge m(1-\zeta)(1+\delta)\left(1 - O\left(\frac{\ln\ln m}{\ln m}\right)\right)\end{aligned}$$

For ζ sufficiently small and m sufficiently large the last expression is greater than m, yielding that a sample of m points misses at least one bucket with probability $1 - o(1)$. Therefore, the largest gap K_m is with probability $1 - o(1)$ at least $(1-\epsilon)\frac{\ln m}{m}$.

Next, we show that for m sufficiently large, K_m is with probability $1-o(1)$ below the upper bound $(1+\epsilon)\frac{\ln m}{m}$. We consider $3/\epsilon$ bucketings $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{3/\epsilon}$. Each bucketing $\mathcal{B}_i$, $i = \{1, 2, \ldots, (3/\epsilon)\}$, is a division of the unit circle into $b = \frac{m}{(1+\epsilon/3)\ln m}$ equal length buckets; each bucket has length $\ell = (1+\epsilon/3)\frac{\ln m}{m}$. The bucketing $\mathcal{B}_i$ will have its left end-point of the first bucket at position $i(\ell\epsilon/3)$. The position of the left end-point of the first bucket of a bucketing is called the *offset* of the bucketing.

We first show that there exists $\zeta > 0$ such that $m \ge (1+\zeta)b\ln b$ for all sufficiently large m. Indeed,

$$\begin{aligned}(1+\zeta)b\ln b &= (1+\zeta)\frac{m}{(1+\epsilon/3)\ln m}\ln\frac{m}{(1+\epsilon/3)\ln m}\\ &\le \frac{1+\zeta}{1+\epsilon/3}m\left(1 - O\left(\frac{\ln\ln m}{\ln m}\right)\right).\end{aligned}$$

For any $\zeta < \epsilon/3$ and sufficiently large m the last expression is greater than m.

The existence of such ζ and Lemma 6 guarantee that for all sufficiently large m, for of each bucketing $\mathcal{B}_i$, with probability $1-o(1)$, each bucket is hit by a sample point. We now apply union bound and get that, for all sufficiently large m, with probability $1 - (3/\epsilon)o(1) = 1-o(1)$, for each bucketing $\mathcal{B}_i$, each bucket is hit by at least one sample point. Consider any sample S such that for each bucketing, each bucket is hit by at least one point of S. Then, the largest gap in S can not be bigger than the bucket size plus the difference of offsets between two adjacent bucketings, since otherwise the largest gap would demonstrate an empty bucket in at least one of the bucketings. In other words, the largest gap K_m is at most

$$(\ell\epsilon/3)+\ell = (1+\epsilon/3)\ell = (1+\epsilon/3)^2\frac{\ln m}{m} < (1+\epsilon)\frac{\ln m}{m}$$

for any $\epsilon < 1$. □

Proof of the Theorem. Consider any two disjoint intervals $U, V \subseteq [0,1]$ such that for any $x \in U$ and any $y \in V$, $\frac{f(x)}{f(y)} < p < 1$ for some $p \in (0,1)$. We claim that with probability $1 - o(1)$, the largest gap in U is bigger than the largest gap in V.

If we draw an i.i.d. sample m points from μ, according to the law of large numbers for an arbitrarily small $\chi > 0$, the ratio between the number of points m_U in the interval U and the number of points m_V in the interval V with probability $1 - o(1)$ satisfies

$$\frac{m_U}{m_V} \le p(1+\chi)\frac{|U|}{|V|}. \tag{8}$$

For a fixed χ, choose a constant $\epsilon > 0$ such that $\frac{1-\epsilon}{1+\epsilon} > p + \chi$.

From Lemma 8 we show that with probability $1 - o(1)$ the largest gap between adjacent sample points falling into U is at least $(1-\epsilon)|U|\frac{\ln m_U}{m_U}$. Similarly, with

probability $1-o(1)$ the largest gap between adjacent sample points falling into V is at most $(1+\epsilon)|V|\frac{\ln m_V}{m_V}$. From (8) it follows that the ratio of gap sizes with probability $1-o(1)$ is at least

$$\frac{(1-\epsilon)|U|\frac{\ln m_U}{m_U}}{(1+\epsilon)|V|\frac{\ln m_V}{m_V}} > \frac{1-\epsilon}{1+\epsilon}\frac{1}{p+\chi}\frac{\ln m_U}{\ln m_V} = (1+\gamma)\frac{\ln m_U}{\ln m_V}$$

$$\geq (1+\gamma)\frac{\ln((p+\chi)\frac{|U|}{|V|}m_V)}{\ln m_V}$$

$$= (1+\gamma)\,(1+O(1)/\ln m_V) \to (1+\gamma) \quad \text{as } m\to\infty$$

for a constant $\gamma > 0$ such that $1+\gamma \leq \frac{1-\epsilon}{1+\epsilon}\frac{1}{p+\chi}$. Hence for sufficiently large m with probability $1-o(1)$, the largest gap in U is strictly bigger than the largest gap in V.

