LARGE-WIDTH BOUNDS FOR LEARNING HALF-SPACES ON DISTANCE SPACES

MARTIN ANTHONY, JOEL RATSABY

Abstract. A half-space over a distance space is a generalization of a half-space in a vector space. An important advantage of a distance space over a metric space is that the triangle inequality need not be satisfied, which makes our results potentially very useful in practice. Given two points in a set, a half-space is defined by them, as the set of all points closer to the first point than to the second. In this paper we consider the problem of learning half-spaces in any finite distance space, that is, any finite set equipped with a distance function. We make use of a notion of 'width' of a half-space at a given point: this is defined as the difference between the distances of the point to the two points that define the half-space. We obtain probabilistic bounds on the generalization error when learning half-spaces from samples. These bounds depend on the empirical error (the fraction of sample points on which the half-space does not achieve a large width) and on the VC-dimension of the effective class of half-spaces that have a large sample width. Unlike some previous work on learning classification over metric spaces, the bound does not involve the covering number of the space, and can therefore be tighter.

keywords: Large width learning, distance and metric spaces, half spaces, pseudo rank, margin

1. Introduction

In [3], we obtained generalization error bounds for learning binary classifiers on a finite metric space $\mathcal{X}$ using the class of all binary functions on $\mathcal{X}$; and [6] obtained error bounds for multi-category classification on infinite metric spaces. In both papers, the bounds involved the covering number of the metric space, which in general is not known or not easy to compute, though can be approximated numerically.

Date: March 8, 2018.

In the current paper we consider learning binary classification on finite distance spaces, that is, finite sets equipped with a distance function (often called 'dissimilarity measure' [10]) where the classifiers are "half-spaces". An important advantage of a distance space over a metric space is that the triangle inequality need not be satisfied, which makes our results potentially very useful in practice. Our definition of distance function is quite loose in that it does *not* need to satisfy any of the non-negativity, symmetry or reflexivity properties of a proper distance function [10]. We still call it a distance because, as far as we can expect in applying our learning results, any useful space has at least the non-negativity property.

Since the distance space is not necessarily equipped with an inner product, by a half-space we do not mean the usual linear half-space that is defined in a Euclidean space, but, rather, a more general definition which is based on the distances to two points in the space.

The standard large-margin results for learning large-margin classifiers [9, 1] by thresholding real-valued functions have error bounds that depend on the covering numbers of function classes. Interestingly, and in contrast, while we also threshold real-valued functions (which we call 'width'-functions), we are able here to provide bounds that do not involve covering numbers, neither of the class of width functions nor of the underlying distance space itself. Depending on some characteristics of the distance space (which involve the positive components a matrix-based representation of the class of half spaces on a distance space) the upper bounds can be tighter and hence more useful in practice.

There have been several works on large-margin learning over metric spaces. In [12], the ubiquitous SVM approach is chosen and they embed the metric space in a Banach or Hilbert space followed by learning linear classifiers on this space. They provide upper bounds on the Rademacher averages, which can be used to obtain generalization error bounds on learning [7]. The bounds involve the covering number of the metric space. In [3], the problem of learning the class of all binary classifiers over finite metric spaces is considered. The learning error bounds involve covering numbers of the metric space. In [6, 4], multi-category classification over metric spaces is considered and the error bounds also involve the covering number of the metric space. Other work on learning over metric spaces includes [11] (see also references within) which considers learning nearest-neighbor classifiers in semi-metric spaces using compression schemes which involves bounds on packing numbers by exponentials in the density-dimension of the space.

In the current paper, we offer an alternative approach to learning 'linear' classifiers on metric spaces. (In addition, the present paper addresses the more general setting of a distance space.) In contrast to [12], which also deals with learning linear classifiers on a metric space, we do not restrict to learning via SVM and, being focused on learning half-spaces, we are able to take advantage of the simpler hypothesis space (in comparison to that of [3]) and obtain an error-bound that does not involve a covering number of the underlying metric space (as in [3]) or any special dimensions, such as fat-shattering or doubling dimension of the space (as in [11]) that may be hard to estimate. Instead we use a new characterization of the class of half spaces on a distance space which is based on positive entries of a matrix that can be automatically computed directly from the distance space. Also, the fact that we deal with a distance space, rather than a metric space, means that the triangle inequality need not be satisfied. This and the fact that the error bound is computable directly from these positive entries make our result widely applicable. Also, as mentioned above, our use of the concept of distance is loose, so that, for instance, not only is it the case that the triangle inequality need not be satisfied, but also none of the three standard properties of a distance need to be satisfied either. From a theoretical perspective, our analysis is less involved than in the above works. In particular, we do not need to introduce any new dimensions or use covering (or packing) bounds. Instead, we analyze the hypothesis-class complexity directly by elementary combinatorics. Hence we believe that the paper can also serve as a reference point, useful for the theorist, to compare other approaches for computing error bounds on abstract spaces.

2. Setup

2.1. **The classifiers.** We consider a finite distance space $\mathcal{X} := \{x_1, \ldots, x_N\}$ with distance d, a function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$, and a binary set $\mathcal{Y} = \{-1, 1\}$ of possible classifications of the points of the distance space. Let us assume that the distance is normalized such that

$$\operatorname{diam}(\mathcal{X}) := \max_{1 \le i,j \le N} d(x_i, x_j) = 1.$$

A *prototype* $p \in \mathcal{X}$ is a point in the distance space that has an associated label $\sigma \in \mathcal{Y}$. We denote by p^+, $p^- \in \mathcal{X}$, prototypes whose labels are 1 and -1, respectively. When the label of a prototype is not explicitly mentioned, we write p.

Let

$$\Pi := \big\{[p^+, 1], [p^-, -1]\big\} \subset \mathcal{X} \times \mathcal{Y}$$

be a pair of two oppositely labeled prototypes (together with their labels). We denote by h_Π a classifier which is defined as follows: given $x \in \mathcal{X}$,

$$h_\Pi(x) := h_{\Pi,\sigma}(x) = \begin{cases} 1 & \text{if } d(x,p^+) < d(x,p^-) \\ -1 & \text{otherwise.} \end{cases} \tag{2.1}$$

For $s \in \mathcal{Y}$, the set $\{x : h_\Pi(x) = s\}$ is referred to as a *half space* since it generalizes the special case of a half-space when the distance space is a vector space; for instance, when the space is a vector space equipped with an inner product (such is commonly considered in learning with Support Vector Machines), a half-space can be specified not only based on two points as in (2.1) but also based on the inner product of a parameter vector (the vector difference between these two points) and the point to be classified.

2.2. **Width and error bounds.** We work in the framework of the popular 'PAC' model of computational learning theory (see [13, 8]). We denote by $\{(X_j, Y_j)\}_{j=1}^m$ a random sample which consists of i.i.d. pairs (X_j, Y_j) , $X_j \in \mathcal{X}$, and $Y_j \in \mathcal{Y}$, $1 \leq j \leq m$, each distributed according to any fixed probability distribution $P_{X,Y}$, which is not assumed to be known. We denote by

$$\xi = \{(x_j, y_j)\}_{j=1}^m$$

a realization of the random labeled sample. Learning is conducted on the basis of ξ.

In the PAC framework, a typical result would state that, for all $\delta \in (0,1)$, with probability at least $1-\delta$, for all classifiers h_Π, the generalization error of h_Π is bounded from above by the empirical classification error (based on the sample ξ) plus some added deviation ϵ which decreases in m and δ.

A prototype p may be any point in $\mathcal{X}$; in particular, it could be one that depends on the sample ξ directly or via some learning algorithm.

We define the *width* of h_Π at a point $x \in \mathcal{X}$ as follows,

$$w_{h_{\Pi,\sigma}}(x) := \max\left\{d\left(x,p^+\right), d\left(x,p^-\right)\right\} - \min\left\{d\left(x,p^+\right), d\left(x,p^-\right)\right\}. \tag{2.2}$$

Since x is classified by h_Π according to the label of the closer of the two prototypes, then the width of h at x is the difference between the distance to the nearest-unlike-prototype of x (this is the farther prototype) and the distance to the nearest-prototype to x. (Here, unlike means of a different classification by h_Π than x.)

The corresponding *signed width* (or margin) function is defined as

$$f_\Pi(x) := f_{\Pi,\sigma}(x) = f_{h_\Pi}(x) = h_\Pi(x) w_{h_\Pi}(x). \tag{2.3}$$

Note that for x equidistant from two oppositely labeled prototypes p^+, $p^- \in \Pi$, the value of the margin $f_{\Pi,\sigma}(x)$ at this x is zero. This definition is intuitive. From (2.2) and (2.3) it follows that

$$f_\Pi(x) = d(x, p^-) - d(x, p^+).$$

(This measure of signed width was introduced in [2].) Henceforth, let $\gamma > 0$ be a width parameter and let $\Pi \subset \mathcal{X}$ be any set of two points in $\mathcal{X}$.

