Quantifying Classification Performance using Combinatorial Geometry and Local Data Analysis

Christopher Lee, Mudassir Shabbir, and Waseem Abbas

***Abstract*—Understanding the theoretical limits of learning mechanisms and determining their fundamental capabilities remains a key challenge in machine learning. This paper presents a novel method for deriving optimal performance bounds for both linear and non-linear models by exploiting the combinatorial geometry of datasets. For a given dataset consisting of points in a d-dimensional Euclidean space, our approach uses local computations on small subsets of points—each comprising $(d+2)$ points—to predict the global performance of classifiers. We derive optimal training error bounds for linear classifiers by analyzing linear separability within these local subsets. For Support Vector Machines (SVMs), we establish margin bounds that align with their performance on similarly sized subsets. To address noisy and imprecise data, we extend our framework to analyze the linear separability of spheres in d-dimensional space. Furthermore, we generalize the method to non-linear classifiers by leveraging hypersphere boundary separation. Theoretical analysis and experimental results demonstrate that these bounds can be efficiently computed with a limited sampling of local subsets of data, enabling practical application of the method. By uncovering the geometric structure underlying data, this research provides critical insights into the predictive capabilities of machine learning models.**

***Index Terms*—Combinatorial geometry, Helly theorem, classification performance, Kirchberger theorem, support vector machines (SVMs)**

I. Introduction

HOW *can we rigorously evaluate the fundamental capabilities of a learning mechanism?* This question lies at the core of machine learning theory, which continues to revolutionize numerous fields. A pivotal concept addressing this question is the concept of VC-dimension (Vapnik–Chervonenkis dimension) [1]–[3]. The VC-dimension provides theoretical bounds on a classifier's performance by quantifying the maximum number of data points a model can "shatter", that is, classify correctly under all possible label arrangements. In essence, the VC-dimension measures a model's capacity to fit even the most intricate patterns within a dataset. For example, in a binary classification problem, where the goal is to draw an optimal decision boundary to separate data points into positive and negative classes, the VC-dimension of a hypothesis class or model indicates the largest number of data points that can be correctly classified, regardless of their arrangement, thereby capturing the core capabilities of a model's learning potential.

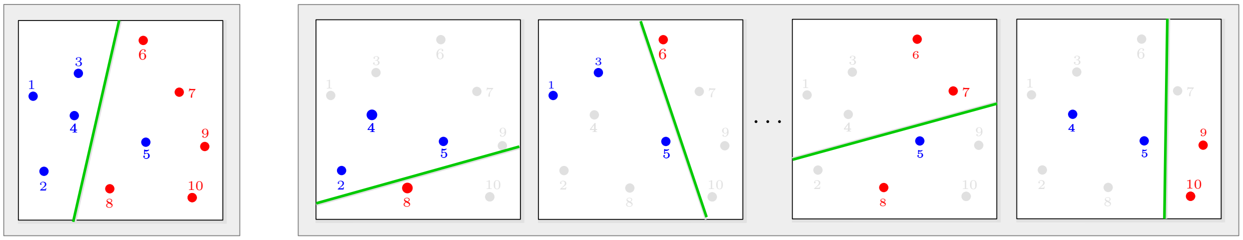
Despite its versatility, the VC-dimension does not incorporate the distribution of data points, which can pose challenges when using it as a practical measure to quantify the classification power of classifiers [4]–[9]. For example, the VC-dimension of a linear classifier is $d+1$, i.e., there exists a set of $d+2$ points with a particular labeling for which a linear classifier cannot learn the decision boundary of a binary classification problem. However, a linear classifier may effectively and successfully classify large datasets of points. The fact that some rare prohibitive point configurations exist is of little consequence in practice. Therefore, it is more desirable from a practical viewpoint to design a satisfactory bound that takes into account the arrangement of a given dataset.

