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
Departement of Computer Science
University of Aarhus
Denmark

Abstract

We define a novel, basic, unsupervised learning problem – learning hyperplane passing through the origin with the lowest probability density. Namely, given a random sample generated by some unknown probability distribution, the task is to find a hyperplane passing through the origin with smallest integral of the probability density on the hyperplane. 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 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 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.

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

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 hyperplane passing through the origin with lowest probability density. 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 (Devroye and Lugosi, 2001), through estimation of level sets (Ben-David and Lindenbaum, 1997; Tsybakov, 1997; Singh et al., 2008), densest region detection (Ben-David et al., 2002), 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 separators is relevant is semi-supervised learning (Chapelle et al., 2006). 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 classifi-

cation depends on assuming some relationship between the unlabeled data distribution and the class membership of data points; see (Ben-David et al., 2008) 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. Transductive Support Vector Machine (TSVM) by Joachims (1999) 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 (Ben-David and von Luxburg, 2008; Shamir and Tishby, 2008).

An algorithm for the lowest-density-hyperplane problem takes as an input a finite sample generated by some distribution and has to output a hyperplane passing through the origin with the smallest integral of the probability density *on* the hyperplane. We investigate two notions of success for these 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.

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 a simpler, but related problem over the real line and prove their consistency. Section 4 extends these results to show the learnability of lowest-density hyperplanes for probability distributions over $\mathbb{R}^d$ 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

For dimension $d \geq 2$, a *learning algorithm* L for the lowest-density-hyperplane problem is a function that takes as an input a finite sample $S \in \bigcup_{m=1}^{\infty}(\mathbb{R}^d)^m$ and outputs a unit weight vector $\mathbf{w} \in \mathbb{R}^d$ which is the normal of a homogeneous[1] hyperplane $\mathbf{w}^T\mathbf{x} = 0$. We assume that S is an i.i.d. sample from a probability measure μ over $\mathbb{R}^d$ with *continuous* density function $f : \mathbb{R}^d \to \mathbb{R}_0^+$. The goal of the algorithm is to find a hyperplane such that the integral of the density function over the hyperplane is as small as possible.

More precisely, for a unit vector $\mathbf{w} \in \mathbb{R}^d$ we define the hyperplane $h(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \ : \ \mathbf{w}^T\mathbf{x} = 0\}$ and the "density on the hyperplane"

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

The mapping $\overline{f} : \mathbf{w} \mapsto \overline{f}(\mathbf{w})$ is a continuous function defined on the $(d-1)$-sphere $\mathcal{S}^{d-1} = \{\mathbf{w} \in \mathbb{R}^d \ : \ \|\mathbf{w}\|_2 = 1\}$ which is compact. In particular, $\overline{f}$ attains global minimum at some point $\mathbf{w}^*$. The minimum of $\overline{f}$ is never unique, since $\overline{f}$ satisfies the obvious symmetry $\overline{f}(\mathbf{w}) = \overline{f}(-\mathbf{w})$ for any $\mathbf{w} \in \mathcal{S}^{d-1}$. However, if up to this symmetry the global minimum $\mathbf{w}^*$ is unique, the algorithm L is required to output $\mathbf{w}$ which is with high probability "close" to $\mathbf{w}^*$.

To define what "close" means, we consider several distance measures[2] over $\mathcal{S}^{d-1}$. For a weight vector $\mathbf{w} \in \mathbb{R}^d$ we define the half-spaces $h^+(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \ : \ \mathbf{w}^T\mathbf{x} \geq 0\}$ and $h^-(\mathbf{w}) = \{\mathbf{x} \in \mathbb{R}^d \ : \ \mathbf{w}^T\mathbf{x} \leq 0\}$. We define the following three distance measures.

Definition 1. *For* $\mathbf{w}, \mathbf{w}' \in \mathcal{S}^{d-1}$ *let*

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

Note that all three distance measures above respect the symmetry of $\overline{f}$ i.e. $D(\mathbf{w}, -\mathbf{w}) = 0$.