In this paper we consider the problem of learning half-spaces on $\mathcal{X}$ that have a large width on the sample ξ, or more generally, half-spaces that have a large width on a high proportion of the sample points. We obtain a bound on the classification error of such classifiers with dependence on this proportion and on the width parameter value. Define the function

$$\operatorname{sgn}(a) := \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a \leq 0. \end{cases}$$

For the purpose of bounding the generalization error it is convenient to express the classification $h_\Pi(X)$ in terms of the signed width as follows,

$$h_\Pi(X) = \operatorname{sgn}\left(f_\Pi(X)\right).$$

Therefore the generalization error $\operatorname{er}_P(h_\Pi)$ can be bounded as follows

$$\begin{aligned} \operatorname{er}_P(h_\Pi) &= P\left(h_\Pi(X) \neq Y\right) && (2.4) \\ &= P\left(Y f_\Pi(X) < 0\right) + P\left(Y = 1, f_\Pi(X) = 0\right) \\ &\leq P\left(Y f_\Pi(X) \leq 0\right). && (2.5) \end{aligned}$$

In order to obtain an upper bound on the learning error we are interested in the probability of the 'bad event' that there is some γ and some Π for which the generalization error of h_Π is larger than the empirical measure of that error by some deviation ϵ. That is, we want to upper bound the following probability

$$P_{X,Y}^m \left(\left\{ \xi : \exists\gamma, \exists\Pi, P\left(Y f_\Pi(X) \leq 0\right) > \frac{1}{m} \sum_{j=1}^m \mathbb{I}\left\{Y_j f_\Pi(X_j) \leq \gamma\right\} + \epsilon \right\} \right). \tag{2.6}$$

In the PAC framework, this probability is bounded by δ provided that we choose ϵ to be a function of m, γ and δ. For simpler notation, we sometimes keep the dependence of ϵ on these parameters implicit or express it in terms of one of these parameters when deemed important in the context.

This can be expressed as the following probability:

$$P_{X,Y}^m\left(\left\{\xi:\ \exists\gamma,\ \exists\Pi,\ P\left(Yf_\Pi(X)>0\right)<\frac{1}{m}\sum_{j=1}^m\mathbb{I}\left\{Y_jf_\Pi(X_j)>\gamma\right\}-\epsilon\right\}\right). \tag{2.7}$$

Let us fix γ for now, and deal with bounding the following probability:

$$P_{X,Y}^m\left(\left\{\xi:\ \exists\Pi,\ P\left(Yf_\Pi(X)>0\right)<\frac{1}{m}\sum_{j=1}^m\mathbb{I}\left\{y_jf_\Pi(x_j)>\gamma\right\}-\epsilon\right\}\right). \tag{2.8}$$

3. Representing the bad event using sets

3.1. **Related sets.** Define the set $M_{\Pi,\gamma}\subset\mathcal{X}\times\mathcal{Y}$ as follows,

$$M_{\Pi,\gamma}:=\{(x,y):\ yf_\Pi(x)>\gamma\}$$

and let

$$M_{\Pi,\gamma}^+:=\{(x,1):\ f_\Pi(x)>\gamma\} \tag{3.1}$$

$$M_{\Pi,\gamma}^-:=\{(x,-1):\ f_\Pi(x)<-\gamma\}. \tag{3.2}$$

Note that

$$\begin{aligned}M_{\Pi,\gamma}&=\left\{M_{\Pi,\gamma}\bigcap\{(x,y):y=1\}\right\}\bigcup\left\{M_{\Pi,\gamma}\bigcap\{(x,y):y=-1\}\right\}\\&=M_{\Pi,\gamma}^+\bigcup M_{\Pi,\gamma}^-.\end{aligned} \tag{3.3}$$

We can see that

$$P\left(Yf_\Pi(X)>\gamma\right)=P\left(M_{\Pi,\gamma}\right)$$

and

$$\frac{1}{m}\sum_{j=1}^m\mathbb{I}\left\{Y_jf_\Pi(X_j)>\gamma\right\}=P_m\left(M_{\Pi,\gamma}\right)$$

where P_m denotes the empirical measure based on a random sample of cardinality m drawn i.i.d. according to P^m. Thus (2.7) is expressed as,

$$P_{X,Y}^m\left(\{\xi:\ \exists\gamma,\ \exists\Pi,\ P\left(M_{\Pi,0}\right)<P_m\left(M_{\Pi,\gamma}\right)-\epsilon\}\right).$$

Let $\epsilon = \epsilon(\gamma)$ depend on γ (in a way to be specified later) and define the set $E(\gamma) \subseteq (\mathcal{X} \times \mathcal{Y})^m$ as

$$E(\gamma) := \{\xi : \exists \Pi,\, P\left(M_{\Pi,0}\right) < P_m\left(M_{\Pi,\gamma}\right) - \epsilon(\gamma)\}. \tag{3.4}$$

Then substituting $\epsilon(\gamma)$ for ϵ in (2.8) will mean that (2.8) equals the probability $P^m_{X,Y}\left(E(\gamma)\right)$. It follows that (2.6) will then be bounded from above by

$$P^m_{X,Y}\left(\bigcup_{\gamma\in(0,\mathrm{diam}(\mathcal{X})]} E(\gamma)\right). \tag{3.5}$$

Let L be an integer (to be specified below). For integers $0 \le l \le L+1$, let η_l be a decreasing sequence such that the following conditions hold:

(1) $0 \le \eta_l \le 1$
(2) $\eta_0 = 1$, $\eta_{L+1} = 0$.

Let

$$C := \sum_{l=1}^{L} \eta_l.$$

While all the above quantities L, η_l and C may depend on $\mathcal{X}$, we keep this dependence implicit in the notation.

Define $\Gamma_l := (\eta_l, \eta_{l-1}]$ for $1 \le l \le L$ and $\Gamma_{L+1} := [0, \eta_L]$. Then (3.5) equals

$$P^m_{X,Y}\left(\bigcup_{l=1}^{L+1}\bigcup_{\gamma\in\Gamma_l} E(\gamma)\right) \le \sum_{l=1}^{L+1} P^m_{X,Y}\left(\bigcup_{\gamma\in\Gamma_l} E(\gamma)\right). \tag{3.6}$$

Define the set $E_l \subseteq (\mathcal{X} \times \mathcal{Y})^m$ as

$$E_l := \{\xi : \exists \Pi,\, P\left(M_{\Pi,\eta_l}\right) < P_m\left(M_{\Pi,\eta_l}\right) - \epsilon(\eta_{l-1})\}.$$

Henceforth, assume that $\epsilon(\gamma)$ is a non-increasing function over each interval Γ_l.

Claim 1. For any $\gamma \in \Gamma_l$, $E(\gamma) \subseteq E_l$.

Proof. We have $M_{\Pi,0} \supseteq M_{\Pi,\eta_l}$ thus $P\left(M_{\Pi,0}\right) \ge P\left(M_{\Pi,\eta_l}\right)$. Also, $M_{\Pi,\eta_l} \supseteq M_{\Pi,\gamma}$ since $\eta_l \le \gamma$, and so $P_m\left(M_{\Pi,\gamma}\right) \le P_m\left(M_{\Pi,\eta_l}\right)$. For $\gamma \le \eta_{l-1}$, by the above assumption on ϵ, $\epsilon(\gamma) \ge \epsilon(\eta_{l-1})$. It follows that $E(\gamma) \subseteq E_l$. □

We therefore have that (3.6) is bounded from above as follows:

$$P_{X,Y}^m\left(\bigcup_{l=1}^{L+1}\bigcup_{\gamma\in\Gamma_l}E(\gamma)\right)\leq\sum_{l=1}^{L+1}P_{X,Y}^m\left(\bigcup_{\gamma\in\Gamma_l}E(\gamma)\right)\leq\sum_{l=1}^{L+1}P_{X,Y}^m\left(E_l\right).$$

The event that there exists Π such that $P\left(M_{\Pi,\eta_l}\right)<P_m\left(M_{\Pi,\eta_l}\right)-\epsilon(\eta_{l-1})$, by (3.3), implies that either of the following events occurs: there exists Π such that

$$P\left(M_{\Pi,\eta_l}^+\right)<P_m\left(M_{\Pi,\eta_l}^+\right)-\epsilon(\eta_{l-1})/2$$

or there exists a Π such that

$$P\left(M_{\Pi,\eta_l}^-\right)<P_m\left(M_{\Pi,\eta_l}^-\right)-\epsilon(\eta_{l-1})/2.$$

Let

$$E_l^+:=\left\{\xi:\exists\Pi,\,P\left(M_{\Pi,\eta_l}^+\right)<P_m\left(M_{\Pi,\eta_l}^+\right)-\epsilon(\eta_{l-1})/2\right\}\tag{3.7}$$

and

$$E_l^-:=\left\{\xi:\exists\Pi,\,P\left(M_{\Pi,\eta_l}^-\right)<P_m\left(M_{\Pi,\eta_l}^-\right)-\epsilon(\eta_{l-1})/2\right\}.\tag{3.8}$$

Then

$$P_{X,Y}^m\left(E_l\right)\leq P_{X,Y}^m\left(E_l^+\right)+P_{X,Y}^m\left(E_l^-\right).$$

3.2. **Bound on the probability.** We now aim to bound from above the first probability $P_{X,Y}^m\left(E_l^+\right)$.

We briefly first recall the definitions of growth function and VC-dimension [14]. Suppose that $\mathcal{C}$ is a collection of subsets of a set Z. Then the *growth function* of $\mathcal{C}$ is the function $\Pi_{\mathcal{C}}:\mathbb{N}\to\mathbb{N}$ defined as follows: for $m\in\mathbb{N}$,

$$\Pi_{\mathcal{C}}(m)=\max\{\Pi_{\mathcal{C}}(S):S\subseteq Z,|S|=m\},$$

where

$$\Pi_{\mathcal{C}}(S)=|\{C\cap S:C\in\mathcal{C}\}|.$$

The *VC-dimension* of $\mathcal{C}$ is (infinity, or) the largest value of m such that $\Pi_{\mathcal{C}}(m)=2^m$. (A set S of size m such that $\Pi_{\mathcal{C}}(S)=2^m$ is said to be *shattered* by $\mathcal{C}$.)