Now, we can choose intervals V_1, V_2 such that $[0,1]\setminus(V_1\cup V_2)$ is an arbitrarily small neighbourhood containing x^*. We can pick an even smaller neighbourhood U containing x^* such that for all $x\in U$ and all $y\in V_1\cup V_2$, $\frac{f(x)}{f(y)} < p < 1$ for some $p\in(0,1)$. Then with probability $1-o(1)$, the largest gap in U is bigger than largest gap in V_1 and the largest gap in V_2. □

4 Learning Linear Cut Separators in High Dimensions

In this section we consider the problem of learning the minimum density homogeneous (i.e. passing through origin) linear cut in distributions over $\mathbb{R}^d$. Namely, assuming that some unknown probability distribution generates i.i.d. finite sample of points in $\mathbb{R}^d$. We wish to process these samples to find the $(d-1)$-dimensional hyperplane, through the origin of $\mathbb{R}^d$, that has the lowest probability density with respect to the sample-generating distribution. In other words, we wish to find how to cut the space $\mathbb{R}^d$ through the origin in the "sparsest direction".

Formally, let $\mathcal{F}_d$ be the family of all probability distributions over the $\mathbb{R}^d$ that have a continuous density function. We wish to show that there exists a linear cut learning algorithm that is consistent for $\mathcal{F}_d$. Note by Theorem 10, no algorithm achieves uniform convergence for $\mathcal{F}_d$ (even for $d=1$).

Define the *soft-margin* algorithm with parameter $\gamma : \mathbb{N}\to\mathbb{R}^+$ as follows. Given a sample S of size m, it counts for every hyperplane, the number of sample points lying within distance $\gamma := \gamma(m)$ and outputs the hyperplane with the lowest such count. In case of the ties, it breaks them arbitrarily. We denote this algorithm by H_γ. Formally, for any weight vector $\mathbf{w}\in\mathcal{S}^{d-1}$ (the unit sphere in $\mathbb{R}^d$) we consider the "γ-strip"

$$h(\mathbf{w},\gamma) = \{\mathbf{x}\in\mathbb{R}^d \;:\; |\mathbf{w}^T\mathbf{x}|\leq\gamma\}$$

and count the number of sample points lying in it. We output the weight vector $\mathbf{w}$ for which the number of sample points in $h(\mathbf{w},\gamma)$ is the smallest; we break ties arbitrarily.

To fully specify the algorithm, it remains to specify the function $\gamma(m)$. As it turns out, there is a choice of the function $\gamma(m)$ which makes the algorithm consistent.

Theorem 9. *If $\gamma(m) = \omega(1/\sqrt{m})$ and $\gamma(m)\to 0$ as $m\to\infty$, then H_γ is consistent for $\mathcal{F}_d$.*

Proof. The structure of the proof is similar to the proof of Theorem 4. However, we will need more technical tools.

First let's fix f. For any weight vector $w\in\mathcal{S}^{d-1}$ and any $\gamma>0$, we define $\overline{\overline{f}}_\gamma(\mathbf{w})$ as the d-dimensional integral

$$\overline{\overline{f}}_\gamma(w) := \int_{h(\mathbf{w},\gamma)} f(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

over γ-strip along $\mathbf{w}$. Note that for any $\mathbf{w}\in\mathcal{S}^{d-1}$,

$$\lim_{m\to\infty}\frac{\overline{\overline{f}}_{\gamma(m)}(\mathbf{w})}{\gamma} = \overline{f}(\mathbf{w})$$

(assuming that $\gamma(m)\to 0$). In other words, the sequence of functions $\left\{\overline{\overline{f}}_{\gamma(m)}/\gamma(m)\right\}_{m=1}^{\infty}$, $\overline{\overline{f}}/\gamma(m) : \mathcal{S}^{d-1}\to\mathbb{R}_0^+$, converges point-wise to the function $\overline{f}:\mathcal{S}^{d-1}\to\mathbb{R}_0^+$.