We say that the algorithm L is consistent w.r.t. a distance measure D, if $D(\mathbf{w}^*, L(S))$ converges in probability to zero as the sample size m increases. Formally, L is *consistent* if for any probability measure μ with

[1] homogeneous = passing through the origin

[2] Technically speaking, by a distance measure we mean a pseudo-metric. Recall that a pseudo-metric D is a metric except that it does not need to satisfy the condition that if $D(\mathbf{w}, \mathbf{w}') = 0$ then $\mathbf{w} = \mathbf{w}'$.

continuous density f such that the minimum $\mathbf{w}^*$ of $\overline{f}$ is (up to symmetry) unique, we have

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

It is not hard to see if L is consistent w.r.t. D_E, then it is consistent w.r.t. to other two distance measures as well. This follows from continuity of $\overline{f}$ and the existence of the density function f. For that reason when talking about consistency we consider only the pseudo-metric D_E and omit the explicit reference to the distance measure.

A natural question is whether one can guarantee the speed of the convergence $D(\mathbf{w}^*, L(S)) \to 0$ which would *not* depend on the probability distribution μ. Such guarantee is called uniform convergence.

Definition 2. *Let $\mathcal{F}$ be a family of probability distributions over $\mathbb{R}^d$ and D a distance measure over S^{d-1}. We say that algorithm L is* $\mathcal{F}$-uniformly convergent *if for every $\epsilon, \delta > 0$, there exists sample size $m(\epsilon, \delta)$ such that for any probability distribution $\mu \in \mathcal{F}$ such that the minimum $\mathbf{w}^*$ of $\overline{f}$ is up to symmetry unique, then for all $m \geq m(\epsilon, \delta)$ we have*

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

For dimension $d = 1$, we study a simpler, but related problem where the probability distribution μ is defined over the unit interval $[0, 1]$ and has continuous density function f. We assume that f attains unique minimum at x^*. Given an i.i.d. sample S from μ, the task of the algorithm L is to output $x \in [0, 1]$ "close" to x^*. To measure closeness, we naturally (re)define distance measures D_E, D_f, D_μ as follows $D_E(x, x') = |x - x'|$, $D_f(x, x') = |f(x) - f(x')|$ and $D_\mu(x, x') = \mu((-\infty, x]\Delta(-\infty, x'])$.

3 The One Dimensional Problem

We consider two natural learning algorithms for the one-dimensional problem. 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 mid-point of the largest gap between two consecutive points the sample. We show its consistency in Section 3.2.

Let $\mathcal{F}_1$ be the family of all probability distributions over the unit interval $[0, 1]$ that have continuous density function. In Section 5 we show there are no algorithms that are $\mathcal{F}_1$-uniformly convergent.

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.

Theorem 3. *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.*

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 $m \geq m_0$

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

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

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

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

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

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

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 (see Anthony and Bartlett, 1999), we have that with probability at least $1 - \delta$ over i.i.d. draws of sample S of size m, for any bucket I,

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

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. □

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 4. *If the number of buckets $k(m) = \omega(m/\log m)$, then B_k is not consistent.*

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

Lemma 5 (The Coupon Collector Problem (Motwani and Raghavan, 1995))**.** *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 4. 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 5, 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 5 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

The *hard-margin* algorithm outputs the mid-point of the largest interval between the adjacent sample points. 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 6. *The hard-margin algorithm is consistent.*

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 (1939). However, we will present a direct and much simpler proof.

Lemma 7. *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 5, 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 5 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 7 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 The High Dimensional Problem

In this section we consider the lowest-density-hyperplane problem for dimension $d\geq 2$. Recall that the task is to find a unit normal vector $\mathbf{w}$ of a homogeneous hyperplane $\mathbf{w}^T\mathbf{x}=0$ such that the "density on hyperplane" $\overline{f}(\mathbf{w})$ is as small as possible. We show that there exists a learning algorithm that is consistent.

We 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 $\mathbf{w}\in\mathcal{S}^{d-1}$ we consider the "γ-strip"

$$h(\mathbf{w},\gamma) = \{\mathbf{x}\in\mathbb{R}^d \;:\; |\mathbf{w}^T\mathbf{x}|\leq\gamma/2\}$$

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 8. *If $\gamma(m)=\omega(1/\sqrt{m})$ and $\gamma(m)\to 0$ as $m\to\infty$, then H_γ is consistent.*

Proof. The structure of the proof is similar to the proof of Theorem 3. However, we will need more technical tools.