Using pairs Π of oppositely labeled prototypes, we define the following classes:

$$\mathcal{M}_{\eta l}^+:=\left\{M_{\Pi,\eta_l}^+:\Pi\subset\mathcal{X}\times\mathcal{Y}\right\},\ \mathcal{M}_{\eta_l}^-:=\left\{M_{\Pi,\eta_l}^-:\Pi\subset\mathcal{X}\times\mathcal{Y}\right\}.$$

We choose for $\epsilon(\gamma)$ in (3.4) the following expression,

$$\epsilon(\gamma):=\sqrt{\frac{32}{m}\left(\mathsf{d}(\gamma)\ln\left(\frac{2em}{\mathsf{d}(\gamma)}\right)+\ln\left(\frac{8(C+1)}{\gamma\delta}\right)\right)}\tag{3.9}$$

where $\mathsf{d}(\gamma)$ is a function which is piecewise constant over the intervals Γ_l; that is, for $\gamma \in \Gamma_l$, $\mathsf{d}(\gamma) = \mathsf{d}(\eta_{l-1})$ (where a specific $\mathsf{d}(\gamma)$ is chosen later). Hence the inequality $\epsilon(\gamma) \geq \epsilon(\eta_{l-1})$ holds for $\gamma \in \Gamma_l$, as required for Claim 1 and for the definition of $\epsilon(\gamma)$ in (3.4).

Remark 2. We actually choose d such that, in addition to the above piecewise constant property, $\mathsf{d}(\eta_l)$ is an upper bound on the VC-dimension of $\mathcal{M}^+_{\eta_l}$ and $\mathcal{M}^-_{\eta_l}$. That is, at the points $\gamma = \eta_l$, $1 \leq l \leq L$, $\mathsf{d}(\eta_l)$ bounds from above the VC dimension of $\mathcal{M}^+_{\eta_l}$ and $\mathcal{M}^-_{\eta_l}$.

Denote by $\Pi_{\mathcal{M}^+_{\eta_l}}(m)$ the growth function of the class $\mathcal{M}^+_{\eta_l}$. From [8] (see also Theorem 3.7 of [1]), for integer $m \geq \mathsf{d}(\eta_l)$,

$$\Pi_{\mathcal{M}^+_{\eta_l}}(m) \leq \left(\frac{em}{\mathsf{d}(\eta_l)}\right)^{\mathsf{d}(\eta_l)}. \tag{3.10}$$

By [14] (see also Theorem 4.3 of [1]), it follows that

$$\begin{aligned} P^m_{X,Y}\left(E^+_l\right) &\leq 4\Pi_{\mathcal{M}^+_{\eta_l}}(2m)\exp\left(-m\epsilon^2/32\right) \\ &\leq 4\left(\frac{2em}{\mathsf{d}(\eta_l)}\right)^{\mathsf{d}(\eta_l)}\exp\left(-m\epsilon^2/32\right). \end{aligned}$$

Substituting η_{l-1} for γ in (3.9) and letting (3.9) be the choice for $\epsilon(\eta_{l-1})$ in (3.7), then from Theorems 3.7, 4.3 of [1], it follows that both $P\left(E^+_l\right)$ and $P(E^-_l)$ are bounded from above by $\eta_{l-1}\delta/2(C+1)$. Then from Claim 1, it follows that (3.6) is bounded from above by

$$\begin{aligned}
\sum_{l=1}^{L+1} P_{X,Y}^m\left(E_l\right) &\leq \sum_{l=1}^{L+1} P_{X,Y}^m\left(E_l^+\right) + \sum_{l=1}^{L+1} P_{X,Y}^m\left(E_l^-\right) && (3.11)\\
&\leq 2\left(\frac{\delta}{2(C+1)}\right)\sum_{l=1}^{L+1}\eta_{l-1}\\
&= \frac{\delta}{C+1}\left(\sum_{l=1}^{L+1}\eta_{l-1}\right)\\
&= \frac{\delta}{C+1}\left(\sum_{l=0}^{L}\eta_{l}\right)\\
&= \frac{\delta}{C+1}\left(\sum_{l=1}^{L}\eta_{l}+1\right)\\
&= \delta. && (3.12)
\end{aligned}$$

In the next section we choose $\mathsf{d}(\gamma)$.

4. Bounding the VC-dimension

In this section we bound the VC dimension of the class $\mathcal{M}_\gamma^+$ and $\mathcal{M}_\gamma^-$. We first introduce some additional notation.

4.1. Half-spaces of $\mathcal{X}$ and matrix representation.

We bound the VC dimension of the class $\mathcal{M}_\gamma^+$ of sets in $\mathcal{X}$. For any z, $z' \in \mathcal{X}$, define the set

$$W_\gamma^{(z,z')} := \{x : d(x,z') - d(x,z) > \gamma\} \tag{4.1}$$

and let the class of such sets be defined as

$$\mathcal{W}_\gamma := \left\{W_\gamma^{(z,z')} : z, z' \in \mathcal{X},\, z \neq z'\right\}.$$

For any pair Π of oppositely labeled prototypes p^+, $p^- \in \mathcal{X}$ we have

$$M_{\Pi,\gamma}^+ = W_\gamma^{(p^+,p^-)} \times \{1\} \tag{4.2}$$

which follows from the fact that the statement $(x,1) \in M_{\Pi,\gamma}^+$ means $h(x) = 1$, and $f_\Pi(x) = d(x,p^-) - d(x,p^+) > \gamma$, and this means $x \in W_\gamma^{(p^+,p^-)}$.

We have

$$M^-_{\Pi,\gamma} = W^{(p^-,p^+)}_\gamma \times \{-1\} \tag{4.3}$$

because the statement $(x,-1) \in M^-_{\Pi,\gamma}$ means $h(x) = -1$, which means $f_\Pi(x) = -\left(d(x,p^+) - d(x,p^-)\right)$, and by definition of $M^-_{\Pi,\gamma}$ we have $f_\Pi(x) < -\gamma$, therefore $d(x,p^+) - d(x,p^-) > \gamma$. This is precisely the definition of the set $W^{(p^-,p^+)}_\gamma$. It follows that in order to indicate if $(x,-1) \in M^-_{\Pi,\gamma}$ it suffices to indicate if $x \in W^{(p^-,p^+)}_\gamma$.

From (4.2),(4.3) it follows that

$$\mathcal{M}^+_\gamma \subseteq \mathcal{W}_\gamma \times \{1\},\ \mathcal{M}^-_\gamma \subseteq \mathcal{W}_\gamma \times \{-1\} \tag{4.4}$$

and we aim to bound the VC-dimension of $\mathcal{W}_\gamma$ in order to bound the VC-dimension of each of $\mathcal{M}^+_\gamma$, $\mathcal{M}^-_\gamma$. (This will work because (4.4) implies that $VC(\mathcal{M}^+_\gamma) \leq VC(\mathcal{W}_\gamma \times \{1\}) = VC(\mathcal{W}_\gamma)$ and similarly for $\mathcal{M}^-_\gamma$.)

Recall that $\mathcal{X} = \{x_1, \ldots, x_N\}$. Since a prototype may be any point in $\mathcal{X}$ then, in general, for any pair of prototypes p, $q \in \mathcal{X}$ there is some $1 \leq i \neq j \leq N$, such that $p = x_i$, $q = x_j$. We write $W_0^{(p,q)}$ as $W_0^{(i,j)}$. Then $W_0^{(i,j)}$ corresponds to the positive elements of the following vector:

$$f^{(i)}_j := \begin{bmatrix} d(x_1,x_j) - d(x_1,x_i) \\ \vdots \\ d(x_N,x_j) - d(x_N,x_i) \end{bmatrix}. \tag{4.5}$$

Note that taking the sign of a vector $f^{(i)}_j$ yields a partition of $\mathcal{X}$ into two parts, which, as mentioned above, is referred to as a half-space.

For a real vector $v \in \mathbb{R}^N$, let $\operatorname{sgn}(v) = [\operatorname{sgn}(v_1), \ldots, \operatorname{sgn}(v_N)]^T$. Hence $\operatorname{sgn}(f^{(i)}_j)$ corresponds to a half-space on $\mathcal{X}$.

Fix any point $x_i \in \mathcal{X}$ and let the $N \times (N-1)$ matrix F_i be defined by

$$F^{(i)} = \left[f^{(i)}_1, \ldots, f^{(i)}_{i-1}, f^{(i)}_{i+1}, \ldots, f^{(i)}_N \right], \tag{4.6}$$

with columns $f^{(i)}_j$, $j \neq i$, $1 \leq j \leq N$.

Define the $N \times N(N-1)$ matrix

$$F := \left[F^{(1)}, \ldots, F^{(N)} \right]. \tag{4.7}$$

The binary matrix

$$\operatorname{sgn}(F) := \left[\operatorname{sgn}(F^{(1)}), \ldots, \operatorname{sgn}(F^{(N)}) \right],$$

where

$$\operatorname{sgn}(F^{(i)}) := \left[\operatorname{sgn}(f_1^{(i)}), \ldots, \operatorname{sgn}(f_N^{(i)})\right],$$

represents the class of all half-spaces on $\mathcal{X}$.