Note that $\overline{\overline{f}}/\gamma(m):\mathcal{S}^{d-1}\to\mathbb{R}_0^+$ is continuous for any m, and recall that $\mathcal{S}^{d-1}$ is compact. Therefore the sequence $\left\{\overline{\overline{f}}_{\gamma(m)}/\gamma(m)\right\}_{m=1}^{\infty}$ converges uniformly to $\overline{f}$. In other words, for every $\zeta>0$ there exists m_0 such that for any $m\geq 0$ and any $\mathbf{w}\in\mathcal{S}^{d-1}$,

$$\left|\frac{\overline{\overline{f}}_{\gamma(m)}(\mathbf{w})}{\gamma(m)} - \overline{f}(\mathbf{w})\right| < \zeta\,.$$

Fix f and $\epsilon,\delta>0$. Let $U = \{\mathbf{w}\in\mathcal{S}^{d-1} \;:\; |\mathbf{w}^T\mathbf{w}^*| > 1-\epsilon\}$ be the "ϵ-double-neighbourhood" of the antipodal pair $\{\mathbf{w}^*,-\mathbf{w}^*\}$. The set $\mathcal{S}^{d-1}\setminus U$ is compact and hence $\alpha := \min\overline{f}(\mathcal{S}^{d-1}\setminus U)$ exists. Since $\mathbf{w}^*,-\mathbf{w}^*$ are the only minimizers of $\overline{f}$, $\alpha > \overline{f}(\mathbf{w}^*)$ and hence $\eta := \alpha - \overline{f}(\mathbf{w}^*)$ is positive.

The assumptions on $\gamma(m)$ imply that there exists m_0 such that for all $m\geq m_0$,

$$2\sqrt{\frac{d+\ln(1/\delta)}{m}} < \frac{\eta}{3}\,\gamma(m) \tag{9}$$

$$\left|\frac{\overline{\overline{f}}_{\gamma(m)}(\mathbf{w})}{\gamma(m)} - \overline{f}(\mathbf{w})\right| < \eta/3 \qquad \text{for all } \mathbf{w}\in\mathcal{S}^{d-1} \tag{10}$$

Fix any $m \geq m_0$. For any $\mathbf{w} \in \mathcal{S}^{d-1} \setminus U$, we have

$$\begin{aligned}
\frac{\overline{\overline{f}}_{\gamma(m)}(\mathbf{w})}{\gamma(m)} &> \overline{f}(\mathbf{w}) - \eta/3 && \text{by (10)} \\
&\geq \overline{f}(\mathbf{w}^*) + \eta - \eta/3 \\
&\text{(by choice of } \eta \text{ and } U) \\
&= \overline{f}(\mathbf{w}^*) + 2\eta/3 \\
&> \frac{\overline{\overline{f}}_{\gamma(m)}(\mathbf{w}^*)}{\gamma(m)} - \eta/3 + 2\eta/3 && \text{by (10)} \\
&= \frac{\overline{\overline{f}}_{\gamma(m)}(\mathbf{w}^*)}{\gamma(m)} + \eta/3 \,.
\end{aligned}$$

From the above chain of inequalities, after multiplying by $\gamma(m)$, we have

$$\overline{\overline{f}}_{\gamma(m)}(\mathbf{w}) > \overline{\overline{f}}_{\gamma(m)}(\mathbf{w}^*) + \eta\gamma(m)/3 \,. \qquad (11)$$

From the well known Vapnik-Chervonenkis bounds [2], we have that with probability at least $1-\delta$ over i.i.d. draws of S of size m we have that for any $\mathbf{w}$,

$$\left| \frac{|h(\mathbf{w},\gamma) \cap S|}{m} - \overline{\overline{f}}_{\gamma(m)}(\mathbf{w}) \right| \leq \sqrt{\frac{d + \ln(1/\delta)}{m}} \,, \qquad (12)$$

where $|h(\mathbf{w},\gamma) \cap S|$ denotes the number of sample points lying in the γ-strip $h(\mathbf{w},\gamma)$.