To elaborate on this, consider a scenario where μ is an unknown probability distribution over a product set $\mathcal{X} \times Y$. Here, $\mathcal{X}$ is a metric space of (potentially very large) dimension d, representing the feature vectors of data points (henceforth, referred to as data points) and $Y = \{-1, +1\}$ denotes the class labels. Let $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ be independent samples drawn from μ. For simplicity, we focus on a binary classification problem, where $y_i \in \{\pm 1\}$. Within this framework, we address the following question: *Can one measure a realistic bound on the classification power of a classifier on this sample "efficiently"?* We assume that n is large enough to prohibit the *global* processing of the complete dataset at once. Instead, the processing relies on numerous *local* queries consisting of a small fraction of data points. Further, the sampling process might be noisy, making it impractical to train a model on these samples alone. Also, we focus mostly on linear classifiers in this paper. With these settings in mind, we ask the following question: *For a given set of samples (X_i, y_i) from an unknown distribution μ, is it possible to deduce from local computations on small subsets of samples whether there exists a linear classifier that correctly classifies every data point?* The answer to this question is in the affirmative. Informally, we know that if the data points are linearly classifiable *locally*, then they are also classifiable *globally*. This conclusion stems from Kirchberger's classical theorem in discrete geometry ([10]–[12]).

Theorem 1.1. ***(Kirchberger Theorem)*** *Given that $\mathcal{A}$ and $\mathcal{B}$ are compact subsets of Euclidean space, E^d, then for every subset $T \subseteq \mathcal{A} \cup \mathcal{B}$, with $|T| \leq d+2$, $T \cap \mathcal{A}$ and $T \cap \mathcal{B}$ can be strictly separated by a hyperplane if and only if $\mathcal{A}$ and $\mathcal{B}$ can be strictly separated by a hyperplane.*

Here, local computation refers to the task of evaluating whether a given $(d+2)$-sized subset of the dataset is linearly separable. Importantly, this computation size is independent of the overall dataset size n. For instance, if one has a

C. Lee is with the Electrical Engineering Department at the University of Texas at Dallas, TX, USA (Email: christopher.lee3@utdallas.edu). M. Shabbir is with the Computer Science Department at the Information Technology Univeristy, Lahore, Punjab, Pakistan (Email: mudassir.shabbir@rutgers.edu). W. Abbas is with the Systems Engineering Department at the University of Texas at Dallas, TX, USA (Email: waseem.abbas@utdallas.edu).



(a)

(b)

Fig. 1: (a) The best possible linear separator (green) for a dataset that is not perfectly linearly separable. (b) The quality of the best possible linear separator on the entire dataset can be determined from the linear separability of subsets of the dataset, each consisting of $d + 2$ points. Here, $d = 2$.

dataset comprising of a large number of data points in a reasonably large dimension d, the theorem indicates that the question of linear separability for the entire dataset can be decided by examining multiple subsets of just $d + 2$ points. Moreover, these computations can be naturally parallelized, offering efficiency in practical applications.

A. Our Results

In this work, we extend the above mentioned result in multiple significant directions. First, we consider the scenario where data is linearly separable. In such cases, the Support Vector Machine (SVM) is a canonical linear classifier with broad applicability across various domains (e.g., [13]–[20]). Specifically, the aim of the SVM is to find the classifier that maximizes the distance–known as the "margin"–between the separating hyperplane and the nearest data points, which are referred to as support vectors. The associated optimization problem is:

$$\textit{Minimize:} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{1}$$

$$\textit{Subject to:} \quad y_i(\mathbf{w} \cdot X_i + b) \geq 1, \quad \textit{for} \;\; i = 1, 2, \ldots, n \tag{2}$$

where, $\mathbf{w}$ represents the weight vector, b represents the bias term, and X_i and y_i are as previously defined. The *margin*, is calculated as $\frac{1}{\|\mathbf{w}\|}$. We address the following question regarding the value of margin an SVM algorithm may achieve for a particular dataset.

Given a linearly separable set of samples (X_i, y_i) and a constant w_0, is there a hyperplane $\mathbf{w}^T X - b = 0$ such that, $y_i(\mathbf{w} \cdot X_i + b) \geq 1$, and the $\frac{1}{\|\mathbf{w}\|} \geq w_0$?