4.2. **Thresholding by γ.** The set $W_0^{(i,j)}$ corresponds to some column of the matrix F. We now define a more general matrix whose columns correspond to the sets $W_\gamma^{(i,j)}$ defined in (4.1), for any fixed $\gamma \geq 0$. Bounding the VC dimension of this matrix means that we obtain a bound on the VC dimension of the class $\mathcal{W}_\gamma$.

For any $1 \leq i \neq j \leq N$, a set $W_\gamma^{(i,j)}$ corresponds to the positive elements of the vector $f_j^{(i)} - \gamma\mathbf{1}$ where $\mathbf{1}$ is an $N \times 1$ vector of all ones. Denote by J an $N \times N(N-1)$ matrix of all ones. For any $\gamma > 0$, let us consider the $N \times N(N-1)$ matrix

$$F_\gamma := F - \gamma J = \left[f_2^{(1)} - \gamma\mathbf{1}, \ldots, f_{N-1}^{(N)} - \gamma\mathbf{1}\right]. \tag{4.8}$$

The matrix F_γ corresponds to the class $\mathcal{W}_\gamma$ of sets, where for column $f_j^{(i)} - \gamma\mathbf{1}$, the positive elements of the vector correspond to the elements of the set $W_\gamma^{(i,j)}$.

The binary matrix $\operatorname{sgn}(F_\gamma)$ is a class of 'affined' half-spaces on $\mathcal{X}$ (the columns of $\operatorname{sgn}(F_\gamma)$). We aim to bound the VC-dimension of $\operatorname{sgn}(F_\gamma)$.

4.3. **Pseudo-rank.** Let A be an $m \times n$ matrix and define the set of columns of A to be $\operatorname{col}(A)$. Define by the *pseudo-rank* of A, denoted by $\rho(A)$, the number of distinct sign-columns of the matrix $\operatorname{sgn}(A)$, that is,

$$\rho(A) := \left|\{\operatorname{sgn}(u) : u \in \operatorname{col}(A)\}\right|.$$

Let

$$\kappa := \rho(F) = \rho(F_0) \leq N(N-1). \tag{4.9}$$

Consider the matrix F_γ. Clearly, we have the following:

Claim 3. For any $1 \leq r \leq \kappa$, there exists some γ such that $\rho(F_\gamma) \geq r$.

This holds since, at least at $\gamma = 0$, $\rho(F_\gamma) \geq \kappa \geq r \geq 1$.

Remark 4. When $\gamma > 1$, $\rho(F_\gamma) = 1$ because all of the columns of $\operatorname{sgn}(F_\gamma)$ equal the all-(-1) binary vector. Thus $\rho(F_\gamma)$ starts at κ and eventually becomes 1; however, in general the function $\rho(F_\gamma)$ is not necessarily decreasing monotonically nor even non-increasing, as, for instance, the matrix $\operatorname{sgn}(F_\gamma)$ can have a set of identical columns

and when γ increases some of these columns may become distinct, which results in a local increase of $\rho(F_\gamma)$.

Also, the following trivial upper bound holds for every $\gamma \geq 0$,

$$\rho(F_\gamma) \leq \kappa \leq N(N-1). \tag{4.10}$$

Define by the *characteristic points* of F the following γ values:

$$\gamma_r = \sup\left\{\gamma : \rho\left(F_\gamma\right) \geq r, \gamma \in [0,1]\right\}, \, 1 \leq r \leq \kappa, \tag{4.11}$$

where this definition is well-posed, since from Claim 3 it follows that the set over which the supremum is taken is non-empty and bounded above.

The points γ_r characterize the matrix F since they express how a 'perturbed' version of F behaves in terms of how many distinct binary vectors are in the matrix $\operatorname{sgn}(F_\gamma)$. For $r \geq r'$, we have $\{\gamma : \rho\left(F_\gamma\right) \geq r\} \subseteq \{\gamma : \rho\left(F_\gamma\right) \geq r'\}$, and hence $\sup\{\gamma : \rho\left(F_\gamma\right) \geq r\} \leq \sup\{\gamma : \rho\left(F_\gamma\right) \geq r'\}$. Therefore, it follows that the points satisfy

$$\gamma_\kappa \leq \gamma_{\kappa-1} \leq \cdots \leq \gamma_1, \tag{4.12}$$

and we refer to γ_i as the i^{th} characteristic point of F.

Note that in (4.12) the inequalities are not necessarily strict because there may be multiple columns with the same positive elements so that when γ reaches their value, the sign of these columns is the same and therefore there are fewer distinct binary vectors. The next example depicts this. Denote by '$-$' any negative real value, and let $\alpha > 2\beta > 0$. Consider the matrix F_γ at $\gamma = 0$ to be

$$F_0 := \begin{pmatrix} \alpha & \alpha & - & - \\ - & \beta & - & - \\ - & - & \alpha & \alpha \\ - & - & - & \beta \end{pmatrix}.$$

At $\gamma = \beta$ the matrix is

$$F_\beta := \begin{pmatrix} (\alpha-\beta) & (\alpha-\beta) & - & - \\ - & 0 & - & - \\ - & - & (\alpha-\beta) & (\alpha-\beta) \\ - & - & - & 0 \end{pmatrix},$$

so that

$$\operatorname{sgn}(F_\beta) := \begin{pmatrix} + & + & - & - \\ - & - & - & - \\ - & - & + & + \\ - & - & - & - \end{pmatrix}.$$

For any small $0 < \epsilon < \beta$ the matrix $F_{\beta-\epsilon}$ is as follows

$$F_{\beta-\epsilon} := \begin{pmatrix} (\alpha-\beta+\epsilon) & (\alpha-\beta+\epsilon) & - & - \\ - & \epsilon & - & - \\ - & - & (\alpha-\beta+\epsilon) & (\alpha-\beta+\epsilon) \\ - & - & - & \epsilon \end{pmatrix}.$$

Hence $\operatorname{sgn}(F_\beta)$ has two distinct binary vectors but the matrix $\lim_{\gamma\to\beta^-} \operatorname{sgn}(F_\gamma)$ has four distinct binary vectors. So at $\gamma = \beta$, $\rho(F_\gamma)$ has a discontinuity: $\rho(F_\beta) = 2$ and $\lim_{\gamma\to\beta^-} \rho(F_\gamma) = 4$, therefore $\gamma_4 = \gamma_3 = \beta$ so indeed it is not the case that the strict inequality $\gamma_4 < \gamma_3$ holds.

For our purposes, we define the VC-dimension of a $\{1, -1\}$-matrix to be that of the set system in which the indicator functions of the sets correspond to the columns of the matrix (with a 1 indicating membership). So, a set of m rows is shattered when the sub-matrix induced by those rows has all $\{1, -1\}$ vectors of length m as columns.

Proposition 5. *For every* $0 \le \gamma \le 1$, $VC(\operatorname{sgn}(F_\gamma)) \le \log_2(\rho(F_\gamma))$.

Proof. The number of distinct columns of the matrix $\operatorname{sgn}(F_\gamma)$ is, by definition, $\rho(F_\gamma)$. Let S be a set of rows of maximal size m that is shattered by $\operatorname{sgn}(F_\gamma)$. Then the number of distinct columns of $\operatorname{sgn}(F_\gamma)$ must be at least 2^m. This implies that $\rho(F_\gamma) \ge 2^m$, that is, $m \le \log_2 \rho(F_\gamma)$. So the largest set that can be shattered by $\operatorname{sgn}(F_\gamma)$ is of size no larger than $\log_2 \rho(F_\gamma)$. □

4.4. **The matrix F'_γ.** Denote by $0 < a_L < a_{L-1} < \cdots < a_1 < 1$ the set of L distinct positive entries of F where a_1 and a_L are the maximum and minimum positive entries of F which are less than 1 and greater than 0, respectively. For $1 \le i \le L$, define the multiplicity of a_i, as the number of times that a_i appears in F and denote it by m_i. Let

$$M := \sum_{l=1}^{L} m_i \tag{4.13}$$

be the number of positive components of F.

We refer to $S_F = \{a_1, a_2, \ldots, a_L\}$ as the *positive set* of F.

For $0 \leq i \leq N$, let us denote by the i^{th} *shell* of the binary N-dimensional cube $\{-1,1\}^N$ (N-cube) the set of all vertices that have i components that are 1.

We next define a few quantities, leaving the dependence on N implicit. For $1 \leq j \leq N$ denote by

$$c_j := \binom{N}{j}$$

the number of vertices in the j^{th} shell.

Denote by

$$b_n := \sum_{j=1}^{n} c_j \tag{4.14}$$

the number of vertices of the cube contained in the first n shells and we let $b_0 := 0$

Let us define

$$\ell_n := \sum_{j=1}^{n} j c_j, \tag{4.15}$$

the total 'weight' (number of 1-entries) in all of the vertices of the cube that are in the first n shells.

For a positive integer m define

$$Q(m) := \min\{q \geq 1 : \ell_q \geq m\}. \tag{4.16}$$

For instance, if $m = 17$, $N = 4$, then $1\binom{4}{1} + 2\binom{4}{2} + 1 = 17$ so $Q(17) = 3$.

Let

$$\Delta := m - \ell_{Q(m)-1},$$

and then define $\lceil m \rceil$ as the following 'rounded' value:

$$\lceil m \rceil := \begin{cases} m & \text{if } Q(m) = 1 \text{ or } \Delta \bmod Q(m) = 0 \\ m + (Q(m) - \Delta \bmod Q(m)) & \text{otherwise.} \end{cases}$$

So, $\lceil m \rceil$ is the smallest integer greater than or equal to m which is the total weight (number of 1-entries) in a set of vertices of the cube, where that set is formed by first populating the first shell, then the second, and so on. So, m 1-entries might not be enough to form all the vertices of shells 1 to $Q(m)-1$ and then some vertices in shell $Q(m)$: an additional (at most $Q(m) - 1$) 1-entries might be necessary (to 'complete' a vertex in shell $Q(m)$).