Fix any sample S satisfying the inequality (12). We have, for any $\mathbf{w} \in \mathcal{S}^{d-1} \setminus U$,

$$\begin{aligned}
\frac{|h(\mathbf{w},\gamma) \cap S|}{m} &\geq \overline{\overline{f}}_{\gamma(m)}(\mathbf{w}) - \sqrt{\frac{d + \ln(1/\delta)}{m}} \\
&> \overline{\overline{f}}_{\gamma(m)}(\mathbf{w}^*) + \frac{\eta\gamma(m)}{3} - \sqrt{\frac{d + \ln(1/\delta)}{m}} \\
&\geq \frac{|h(\mathbf{w}^*,\gamma) \cap S|}{m} - \sqrt{\frac{d + \ln(1/\delta)}{m}} + \frac{\eta\gamma}{3} \\
&\quad - \sqrt{\frac{d + \ln(1/\delta)}{m}} \\
&> \frac{|h(\mathbf{w}^*,\gamma) \cap S|}{m}
\end{aligned}$$

Since $|h(\mathbf{w},\gamma) \cap S| > |h(\mathbf{w}^*,\gamma) \cap S|$, the algorithm must not output a weight vector $\mathbf{w}$ lying in $\mathcal{S}^{d-1} \setminus U$. In other words, the algorithm's output, $H_\gamma(S)$, lies in U i.e. $|H_\gamma(S)^T \mathbf{w}^*| > 1 - \epsilon$.

We have proven, that for any $\epsilon, \delta > 0$, there exists m_0 such that for all $m \geq m_0$, if a sample S is drawn i.i.d. from f, then $|H_\gamma(S)^T \mathbf{w}^*| > 1 - \epsilon$. In other words, H_γ is consistent for f. □

5 The Impossibility of Uniform Convergence

In this section we show a negative result that roughly says one cannot hope for an algorithm that can achieve ϵ accuracy and $1-\delta$ confidence for sample sizes that only depend on these parameters and not on properties of the probability measure.

Theorem 10. *No linear cut learning algorithm is uniformly convergent for $\mathcal{F}_1$ with respect to any of the distance functions D_E, D_f and D_μ.*

Proof. For a fixed $\delta > 0$ we show that for any $m \in \mathbb{N}$ there are distributions with density functions f and g such that no algorithm using a random sample of size at most m drawn from one of the distributions chosen uniformly at random, can identify the distribution with probability of error less than $1/2$ with probability at least δ over random choices of a sample.

Since for any δ and m we find densities f and g such that with probability more than $(1-\delta)$ the output of the algorithm is bounded away by $1/4$ from either $1/4$ or $3/4$, for the family $\mathcal{F}_1$ no algorithm converges uniformly w.r.t. any distance measure.

Consider two partly linear density functions f and g defined in $[0,1]$ such that for some n, f is linear in the intervals $[0, \frac{1}{4} - \frac{1}{2n}]$, $[\frac{1}{4} - \frac{1}{2n}, \frac{1}{4}]$, $[\frac{1}{4}, \frac{1}{4} + \frac{1}{2n})]$, and $[\frac{1}{4} + \frac{1}{2n}, 1]$, and satisfies

$$f(0) = f\left(\frac{1}{4} - \frac{1}{2n}\right) = f\left(\frac{1}{4} + \frac{1}{2n}\right) = f(1),\ f(\frac{1}{4}) = 0$$

and g is the reflection of f w.r.t. to the centre of the unit interval, i.e. $f(x) = g(1-x)$.

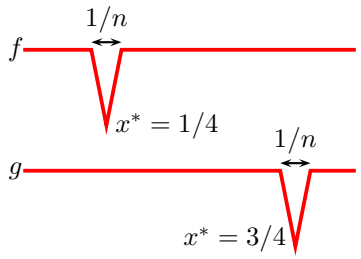

Figure 1: f is uniform everywhere except a small neighbourhood around $1/4$ where it has a sharp 'v' shape. And g is the reflection of f about $x = 1/2$.