In the context of this question, and in the spirit of the Kirchberger theorem, we prove:

Theorem 1.2. ***(SVM-Kirchberger)*** *Let $\mathcal{A}$ and $\mathcal{B}$ be disjoint, non-empty compact sets of E^d. Then $\mathcal{A} \cup \mathcal{B}$ is strictly linearly separable with margin w_0 if and only if for each subset $T \subset \mathcal{A} \cup \mathcal{B}$ of $d + 2$ or fewer points, there exists a linear SVM of margin w_0 that strictly separates $T \cap \mathcal{A}$ and $T \cap \mathcal{B}$.*

To address a more realistic scenario where the data may not be perfectly linearly separable, we introduce a significant advancement: a *fractional extension of the Kirchberger theorem*. This fractional version offers a way to quantify the margin of error—specifically, it provides a bound on the number of misclassified samples based on local data of $(d + 2)$ samples. The concept is illustrated in Figure 1. As a result, this allows us to infer a global property (misclassification) of the dataset based on local $(d + 2)$-sample data sets. The formal statement of our result is as follows:

Theorem 1.3. ***(Fractional Kirchberger)*** *Consider a dataset $\mathcal{A} \cup \mathcal{B} \subset E^d$, with $|\mathcal{A} \cup \mathcal{B}| = n$. For $\alpha \in (0, 1]$, if an α fraction of all $(d + 2)$-member subsets of $\mathcal{A} \cup \mathcal{B}$ are strictly linearly separable, then there exists a constant β, such that at least βn members of $\mathcal{A} \cup \mathcal{B}$ are also linearly separable. Moreover, $\beta \geq 1 - (1 - \alpha)^{1/(d+2)}$, and this bound on β is optimal.*

Proceeding further, we extend the fractional Kirchberger theorem into the realm of SVMs, and show the following:

Theorem 1.4. ***(SVM Fractional Kirchberger)*** *Let $\mathcal{A}$ and $\mathcal{B}$ be disjoint, non-empty compact sets of E^d, with $|\mathcal{A} \cup \mathcal{B}| = n$. For $\alpha \in (0, 1]$, assume an α fraction of the $(d+2)$-member subsets of $\mathcal{A} \cup \mathcal{B}$ can be linearly classified by an SVM algorithm with a margin of w_0. Then, there exists a constant $\beta(\alpha, d) > 0$ such that βn members of $\mathcal{A} \cup \mathcal{B}$ can be accurately classified by a soft-margin SVM with a margin of w_0.*

To clarify, a soft-margin SVM modifies the original constraints, allowing for some degree of misclassification. Specifically, the constraints become, $y_i(w \cdot X_i + b) \geq 1 - \xi_i$, where ξ_i is the slack variable, and the objective function includes an additional penalty term, $C \sum_{i=1}^{n} \xi_i$, to account for each misclassified point. Finally, we apply our methods to the case of *separation by hypersphere* in d-dimensions as an example of a non-linear classifier. We prove a fractional Kirchberger type theorem for the hypersphere separation.

Theorem 1.5. ***(Fractional Hypersphere Separation)*** *For any $\alpha \in (0, 1]$, and finite disjoint point sets $\mathcal{A}, \mathcal{B} \subset E^d$, if $\alpha\binom{n}{d+3}$ of the distinct $(d + 3)$-member subsets of $\mathcal{A} \cup \mathcal{B}$ are strictly spherically separable, then there exists a constant $\beta \in (0, 1]$ such that βn points of $\mathcal{A} \cup \mathcal{B}$ can be strictly separated by a hypersphere.*

Each of Theorems 1.2, 1.3, 1.4, and 1.5 leverages the combinatorial geometric properties of local data, specifically α, the proportion of linearly separable $(d + 2)$-sized subsets, to provide tight lower bounds on β, the number of correctly classified data points achievable by an optimal classifier on the global dataset. The significant practical implication of these results is that they offer a quantitative guarantee of classification performance without the need to undertake the relatively expensive task of training an SVM on a large dataset. While local tests of linear separability on small subsets are computationally simple, the number of subsets requiring tests grows prohibitively large with the size of the dataset. To address this issue, we demonstrate that the bounds on the global dataset can be accurately captured by sampling a *small proportion* of the local data subsets. We experimentally verify the effectiveness and fidelity of the resulting sample bounds. Our methods are versatile and capable of handling real-world complications such as imprecision in data, making them both theoretically and practically relevant. The remainder of the paper is organized as follows: We provide a review of the requisite background in Section II. In Section III we prove and discuss our main results. In Section IV, we demonstrate the computational practicality of their implementation, and Section V concludes the paper.