For instance (continuing the above example), since $Q(17) = 3 \geq 2$ then $\lceil 17 \rceil = 17 + (3 - 1 \bmod 3) = 19$, and indeed 19 1-entries are required in the vertices in shells 1 and 2, and in the first vertex of the 3rd shell.

Define integers R_i as follows:

$$\begin{aligned} R_0 &= 0, \\ R_i &= \lceil R_{i-1} + m_i \rceil, \quad 1 \leq i \leq L. \end{aligned} \tag{4.17}$$

Define the values α_j, $1 \leq j \leq R_L$ as follows: for $1 \leq i \leq L$,

$$\alpha_j = a_i, \quad \text{for all } j \in \{R_{i-1} + 1, \ldots, R_i\}. \tag{4.18}$$

We now construct a matrix F' with the same positive set, S_F, as F; that is, its positive components are the α_i values. F' is defined as follows:

$$F' := \begin{pmatrix}
- & - & - & \cdots & \alpha_{\ell_1} & - & - & \cdots & \alpha_{\ell_2} & - & - & \cdots & \alpha_{\ell_3} & \vdots & \vdots & - & \cdots & - \\
\vdots & \vdots & \vdots & & - & \vdots & \vdots & & \alpha_{\ell_2-1} & \vdots & \vdots & & \alpha_{\ell_3-1} & \vdots & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & - & \vdots & \vdots & & \alpha_{\ell_3-2} & \vdots & \vdots & \vdots & & - \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & - & \vdots & \vdots & \vdots & & \alpha_{R_L} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & - & & \vdots & \vdots & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots & - & \cdots & \vdots & - & \alpha_{\ell_2+6} & \cdots & \vdots & \vdots & \cdots & \alpha_{\ell_{K-1}+K} & \cdots & \alpha_{R_L-1} \\
\vdots & \vdots & - & & \vdots & - & \alpha_{\ell_1+4} & & \vdots & \alpha_{\ell_2+3} & - & & \vdots & \vdots & \vdots & \vdots & & \vdots \\
\vdots & - & \alpha_2 & & \vdots & \alpha_{\ell_1+2} & - & & \vdots & \alpha_{\ell_2+2} & \alpha_{\ell_2+5} & & \vdots & \vdots & \vdots & \alpha_{\ell_{K-1}+2} & & \alpha_{R_L-K+1} \\
- & \alpha_1 & - & \cdots & - & \alpha_{\ell_1+1} & \alpha_{\ell_1+3} & \cdots & - & \alpha_{\ell_2+1} & \alpha_{\ell_2+4} & \cdots & - & \vdots & \vdots & \alpha_{\ell_{K-1}+1} & \cdots & \vdots
\end{pmatrix}$$

where we write the symbol '$-$' to indicate any arbitrarily chosen negative real value and $K = Q(R_L)$ (where Q is as described above). Note that in defining F', any fixed ordering of binary vectors, grouped according to increasing weight (shell) is appropriate, and the highest shell (the K^{th} shell) may be only partially full.

Example. To illustrate with a small example, suppose that $N = 4$ and that the positive set S_F is $a_6 < a_5 < a_5 < a_4 < a_3 < a_2 < a_1$, and that each a_i has multiplicity $m_i = 1$, except for a_6, which has multiplicity 3. Then we have:

$$R_0 = 0,\ R_1 = \lceil 1 \rceil = 1,\ R_2 = \lceil 2 \rceil = 2, R_3 = \lceil 3 \rceil = 3, R_4 = \lceil 4 \rceil = 4,$$

$$R_5 = \lceil 5 \rceil = 6,\ R_6 = \lceil 6 + 3 \rceil = \lceil 9 \rceil = 10.$$

The corresponding values of α_i are therefore:

$$\alpha_1 = a_1,\ \alpha_2 = a_2,\ \alpha_3 = a_3,\ \alpha_4 = a_4,$$

$$\alpha_5 = \alpha_6 = a_5,$$

$$\alpha_7 = \alpha_8 = \alpha_9 = \alpha_{10} = a_6.$$

A suitable F' is then

$$\begin{pmatrix} - & - & - & - & a_4 & - & - & a_6 \\ - & - & - & a_3 & - & - & a_6 & - \\ - & - & a_2 & - & - & a_5 & - & - \\ - & a_1 & - & - & - & a_5 & a_6 & a_6 \end{pmatrix}.$$

For positive integer m let us define

$$\lambda(m) := b_{Q(m)-1} + \frac{\lceil m \rceil - \ell_{Q(m)-1}}{Q(m)} \tag{4.19}$$

and define $\lambda(0) := 0$. Note that $\lceil R_i \rceil = R_i$ because $R_i - \ell_{Q(R_i)-1}$ is an integer multiple of $Q(R_i)$ and hence for $1 \le i \le L$, we have

$$\lambda(R_i) := b_{Q(R_i)-1} + \frac{R_i - \ell_{Q(R_i)-1}}{Q(R_i)}. \tag{4.20}$$

Then, for each i, $\lambda(R_i)$ equals the number of columns of F' that are occupied by entries $\alpha_1, \ldots, \alpha_{R_i}$. In the above example,

$$\lambda(R_6) = \lambda(10) = b_1 + \frac{10 - \ell_1}{2} = 4 + \frac{10-4}{2} = 7.$$

Note that the choice of values (4.18) ensures that every column of F' consists of at most a single value $a_i \in S_{F'}$ for some $1 \le i \le L$; that is, no column has a mix of several different a_i values.

Henceforth denote by κ' the pseudo-rank of F' and let v_i be the columns of F'. Let us evaluate the characteristic points γ'_r of F', $1 \le r \le \kappa'$. As in (4.12), they are ordered from the lowest $\gamma'_{\kappa'}$ to largest γ'_1. We start with $\gamma > 1$ and in this case $\rho(F'_\gamma) = 1$ since the matrix $\operatorname{sgn}(F'_1)$ consists of all-(-1) components. We begin to decrease γ. As long as γ is larger than a_1, $\rho(F'_\gamma) = 1$ but when $\gamma < a_1$, $\lambda(R_1)$ new binary vectors are formed, namely, the vectors $\operatorname{sgn}(v_1), \ldots, \operatorname{sgn}(v_{\lambda(R_1)})$, which differ from the all-(-1) vector, and thus $\rho(F'_\gamma)$ changes from a value of 1 to $\lambda(R_1)+1$. The characteristic points for $r \in \{2, \ldots, \lambda(R_1)+1\}$ are $\gamma'_r = a_1$ since a_1 is the supremum of all values γ such that $\rho(F'_\gamma) \ge r$.

As we continue to decrease γ and pass the value a_2, we have $\lambda(R_2) - \lambda(R_1)$ new binary vectors that are formed, namely, the vectors $\operatorname{sgn}(v_{\lambda(R_1)+2})$, ..., $\operatorname{sgn}(v_{\lambda(R_2)+1})$, all of which differ from the previously-created vectors. So $\gamma'_r = a_2$, for

$$r \in \{\lambda(R_1) + 2, \ldots, \lambda(R_2) + 1\}.$$

In general, when we pass $\gamma = a_k$ from above we have $\lambda(R_k) - \lambda(R_{k-1})$ new binary vectors formed relative to when $\gamma \geq a_k$, so ρ jumps up by this amount. By definition, κ' equals the number of distinct binary columns of sgn (F'). Hence

$$\kappa' = \lambda(R_L) + 1. \tag{4.21}$$

We can express the characteristic points of F'_γ as follows: for all $1 \leq k \leq L$,

$$\gamma'_r = a_k, \text{ for all } r \text{ that satisfy: } r \in \{\lambda(R_{k-1}) + 2, \ldots, \lambda(R_k) + 1\}, \tag{4.22}$$

where we define $\lambda(R_0) = 0$. Note that, by construction of F', every a_k is a characteristic point and every characteristic point must be some a_k. An a_k could be the characteristic point γ'_r for a range of values of r as (4.22) indicates. Note also that in (4.22) the largest value of r is κ' and $\gamma'_{\kappa'} = a_L$.

Figure 4.1 shows an example of $\rho(F'_\gamma)$ with its characteristic points. Note that the graph is always non-increasing and piecewise constant, not just for this example, because of the special form of the matrix F'_γ.

4.5. **Relating F and F'.** The matrices F and F' have the same positive sets but can have different pseudo-ranks.

Theorem 6. *Suppose a matrix F with positive set S_F is given, and construct a matrix F' (using the elements of S_F) as described above. Then*

$$\rho(F_\gamma) \leq \min\left\{\rho(F'_\gamma), N(N-1)\right\}$$

for every γ.

To prove Theorem 6, we first have the following preliminary result about binary matrices. Here, by the *weight* of a column of a $\{-1, 1\}$-matrix, we mean the number of 1 entries in the column and by the weight of the matrix we mean the total number of 1 entries in the matrix.