Fix the probability measure μ and (hence also f). Note that for any $\mathbf{w}\in\mathcal{S}^{d-1}$,

$$\lim_{m\to\infty}\frac{\mu(h(\mathbf{w},\gamma(m)))}{\gamma(m)} = \overline{f}(\mathbf{w})\;.$$

In other words, the sequence of functions $\{\mu(h(\cdot,\gamma(m)))/\gamma(m):\mathcal{S}^{d-1}\to\mathbb{R}_0^+\}_{m=1}^\infty$, converges point-wise to the function $\overline{f}:\mathcal{S}^{d-1}\to\mathbb{R}_0^+$. Note that $\mu(h(\cdot,\gamma(m)))/\gamma(m):\mathcal{S}^{d-1}\to\mathbb{R}_0^+$ is continuous for any m and $\mathcal{S}^{d-1}$ is compact. Therefore the sequence $\{\mu(h(\cdot,\gamma(m)))/\gamma(m)\}_{m=1}^\infty$ 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 $\mathbf{w}\in\mathcal{S}^{d-1}$,

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

Fix $\epsilon,\delta>0$. 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 $m\geq m_0$,

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

$$\forall\mathbf{w}\in\mathcal{S}^{d-1}\quad \left|\frac{\mu(h(\mathbf{w},\gamma(m)))}{\gamma(m)} - \overline{f}(\mathbf{w})\right| < \eta/3 \quad (10)$$

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

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

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

$$\mu(h(\mathbf{w},\gamma(m))) > \mu(h(\mathbf{w}^*,\gamma(m))) + \eta\gamma(m)/3\;. \quad (11)$$

From the well known Vapnik-Chervonenkis bounds (see Anthony and Bartlett, 1999), we have that with probability at least $1-\delta$ over i.i.d. draws of S of size m we have that for any $\mathbf{w}$,

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

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 $\mathbf{w} \in \mathcal{S}^{d-1} \setminus U$,

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

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. □

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 9. *No learning algorithm is $\mathcal{F}_1$-uniformly convergent w.r.t. any of the distance measures 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 $3/4$, no algorithm is $\mathcal{F}_1$-uniformly convergent 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. $f(x) = g(1-x)$.

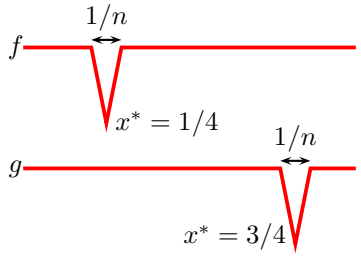

Figure 1: f is uniform everywhere except a small neighbourhood around $1/4$ where it has a sharp 'v' shape. And g is the reflection of f about $x = 1/2$.

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 one-dimensional 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

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

Shai Ben-David and Michael Lindenbaum. Learning distributions by their density levels: A paradigm for learning without a teacher. *Journal of Computer and System Sciences*, 55(1):171–182, 1997.

Shai Ben-David and Ulrike von Luxburg. Relating clustering stability to properties of cluster boundaries. In *Prooceedings of Conference on Learning Theory (COLT)*, 2008.

Shai Ben-David, Nadav Eiron, and Hans-Ulrich Simon. The computational complexity of densest region detection. *Journal of Computer and System Sciences*, 64(1):22–47, 2002.

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Prooceedings of Conference on Learning Theory (COLT)*, 2008.

Olivier Chapelle, Bernard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

Luc Devroye and Gábor Lugosi, editors. *Combinatorial Methods in Density Estimation*. Springer, 2001.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 200–209, 1999.

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

Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

Ohad Shamir and Naftali Tishby. Model selection and stability in k-means clustering. In *Prooceedings of Conference on Learning Theory (COLT)*, 2008.

Aarti Singh, Clayton Scott, and Robert Nowak. Adaptive hausdorff estimation of density level sets. In *Prooceedings of Conference on Learning Theory (COLT)*, 2008.

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