II. From Local to Global: A Discrete Geometric Perspective

In this section, we outline the preliminaries and background necessary to understand the key concepts underpinning our results. We begin by introducing point-hyperplane duality, a fundamental concept in our proofs, which elegantly transforms points into flat affine subspaces and vice versa. This concept is widely known as *point-line duality* in the literature (e.g., [21], [22]). To provide a clear understanding, we include a concise introduction with a small illustrative example in Section II-A. Our results also draw upon the fractional versions of the well-known Helly theorem from discrete geometry. To contextualize these theorems, we briefly review the the Fractional Helly Theorem in Section II-B. For further details, we refer the readers to [23]. Before discussing these prerequisites, we define some preliminary notions that will be employed throughout this paper. E^d denotes Euclidean space in d-dimensions.

Definition 2.1. ***(Hyperplane)*** *A hyperplane h in E^d is defined for real coefficients $a_1, a_2, \cdots, a_d, a_{d+1}$, not all identically equal to 0, as:*

$$h = \{x \in E^d : a_1x_1 + a_2x_2 + \cdots + a_dx_d + a_{d+1} = 0\}.$$

Definition 2.2. ***(Signed Halfspace)*** *For hyperplane $h \in E^d$, the positive open half-space h^+ is defined as:*

$$h^+ = \{x \in E^d : a_1x_1 + a_2x_2 + \cdots + a_dx_d + a_{d+1} > 0\},$$

and the negative open half-space h^- is defined as:

$$h^- = \{x \in E^d : a_1x_1 + a_2x_2 + \cdots + a_dx_d + a_{d+1} < 0 \}.$$

Definition 2.3. ***(Strict Linear Separability)*** *Let $\mathcal{A}$ and $\mathcal{B}$ be disjoint point sets in E^d. If $\mathcal{A}$ and $\mathcal{B}$ are strictly linearly separable, then then there exists a hyperplane h and associated open half-spaces h^+ and h^-, such that $\mathcal{A} \subset h^+$, $\mathcal{B} \subset h^-$.*

The term *fractional separability* will be used in this paper to refer to the ratio of correctly classified data points to the cardinality of the dataset as achieved by an optimal classifier. The optimal classifier maximizes this ratio among all possible classifiers of a given type.

A. Point-Hyperplane Duality

In the following sections, we will rely extensively on the point-hyperplane duality of projective geometry. When dealing with a set of points in a Euclidean space, referred to as the *primal space*, we can create another Euclidean space, the *dual space*. In this dual space, a unique relationship exists between points in the primal space and hyperplanes in the dual space, and vice versa. This duality transformation has two essential qualities: (a) *Preservation of incidences:* It ensures that the relationships or incidences between points and hyperplanes remain intact. (b) *Consistency in order:* It maintains the order of incidences, which can either be identical or opposite to that in the primal or reference space. This latter quality is particularly important for connecting separating hyperplanes to points in set intersections. Much of the literature on geometric duality focuses on the point-line duality mapping, denoted as $\pi : E^2 \mapsto E^{2*}$. In this mapping, each point p (or line l) in the primal space corresponds to a line $p*$ (or point $l*$) in the dual space, respectively, with the same incidence and order properties as mentioned earlier. In this paper, we adopt similar notation and language to describe duality transformations in E^d. Here, we introduce preliminary definitions and notation related to point-hyperplane duality, followed by an illustrative example.

Definition 2.4. ***(Duality Transform)*** *We define the duality transform $\mathcal{D} : E^d \mapsto E^{d+1}$ in the following manner for point $p = (p_1, p_2, \cdots, p_d) \in E^d$, and hyperplane $h = \{x \in E^d : a_1x_1 + a_2x_2 + \cdots + a_dx_d + a_{d+1} = 0\}$:*

$$\mathcal{D}(p) = p^* = \{x \in E^{d+1} : \langle x, p\rangle = 0\},$$

$$\mathcal{D}(h) = h^* = (a_1, a_2, \cdots, a_{d+1}) \in E^{d+1}.$$

Furthermore, incidences are preserved such that:

1) $p \in h \iff h^* \in p^*$
2) $p \in h^+ \iff h^* \in p^{*+}$ and $p \in h^- \iff h^* \in p^{*-}$

Definition 2.5. ***(Signed Duality Transforms)*** *We will use the notation $\mathcal{D}^+ : E^d \mapsto E^{d+1}$ to denote the positive dual transformation from a point $p \in E^d$ to a signed halfspace of E^{d+1} such that*

$$\mathcal{D}^+(p) = p^{*+} = \{x \in E^{d+1} : \langle x, p\rangle > 0\}$$

and $\mathcal{D}^- : E^d \mapsto E^{d+1}$ to denote the negative dual transformation from a point $p \in E^d$ to a signed halfspace of E^{d+1} such that

$$\mathcal{D}^-(p) = p^{*-} = \{x \in E^{d+1} : \langle x, p\rangle < 0\}.$$

To illustrate the utility of the above duality transforms in determining the equation of a separating hyperplane, we present the following example:

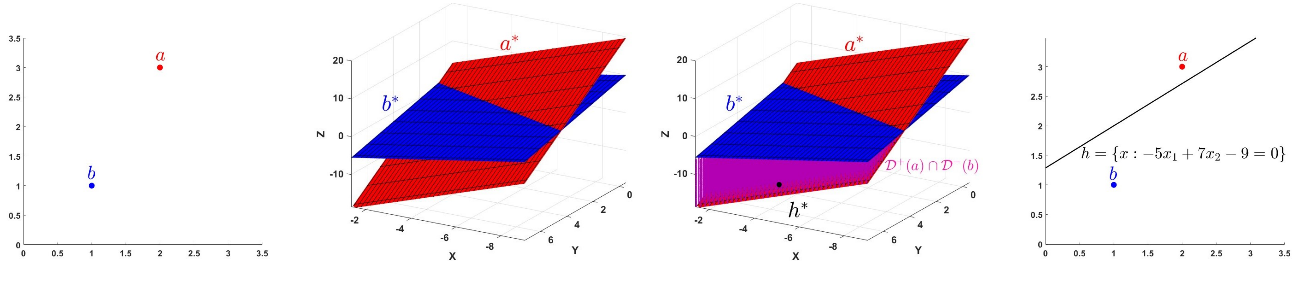
Example: Consider points $a = (2, 3)$ and $b = (1, 1)$ in E^2, as shown in Figure 2. We can find the equation of a separating hyperplane in the following way:

(i) The signed duality transform of a yields:

$$\mathcal{D}^+(a) = a^{*+} = \{x \in E^3 : 2x_1 + 3x_2 + x_3 > 0\}.$$

(ii) The signed duality transform of b yields:

$$\mathcal{D}^-(b) = b^{*-} = \{x \in E^3 : x_1 + x_2 + x_3 < 0\}.$$



(a) (b) (c) (d)

Fig. 2: (a) Points a and b in primal space E^2. (b) a^* and b^* in dual space E^3. (c) $h^* = (-5, 7, -9)$ is a point in the intersection of $\mathcal{D}^+(a)$ and $\mathcal{D}^-(b)$. (d) h is a linear separator for points a and b.

(iii) Select a point h^* such that $h^* \in \mathcal{D}^+(a) \cap \mathcal{D}^-(b)$. The point $h^* = (-5, 7, -9)$ satisfies this condition.

(iv) Take the dual transformation of h^* to produce $\mathcal{D}(h^*) = h = \{x \in E^2 : -5x_1 + 7x_2 - 9 = 0\}$.

(v) h strictly linearly separates a and b.

Set Notation – For convenience, we use $\mathcal{A}$ and $\mathcal{B}$ to denote the sets of data points in the positive class ($y_i = +1$) and negative class ($y_i = -1$), respectively. This notation will be used consistently without the loss of generality.

B. Fractional Helly

Central to this work is the relationship between the global properties of a dataset and the local properties of its subsets. In the field of combinatorial geometry, this relationship is the focus of problems known as "Helly-type" problems. Helly-type problems broadly address the global properties of a set of geometric objects that are implied by properties of its local subsets, typically of some fixed cardinality [24]–[26]. For example, Helly Theorem states that a set of convex objects in E^d have a non-empty intersection if and only if *all* of its $(d+1)$-member subsets have a non-empty intersection. The crux of the arguments presented in Section III will be based on a fractional variety of Helly's Theorem. Our exploration centers on scenarios where not all subsets of cardinality $(d+1)$ or fewer share a common intersection. Instead, only an α fraction of the $(d+1)$ sized subsets exhibit this property. In this section, we use α to denote the fraction of $(d+1)$-member subsets of a set of convex objects that share an intersection, and β to denote the fraction of the entire set that shares a common intersection. This concept was first elucidated by Liu and Katchalski [27], who demonstrated that for a given value of α, a lower bound can be guaranteed for β. Subsequently, Kalai proved an optimal value of β in terms of α and d as an application of his work on the Least Upper Bound Theorem [28], which we state here and refer to as the "Fractional Helly Theorem":