Lemma 7. *Let M be a positive integer and let $r := \lambda(M) - b_{Q(M)-1}$. Let $A^*(M)$ be any matrix with the property that it solely contains all columns of weight 0, 1, ..., up to and including $Q(M) - 1$, and r columns of weight $Q(M)$. Then, for any $\{-1, 1\}$-matrix A of weight M, the number of distinct columns of A is no more than the number of distinct columns of $A^*(M)$.*

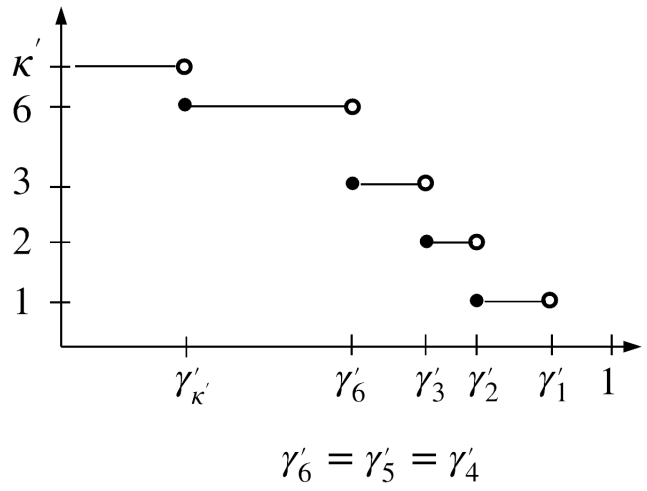

FIGURE 4.1. Example of pseudo-rank function for F'_γ with characteristic points γ'_r. At a point γ'_r, the symbol $\bullet$ for $\rho(F'_{\gamma'_r})$ indicates that $\rho(F'_{\gamma'_r})$ is well defined, and $\circ$ indicates that only the limit exists $\lim_{\gamma\to\gamma'^{-}_r} \rho(F'_\gamma)$.

Proof. We describe an algorithm to produce $A^*(M)$ given A, which is displayed in Algorithm 1 and Procedures 2, 3, 4. Suppose A has N rows and, for $1 \le i \le N$, let

$$S_i = \{\mathbf{x} \in \{-1, 1\}^N : \mathrm{wt}(\mathbf{x}) = i\},$$

where $\mathrm{wt}(\mathbf{x})$ is the weight of $\mathbf{x}$, the number of 1 entries. (We call S_i the ith shell.) We will describe an algorithmic procedure for successively making changes in the matrix A in such a way that, at each step, the new matrix has the same weight and at least as many distinct columns as the previous matrix. We show the procedure ends with a matrix that has no more distinct columns than a matrix of type $A^*(M)$. The result will then follow.

The procedure starts with $A_0 = A$ and we will denote successive matrices by $A_0, A_1, \ldots A_s$. For each stage i, we define D_i to be the (multi-)set of columns of A_i of weight at least 1 which are duplicates of other columns, that is, for each column $u \in D_i$ there is exactly one $v \notin D_i$ such that $u = v$. We define E_i to be the (multi-)set of columns of A where each column in E_i has a weight greater than $Q(M)$. The procedure works as follows, at stage i: if there is a $\{-1, 1\}$-vector $\mathbf{x}$ of weight $j < Q(M)$ which is not a column of A_i, then we create a new (additional) column equal to $\mathbf{x}$ and (to maintain the number of 1s in the resulting matrix A_{i+1}) we replace a total of j 1s by -1s in any vectors in $D_i \setminus E_i$ and also, if necessary,

in any vectors in E_i. (We try to use the duplicate columns of weight no more than $Q(M)$ first.) We can think of this as 'transferring' 1 entries from these columns to make the new one. This action potentially makes changes to columns in D_i and E_i, so these sets are updated accordingly to D_{i+1} and E_{i+1}. The number of distinct columns in A_{i+1} is not smaller than the number in A_i. This is because changing any column in D_i (which is already present elsewhere as a column) cannot decrease the number of distinct columns, and changing any column in E_i will, at worst, create a duplicate column: however, given that the new column $\mathbf{x}$ has been introduced into A_{i+1}, the net effect would be (at worst) for the number of distinct columns to remain the same.

Because $Q(M) = \max\{q : M > \ell_{q-1}\}$, there are enough 1 entries in the matrices so that, eventually, all columns of weight less than $Q(M)$ are present. Call A_s the matrix that results once all such columns first appear. Any additional columns of A_s will be either (i) duplicates of columns of any weight, or (ii) columns of weight $Q(M)$, or (iii) columns of weight greater than $Q(M)$.

We can 'transfer' all the 1 entries of the columns of type (i) or (iii) to create new columns of weight $Q(M)$ and at most one duplicate of a column of weight less than $Q(M)$. The resulting matrix will have no more distinct columns than any matrix of type $A^*(M)$, and this implies the result. (We will not need to have any columns of weight greater than $Q(M)$ because $M \leq \ell_{Q(M)}$.) □

Algorithm 1 Produce $A^*(M)$

Input: Matrix A of N-dimensional binary columns with $\mathrm{wt}(A) = M$
Output: Binary matrix $A^*(M)$

$i := 0$, $A_0 := A$
while $|\{x : x \in A_i, \mathrm{wt}(x) < Q(M)\}| < b_{Q(M)-1}$ **do**
  $D_i = \mathrm{getD}(A_i)$
  $E_i = \mathrm{getE}(A_i)$
  $s := 0$ // current supply of 1s
  if $\exists x \in \{-1, 1\}^N$, $x \notin A_i$, $\mathrm{wt}(x) < Q(M)$ **then**
    $n := \mathrm{wt}(x)$ // need n 1s
    $s :=$ **getOnes**$(A_i, D_i \setminus E_i, n)$ // A_i may change
    if $s < n$ **then**
      $s := s +$ **getOnes**$(A_i, E_i, n - s)$ // A_i may change
      // s now equals n
    end if
    construct a new N-dimensional column equal to x using s 1s
    $A_{i+1} := A_i \bigcup \{x\}$ // add new column x to A_i
    $i := i + 1$
  end if
end while
Denote $A_i = \{u^{(1)}, \ldots, u^{(r)}\}$
$D_i = \mathrm{getD}(A_i)$
$E_i = \mathrm{getE}(A_i)$
while $\mathrm{not}\Big(|D_i| = 0$ or $\big(D_i = \{j\}$ and $\mathrm{wt}(u^{(j)}) < Q(M)\big)\Big)$ **do**
  if $\exists x \in \{-1, 1\}^N$, $x \notin A_i$, $\mathrm{wt}(x) = Q(M)$ **then**
    $n := \mathrm{wt}(x)$ // need n 1s
    $s :=$ **getOnes**$(A_i, D_i \setminus E_i, n)$ // A_i may change
    if $s < n$ **then**
      $s := s +$ **getOnes**$(A_i, E_i, n - s)$ // A_i may change
      // s now equals n
    end if
    construct a new N-dimensional column equal to x using s 1s
    $A_{i+1} := A_i \bigcup \{x\}$ // add new column x to A_i
    $i := i + 1$
  end if
end while
return A_i

Procedure 2 getD(V)

Input: List of N-dimensional binary vectors $V = \{v_1, \ldots, v_r\}$
Output: List D of indices of duplicate vectors in V

$D = \emptyset$ // Initialize D to an empty list
$S = \emptyset$ // Initialize S to an empty set
for $1 \leq i \leq r$ **do**
  if $v_i \in S$ **then**
    add i to end of D
  else
    add v_i to S
  end if
end for
return D

Procedure 3 getE(V)

Input: List of N-dimensional binary vectors $V = \{v_1, \ldots, v_r\}$
Output: $E := \{i : \mathrm{wt}(v_i) > Q(M)\}$

$E = \emptyset$ // Initialize E to an empty list
for $1 \leq i \leq r$ **do**
  if $\mathrm{wt}(v_i) > Q(M)$ **then**
    add v_i to end of E
  end if
end for
return E

Procedure 4 getOnes(V, B, n)

Input: List of N-dimensional binary vectors $V = \{v^{(1)}, \ldots, v^{(r)}\}$, list of integers $B = \{i_j : 1 \leq i_j \leq r, 1 \leq j \leq l\}$, and integer n

Output: integer s = number of 1 converted to -1 in V

```
1: s:=0
2: for 1 ≤ j ≤ l do
3:    for 1 ≤ k ≤ N do
4:       if v_k^(i_j) = 1 then
5:          v_k^(i_j) = −1
6:          s := s + 1
7:       end if
8:       if s=n then
9:          goto 13
10:      end if
11:   end for
12: end for
13: return s
```

We can then prove Theorem 6.

Proof of Theorem 6.

Proof. When $\gamma > 1$, $\operatorname{sgn}(F_\gamma)$ and $\operatorname{sgn}(F'_\gamma)$ both have weight 0. As γ is decreased, the 1 entries in $\operatorname{sgn}(F_\gamma)$ will be in positions occupied in F by $a_1, \dots a_k$ for some k. The positive entries of F' are, by construction, the same as the positive entries of F, albeit with possibly different multiplicities m_i, and so the 1 entries in $\operatorname{sgn}(F'_\gamma)$ will also be in positions occupied in F' by $a_1, \dots a_k$. Let $M_k := \sum_{i=1}^{k} m_i$ then M_k is the number of 1 entries, and hence it is the weight of the matrix $A_\gamma := \operatorname{sgn}(F_\gamma)$.