Let us lower-bound the probability that a sample of size m drawn from f misses the set $U \cup V$ for $U := [\frac{1}{4} - \frac{1}{2n}, \frac{1}{4} + \frac{1}{2n}]$ and $V := [\frac{3}{4} - \frac{1}{2n}, \frac{3}{4} + \frac{1}{2n}]$. For any $x \in U$ and $y \notin U$, $f(x) \leq f(y)$, and furthermore, f is constant on the set $[0,1] \setminus U$ containing at most the entire probability mass 1. Therefore, for $p_f(Z)$

denoting the probability that a point drawn from the distribution with the density f hits the set Z, we have $p_f(U) \leq p_f(V) \leq \frac{1}{n-1}$, yielding that $p_f(U \cup V) \leq \frac{2}{n-1}$. Hence, an i.i.d. sample of size m misses $U \cup V$ with probability at least $(1 - 2/(n-1))^m \geq (1-\eta)e^{-2m/n}$ for any constant $\eta > 0$ and n sufficiently large. For a proper η and n sufficiently large we get $(1-\eta)e^{-2m/n} > 1 - \delta$. From the symmetry between f and g, a random sample of size m drawn from g misses $U \cup V$ with the same probability.

We have shown that for any $\delta > 0$, $m \in \mathbb{N}$, and for n sufficiently large, regardless of whether the sample is drawn from either of the two distributions, it does not intersect $U \cup V$ with probability more than $1 - \delta$. Since in $[0,1] \setminus (U \cup V)$ both density functions are equal, the probability of error in the discrimination between f and g conditioned on that the sample does not intersect $U \cup V$ cannot be less than $1/2$. □

6 Conclusions and Open Questions

In this paper have presented a novel unsupervised learning problem that is modest enough to allow learning algorithm with asymptotic learning guarantees, while being relevant to several central challenging learning tasks. Our analysis can be viewed as providing justification to some common semi-supervised learning paradigms, such as the maximization of margins over the unlabeled sample or the search for empirically-sparse separating hyperplanes. As far as we know, our results provide the first performance guarantees for these paradigms.

From a more general perspective, the paper demonstrates some type of meaningful information about a data generating probability distribution that can be reliably learned from finite random samples of that distribution, in a fully non-parametric model – without postulating any prior assumptions about the structure of the data distribution. As such, the search for a low-density data separating hyperplane can be viewed as a basic tool for the initial analysis of unknown data. Analysis that can be carried out in situations where the learner has no prior knowledge about the data in question and can only access it via unsupervised random sampling.

Our analysis raises some intriguing open questions. First, note that while we prove the universal consistency of the 'hard-margin' algorithm for Real data distributions, we do not have a similar result for higher dimensional data. Since searching for empirical maximal margins is a common heuristic, it is interesting to resolve the question of consistency of such algorithms.

Another natural research direction that this work calls for is the extension of our results to more complex separators. In clustering, for example, it is common to search for clusters that are separated by sparse data regions. however, such between-cluster boundaries are often not linear. Can one provide any reliable algorithm for the detection of sparse boundaries from finite random samples when these boundaries belong to a richer family of functions?

Our research has focused on the information complexity of the task. However, to evaluate the practical usefulness of our proposed algorithms, one should also carry a computational complexity analysis of the low-density separation task. We conjecture that the problem of finding the homogeneous hyperplane with largest margins, or lowest density around it (with respect to a finite high dimensional set of points) is NP-hard (when the Euclidean dimension is considered as part of the input, rather than as a fixed constant parameter). However, even if this conjecture is true, it will be interesting to find efficient approximation algorithms for these problems.

References

[1] C. Scott A. Singh and R. Nowak. Adaptive hausdorff estimation of density level sets. In *COLT*, 2008.

[2] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[3] Shai Ben-David, Nadav Eiron, and Hans-Ulrich Simon. The computational complexity of densest region detection. *J. Comput. Syst. Sci.*, 64(1):22–47, 2002.

[4] Shai Ben-David and Michael Lindenbaum. Learning distributions by their density levels: A paradigm for learning without a teacher. *J. Comput. Syst. Sci.*, 55(1):171–182, 1997.

[5] Shai Ben-David and Tyler Lu Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, 2008.

[6] Shai Ben-David and Ulrike von Luxburg. Relating clustering stability to properties of cluster boundaries. In *COLT*, 2008.

[7] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[8] Luc Devroye and Gábor Lugosi, editors. *Combinatorial Methods in Density Estimation*. Springer-Verlag, 2001.

[9] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[10] Paul Lévy. Sur la division d'un segment par des points choisis au hasard. *C.R. Acad. Sci. Paris*, 208:147–149, 1939.

[11] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[12] Ohad Shamir and N. Tishby. Model selection and stability in k-means clustering. In *COLT*, 2008.

[13] A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.