Theorem 2.1. ***(Fractional Helly)*** *Let $\mathcal{F}$ be a finite family of convex sets in E^d with $|\mathcal{F}| = n$ and $\alpha \in (0, 1]$. If at least $\alpha\binom{n}{d+1}$ of the subsets of size $d+1$ have non-empty intersection, then at least $\beta(\alpha, d) > 0$ members of $\mathcal{F}$ have a common intersection. In particular, $\beta = 1 - (1-\alpha)^{1/(d+1)}$.*

The derivation of this bound hinges on a fascinating insight involving the representation of a family of convex sets and their intersection patterns as a simplicial complex, often referred to as a 'nerve', within E^d. In this context, each set is represented as a vertex, and intersections between sets are depicted as edges. Notably, this bound is demonstrated to be optimal in the general case. However, the lower bound in Theorem 2.1 provides an asymptotic estimate as n goes to infinity. For a specific value of n, a more precise version of the Fractional Helly Theorem is implied by the Least Upper Bound Theorem [28].

Theorem 2.2. ***(Precise Fractional Helly)*** *Let $\mathcal{F}$ be a family of convex sets in $\mathbb{E}^d$, with $|\mathcal{F}| = n$. If there are $\alpha\binom{n}{k}$ intersecting k-tuples of $\mathcal{F}$, and $\alpha\binom{n}{k} > \sum_{i=0}^{d}\binom{r}{k-i}\binom{n-r}{i}$, then $\mathcal{F}$ has an intersecting subfamily of size at least $d + r + 1$.*

This result yields a tight bound on β when the size of $\mathcal{F}$ is given in addition to the value of α. In the real world scenarios, this value of β is most valuable. After proving Theorem 1.3 in Section III-A, we will show that with an additional constraint, the identical lower bound on the linear separability of point sets can be obtained in terms of the fraction of linearly separable subsets of size $d+1$ rather than $d+2$ using a refinement presented in [29], [30]. The original (non-fractional) refinement of Kirchberger's Theorem due to Houle [11] is as follows:

Theorem 2.3. ***(Refined Kirchberger)*** *Let $\mathcal{A}$, $\mathcal{B}$ denote disjoint, non-empty, families of convex sets of E^d. For an arbitrary non-vertical hyperplane, h, denote by h^+ the upper halfspace and h^- the lower halfspace bounded by h. $\mathcal{A}$ and $\mathcal{B}$ are strictly separable by hyperplane if and only if every $f \subset \mathcal{A} \cup \mathcal{B}$ with $|f| = d + 1$, is strictly separable by a hyperplane h, such that $f \cap \mathcal{A} \subset h^+(h^-)$ and $f \cap \mathcal{B} \subset h^-(h^+)$.*

With this background, we proceed to our main results.

III. Tight Bounds on Linear Classification Performance via Local Data Analysis

The results presented in Section I offer precise lower bounds on a classifier's performance for a given binary class dataset. This is achieved by analyzing a statistic, α, which represents the proportion of linearly separable $(d+2)$-sized subsets of a given datasets. Surprisingly, the knowledge of this local statistic alone allows us to apply a sophisticated insight into the intersection patterns of set systems (Theorem 2.1) to derive a constraint on the global dataset, specifically, the minimum number of points from either class that are correctly classified by an optimal classifier. In this section we present the proofs of the theorems stated in Section I and elaborate on their data science applications. We begin with the fractional version of Kirchberger's theorem, Theorem 1.3, in Section III-A.

Kirchberger's theorem asserts that if a binary class dataset has the property that all of its $(d+2)$-member subsets are linearly separable, then the entire dataset is linearly separable. The result we will now present, the fractional version of Kirchberger's theorem, applies to the scenario in which only a fraction of the $(d+2)$-member subsets are linearly separable. Subsequently, we will present some simulation results for illustration. Subsequently, we present the proof of Theorem 1.2, an extension of Kirchberger's Theorem to the realm of strict linear separation with a margin in Section III-B and show that it has direct relevance to the case of "noisy" data, wherein points of a dataset are subject to bounded error. Moving forward in Section III-B, we prove Theorem 1.4, which provides a lower bound on the performance of a soft-margin SVM classifier—a fractional counterpart to Theorem 1.2. In Section III-C we prove the extension of the fractional bound to classification by hypersphere. A notable feature unifying all the theorems presented in this section, and one that the forthcoming proofs substantiate, is their remarkable dependency solely on fundamental dataset properties: size, dimensionality, and the average performance of linear and hypersphere classifiers on subsets of cardinality $(d+2)$ and $(d+3)$, respectively.