By Lemma 7, the number of distinct columns of A_γ, namely $\rho(F_\gamma)$, is no more than that of any matrix of the type $A^*(M_k)$ described in the Lemma and this number equals $1+b_{Q(M_k)-1}+r = \lambda(M_k)+1$. By construction, the number of distinct columns of the matrix $\operatorname{sgn}(F'_\gamma)$, namely $\rho(F'_\gamma)$, is the sum of the number of columns with a weight at least one (4.20) plus the all-(-1) column, which in total equals $\lambda(R_k)+1$. We have

$$\begin{aligned} M_k = m_1 + m_2 + \cdots + m_k &\leq \lceil m_1 \rceil + m_2 + \cdots + m_k \\ &\leq \lceil \lceil m_1 \rceil + m_2 \rceil + m_3 + \cdots + m_k \\ &\leq \lceil \cdots \lceil \lceil \lceil m_1 \rceil + m_2 \rceil + m_3 \rceil + \cdots + m_k \rceil \\ &= R_k \end{aligned}$$

and therefore, from (4.19), it follows that $\lambda(M_k) \leq \lambda(R_k)$. Combining all of the above, we have

$$\rho(F_\gamma) \leq \lambda(M_k) + 1 \leq \lambda(R_k) + 1 = \rho(F'_\gamma)$$

and combining with (4.10) completes the proof. □

The next lemma states a bound on the VC-dimension of $\operatorname{sgn}(F_{a_l})$, $1 \leq l \leq L$.

Lemma 8. *Let* $1 \leq l \leq L$,

$$VC\left(\operatorname{sgn}\left(F_{a_l}\right)\right) \leq \min\left\{\log_2\left(\lambda\left(R_l\right)+1\right), \log_2(N(N-1))\right\}.$$

Remark 9. The second entry above, $\log_2(N(N-1))$, is the trivial bound on the VC-dimension of F_{a_l} since it is independent of a_l.

Proof. From (4.22) it follows that if $r = \lambda\left(R_l\right)+1$ then $\gamma'_r = a_l$ and thus $\rho\left(F'_{a_l}\right) = \lambda\left(R_l\right)+1$. From Theorem 6 it follows that $\rho\left(F_{a_l}\right) \leq \min\{\rho\left(F'_{a_l}\right), N(N-1)\}$. We apply Proposition 5 to obtain

$$VC\left(\operatorname{sgn}\left(F_{a_l}\right)\right) \leq \log_2\left(\rho\left(F_{a_l}\right)\right).$$

It follows that

$$\begin{aligned} VC\left(\operatorname{sgn}\left(F_{a_l}\right)\right) &\leq \log_2\left(\rho\left(F_{a_l}\right)\right) \\ &\leq \min\left\{\log_2\left(\rho\left(F'_{a_l}\right)\right), \log_2(N(N-1))\right\} \\ &= \min\left\{\log_2\left(\lambda\left(R_l\right)+1\right), \log_2(N(N-1))\right\} \end{aligned}$$

from which the stated claim follows. □

For $0 \leq l \leq L$, let

$$\tau_l := \log_2\left(\lambda\left(R_l\right)+1\right)$$

where $\tau_0 = 0$ because $R_0 = 0$.

We henceforth choose for η_l (as defined in section 3) the value $\eta_l := a_l$. Thus $\Gamma_l := (a_l, a_{l-1}]$ for $1 \leq l \leq L$ and $\Gamma_{L+1} = [0, a_L]$. We choose for $\mathsf{d}(\gamma)$ of section 3.2 the following step function, for $1 \leq l \leq L$,

$$\mathsf{d}(\gamma) := \min\left\{\tau_{l-1}, \log_2(N(N-1))\right\}, \text{ for all } \gamma \in \Gamma_l,\ 1 \leq l \leq L+1$$

where we note that this interval does not contain the discontinuity point a_l. At $\gamma = a_l$, $\mathsf{d}(a_l) = \tau_l$. Thus d is a non-increasing step function with discontinuity points a_l, $1 \leq l \leq L$.

It follows from Lemma 8 that for $1 \leq l \leq L$,

$$VC(\operatorname{sgn}\left(F_{a_l}\right)) \leq \mathsf{d}(a_l)$$

and, $\mathsf{d}(\gamma) = \mathsf{d}(a_{l-1})$ for $\gamma \in \Gamma_l$, as required (see section 3).

As mentioned in section 3.2, the columns of the matrix F_γ correspond to the sets $W_\gamma^{(i,j)}$ defined in (4.1). Hence an upper bound on the VC-dimension of $\operatorname{sgn}(F_\gamma)$ is also an upper bound on the VC dimension of the class $\mathcal{W}_\gamma$, and hence

$$VC(\mathcal{W}_{a_l}) \leq \mathsf{d}(a_l).$$

From (4.4), it follows that

$$VC(\mathcal{M}_{a_l}^+) \leq \mathsf{d}(a_l), \quad VC(\mathcal{M}_{a_l}^-) \leq \mathsf{d}(a_l)$$

which is what we need, as mentioned in Remark 2.

5. Main result

5.1. **The main theorem.** The following is the main result of the paper.

Theorem 10. *Let $N \geq 1$ and let $\mathcal{X} = \{x_i\}_{i=1}^N$ be a finite distance space with a distance $d(x_i, x_j)$, normalized such that $diam(\mathcal{X}) = \max_{1\leq i,j\leq N} d(x_i, x_j) = 1$. Let*

$$f_j^{(i)} := \begin{bmatrix} d(x_1, x_j) - d(x_1, x_i) \\ \vdots \\ d(x_N, x_j) - d(x_N, x_i) \end{bmatrix},$$

$$F^{(i)} = \left[f_1^{(i)}, \ldots, f_{i-1}^{(i)}, f_{i+1}^{(i)}, \ldots, f_N^{(i)}\right]$$

and define the $N \times N(N-1)$ matrix F by

$$F := \left[F^{(1)}, \ldots, F^{(N)}\right].$$

Let $0 = a_{L+1} < a_L < \cdots < a_1 < a_0 = 1$ be the values of the positive entries of F and let $m_l \geq 1$ be the number of times that a_l appears in F, $1 \leq l \leq L$. Define $\Gamma_l := (a_l, a_{l-1}]$, $1 \leq l \leq L$, $\Gamma_{L+1} := [0, a_L]$ and $C := \sum_{l=1}^L a_l$. On the interval $[0, 1]$, define the non-increasing step function

$$\mathsf{d}(\gamma) := \begin{cases} \min\left\{\log_2\left(\lambda\left(R_{l-1}\right)+1\right), \log_2(N(N-1))\right\}, & \gamma \in \Gamma_l,\ 2 \leq l \leq L+1, \\ 0, & \gamma \in \Gamma_1 := (a_1, 1] \end{cases} \tag{5.1}$$

where the R_i are as defined in (4.17). Let $P^m := P_{XY}^m$ be a probability measure over $\mathcal{X} \times \mathcal{Y}$. For any $0 < \delta \leq 1$, with P^m-probability at least $1-\delta$ the following holds for a sample $\xi := \{(x_i, y_i)\}_{i=1}^m \subseteq (\mathcal{X} \times \mathcal{Y})^m$: for all $\gamma \in (0, 1]$, and for any half-space h_Π,

$$P\left(Y f_\Pi(X) \leq 0\right) \leq \frac{1}{m}\sum_{j=1}^m \mathbb{I}\left\{Y_j f_\Pi(X_j) \leq \gamma\right\} + \epsilon, \tag{5.2}$$

where

$$\epsilon := \sqrt{\frac{32}{m}\left(\mathsf{d}(\gamma)\ln\left(\frac{2em}{\mathsf{d}(\gamma)}\right) + \ln\left(\frac{8(C+1)}{\gamma\delta}\right)\right)}. \tag{5.3}$$

5.2. **Discussion of the main theorem.**

Remark 11. We first note that, ignoring the $\ln(1/\mathsf{d})$ and $\ln(1/\gamma\delta)$ terms, the bound ϵ of the theorem is $O\left(\sqrt{(\mathsf{d}(\gamma)\ln(m))/m}\right)$.

Let us assess how large $\mathsf{d}(\gamma)$ is and specifically how it depends on the main parameters N and γ. To start with, consider the second term on the right of (4.20). This second term gives the number of columns in the $(Q(R_i))^{th}$ shell that appear in F' and are formed from the entries $\alpha_1, \alpha_2, \ldots \alpha_{R_i}$. We now make a rough estimate and say that this number is no more than the number of vertices in the $(Q(R_i))^{th}$ shell of the N-cube, which is $\binom{N}{Q(R_i)}$ (depending on the value of R_i, this number may be much smaller and even close to zero). Adding this to the first term of (4.20) it follows that $\lambda(R_i) \leq b_{Q(R_i)}$. Hence

$$\begin{aligned}\lambda(R_i) + 1 &\leq \sum_{j=1}^{Q(R_i)} \binom{N}{j} + 1 \\ &= \sum_{j=0}^{Q(R_i)} \binom{N}{j}.\end{aligned}$$

Bounding $\log_2(N(N-1))$ from above by $2\log_2 N$ it follows that the expression for $\mathsf{d}(\gamma)$ is bounded as follows,

$$\mathsf{d}(\gamma) \leq \min\left\{\log_2\left(\sum_{j=0}^{Q(R_{l-1})}\binom{N}{j}\right), 2\log_2 N\right\} \tag{5.4}$$

for $\gamma \in \Gamma_l$, $2 \leq l \leq L+1$. For l such that $Q(R_{l-1}) > 2$, and for all $N \geq 4$, the right side of (5.4) is $2\log_2 N$ (because $N^2 < \sum_{j=0}^{3}\binom{N}{j}$) and hence it does not depend on l nor on γ. Let us denote by

$$l^* := \max\{l \geq 1 : Q(R_{l-1}) \leq 2\}.$$

The interesting set of values for γ (where the bound on $\mathsf{d}(\gamma)$ is influenced by γ) is the set $\bigcup_{l=1}^{l^*}\Gamma_l$. If γ is too small, that is, it falls outside this set, then $\mathsf{d}(\gamma)$ becomes the trivial upper bound on the VC-dimension of the class of half-spaces. The higher the multiplicity values m_i of a_i, the smaller the value of l^* because it takes fewer positive values a_i to end up with enough 1 entries to fill the first and second shells. This means that γ needs to be larger in order to make the bound (5.4) smaller than the trivial value of $2\log_2 N$ (also, the set $\bigcup_{l=1}^{l^*}\Gamma_l$ becomes smaller).

The rough estimate used above makes γ appear to have a limited effect on $\mathsf{d}(\gamma)$, in that its value is limited to either 0, $\log_2(1+N)$, $\log_2(1+N+\binom{N}{2})$ or $2\log_2 N$. But, to be more exact (without using this simple estimate) $\mathsf{d}(\gamma)$ can actually take any value in the set $\{0, 1, \log_2 3, 4, \log_2 5, \ldots, 2\log_2 N\}$ and therefore γ has a 'smoother' effect on the bound.

There is not much more that can be said in general for $\mathsf{d}(\gamma)$ so let us consider an example of a specific distance space $\mathcal{X}$.

Example 12. Let $\mathcal{X}$ be a distance space whose corresponding matrix F has positive entries $0 < a_L < a_{L-1} \cdots < a_1 < 1$ with multiplicity values $m_j = 1$, for $1 \leq j \leq L$. In this case the matrix F' is simple, in that column number l contains entries that only take the value a_l, $1 \leq l \leq L$. Define the shell index as a function of column index l by $\tilde{Q}(l)$. It equals

$$\tilde{Q}(l) := j \text{ if } b_{j-1} < l \leq b_j$$

where b_j is defined in (4.14). Consider the variable $Q(R_{l-1})$ in (5.4). By definition, R_l is the highest index j of α_j such that $\alpha_j = a_l$. From above, for this example, the only column that has entries a_l is the l^{th} column. Thus $\alpha_{R_{l-1}}$ must be in the $(l-1)^{th}$ column. Hence we have

$$Q(R_{l-1}) = \tilde{Q}(l-1). \tag{5.5}$$

Let us write $\mathcal{Q}(\gamma)$ for Q as a function of γ. Any $\gamma \in \Gamma_l$ maps to $Q(R_{l-1})$ in the bound (5.4). Thus we have

$$\mathcal{Q}(\gamma) := \sum_{l=1}^{L+1} Q(R_{l-1}) \mathbb{I}\{\gamma \in \Gamma_l\}$$

which, with (5.5), implies that

$$\mathcal{Q}(\gamma) = \sum_{l=1}^{L+1} \tilde{Q}(l-1) \mathbb{I}\{\gamma \in \Gamma_l\}.$$

$\mathcal{Q}(\gamma)$ is a non-increasing step function with breaks at points $\gamma = a_{b_n}$ where a_{b_n} is a decreasing subsequence of the sequence $\{a_l\}_{l=1}^{L}$, $1 \leq n \leq Q(R_L)$, where $Q(R_L)$ is the index of the highest shell that is either 'full' or 'partially full' by the columns of $\operatorname{sgn}(F')$. Note that $\mathcal{Q}(\gamma)$ starts at a maximum value of $Q(R_L)$ when $\gamma \in [0, a_{b_{Q(R_L)-1}+1}]$, then decreases to $Q(R_L) - 1$ for $\gamma \in (a_{b_{Q(R_L)-1}+1}, a_{b_{Q(R_L)-2}+1}]$, and so forth, until it eventually decreases down to zero when $\gamma \in (a_1, 1]$. Now, as mentioned above, the interesting set of values for γ is $\bigcup_{l=1}^{l^*} \Gamma_l$. By definition, $\Gamma_{l^*} = (a_{l^*}, a_{l^*-1}]$ so, for this example, we have $l^* = b_2 + 1 = \binom{N}{2} + N + 1$. As long as γ falls

in the interval $\bigcup_{l=1}^{l^*} \Gamma_l = (a_{l^*}, 1]$ then $\mathcal{Q}(\gamma) \leq 2$ and the bound (5.4) is non-trivial for all $N \geq 4$, that is, $\log_2 \left(\sum_{j=0}^{\mathcal{Q}(\gamma)} \binom{N}{j} \right) < 2 \log_2 N$. The smaller the $\left(\binom{N}{2} + N + 1 \right)^{th}$ positive value $a_{\binom{N}{2}+N+1}$, the larger the range of values of γ that yield a non-trivial bound. Thus in this particular example, for $N \geq 4$, for all $\gamma \in (a_{\binom{N}{2}+N+1}, 1]$, $\mathsf{d}(\gamma) \leq \log_2 \left(\sum_{j=0}^{\mathcal{Q}(\gamma)} \binom{N}{j} \right)$ and the bound is smaller than the trivial bound of $2 \log_2 N$. □

If a distance space $\mathcal{X}$ is $\mathbb{R}^n$, then learning the class of half-spaces amounts to learning the class of linear functions on $\mathbb{R}^n$ which, when thresholded, yield half-space classifiers. The bound on the error in this case is $O\left(\sqrt{\frac{\log^2 m}{m\gamma^2}} \right)$ (see [9]), and is useful for learning with Support Vector Machines (linear machines on the feature-space).

Let us compare this rate with that of Theorem 10. The dependence on m is approximately the same (within a $\log m$ factor). Comparing the dependence on γ is more subtle since in the Euclidean case it is $O(\frac{1}{\gamma})$ while in Theorem 10 it is $O(\sqrt{\mathsf{d}(\gamma)})$ and $\mathsf{d}(\gamma)$ decreases with γ with steps that depend on the positive set S_F of the distance space. In the above example, these steps are at a_{b_n}, $1 \leq n \leq Q(R_L)$, but in general, if there are multiplicity values m_i that are larger than 1 then $\mathsf{d}(\gamma)$ can decrease with γ faster, albeit the range of γ that has an effect on the bound is smaller.

As is the case of the Euclidean space, Theorem 10 shows that for a general distance space, the learning bound is independent of any kind of 'metric dimension' or richness quantity such as a covering number of the space (in the case of $\mathbb{R}^n$ this quantity is the dimension n). The only factor in the bound (5.3) (via the bound (5.4) on $\mathsf{d}(\gamma)$), that resembles a complexity of the space is $\log_2(N)$ which comes from the trivial bound on the number of columns of F. However, as mentioned above, this $\log_2(N)$ factor only enters the bound if γ is outside the interesting range. Thus Theorem 10 almost maintains the "dimension-independence" that large-margin learning error bounds are known to offer.

6. Conclusions

We have studied error bounds for classifiers that are generalizations of half-spaces to arbitrary finite distance spaces (a significantly more general and perhaps more applicable setting than that of a metric space) and have obtained bounds that depend on a notion of 'sample width'. We have shown that from the positive points

a_i of a matrix F associated with a finite distance space, we can directly characterize the influence of the width parameter on the error bound for learning half-spaces.

For further work, we think that it should be possible to remove the finiteness assumption on the distance space and to carry out the analysis without a matrix representation (since the notion of pseudo-rank holds also for an infinite class of functions, in the same manner as the growth-function is well-defined for infinite classes of functions).

7. Acknowledgements

This work was supported in part by a research grant from the Suntory and Toyota International Centres for Economics and Related Disciplines at the London School of Economics.

We are very grateful to the referees for their useful comments.

References

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
[2] M. Anthony and J. Ratsaby. A hybrid classifier based on boxes and nearest neighbors. *Discrete Applied Mathematics*, 172:1–11, 2014.
[3] M. Anthony and J. Ratsaby. Learning bounds via sample width for classifiers on finite metric spaces. *Theoretical Computer Science*, 529:2–10, 2014.
[4] M. Anthony and J. Ratsaby. A probabilistic approach to case-based inference. *Theoretical Computer Science*, 589:61–75, 2015.
[5] M. Anthony and J. Ratsaby. Multi-category classifiers and sample width. *Journal of Computer Systems and Sciences*, 82(8):1223–1231, 2016.
[6] M. Anthony and J. Ratsaby. Multi-category classifiers and sample width. *J. Comput. Syst. Sci.*, 82(8):1223–1231, 2016.
[7] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4), 929–965, 1989.
[9] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based learning methods*. Cambridge University Press, 2000.
[10] E. Deza M. Deza. *Encyclopedia of Distances*, volume 15 of *Series in Computer Science*. Springer-Verlag, 2009.
[11] L. Gottlieb, A. Kontorovich, and P. Nisnevitch. Nearly optimal classification for semimetrics. *Journal of Machine Learning Research*, 18(37):1–22, 2017.

[12] M. Hein, O. Bousquet and B. Schólkopf. *Maximal Margin Classification for Metric Spaces. Journal of Computer and System Sciences* 71(3): 333–359, 2005.
[13] V. N. Vapnik. *Statistical Learning Theory.* Wiley, 1998.
[14] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 264–280, 1971.

Department of Mathematics, London School of Economics and Political Science, Houghton Street, London WC2A2AE, U.K.

Department of Electrical and Electronics Engineering, Ariel University of Samaria, ARIEL 40700, ISRAEL