A. Fractional Kirchberger Theorem

Here, we harness the point-hyperplane duality relation (as in Section II-A) in conjunction with the Fractional Helly Theorem II-B. To prove Theorem 1.3, we establish a vital connection: the lower bound on set intersections, as provided by the Fractional Helly Theorem, is equivalent to a lower bound for the number of linearly separable points in a binary class dataset. The essential connection is made by observing that for a point p of a given class, the union of all separating hyperplanes that correctly classify p forms a convex set under the duality transformation. The problem of proving the existence of a common separating hyperplane is thereby re-framed as a problem of proving the existence of a common point of intersection amongst convex sets in the dual space, for which the solution is provided by the Fractional Helly Theorem. Before proving the main result, we formally establish the correspondence between common intersections in the dual space and separating hyperplanes (as demonstrated in the example in Section II-A). We make a simplifying assumption, without loss of generality, that our objective is to find a hyperplane placing $\mathcal{A}$ in its corresponding negative open half-space and and $\mathcal{B}$ in the positive open half-space.

Lemma 1. *For any $A \subseteq \mathcal{A}$ and $B \subseteq \mathcal{B}$, A and B are strictly linearly separated by a hyperplane h in E^d if and only if the corresponding dual point h^* in the transformed space E^{d+1}, lies in the intersection of all negative dual transforms of points in A, as well as, in the intersection of positive dual transforms of points in B, i.e., $h^* \in \bigcap_{a\in A} \mathcal{D}^-(a)$, $\quad h^* \in \bigcap_{b\in B} \mathcal{D}^+(b)$.*

Proof. If the transformed point $h^* \in \bigcap_{a\in A} \mathcal{D}^-(a)$, in E^{d+1}, then by Definition 4, we have that, each point $a \in A$, $a_1h_1 + a_2h_2 + \cdots + a_dh_d + h_{d+1} < 0$. Similarly, for each point $b \in B$, we have that, $b_1h_1 + b_2h_2 + \cdots + b_dh_d + h_{d+1} > 0$. Recall that $h = x_1h_1 + x_2h_2 + \cdots + x_dh_d + h_{d+1} = 0$, in E^d. Thus, $A \subset h^-$ and $B \subset h^+$, in the primal space, and therefore, h strictly separates $A \cup B$. Alternatively, if $A \cup B$ are strictly separated by a hyperplane h', then for $h' = \{x \in E^d : h_1x_1+h_2x_2+\cdots+h_dx_d+h_{d+1} = 0\}$, Definitions 2 and 3 assert that for all $a \in \mathcal{A}$, $h_1a_1+h_2a_2+\cdots+h_da_d+h_{d+1} < 0$, and for all $b \in \mathcal{B}$ $h_1b_1+h_2b_2+\cdots+h_db_d+h_{d+1} > 0$. Then by Definition 6, $h'^* \in \mathcal{D}^-(a)$ and $h'^* \in \mathcal{D}^+(b)$. In words, if point a is below hyperplane h' and point b is above hyperplane h' in E^d then hyperplane a^* is below point h'^* and hyperplane b^* is above point h'^* in E^{d+1}. Therefore, h'^* is a point in the intersection of negative/positive halfspaces bounded by a^* and b^*. ∎

We now prove one of our key results, that establishes a precise lower bound on the misclassification error for any linear classifier. This result directly relates to a fundamental aspect of support vector machines. At a high level, it provides a simple way to quantify the extent to which a dataset is non-linearly separable. We now use α to denote the fraction of linearly separable $(d+2)$-member subsets of a binary class dataset and β to denote the greatest fraction of the entire dataset that is linearly separable.

Theorem 1.3. ***(Fractional Kirchberger)*** *Consider a dataset $\mathcal{A} \cup \mathcal{B} \subset E^d$, with $|\mathcal{A} \cup \mathcal{B}| = n$. For $\alpha \in (0, 1]$, if an α fraction of all $(d+2)$-member subsets of $\mathcal{A} \cup \mathcal{B}$ are strictly linearly separable, then there exists a constant β, such that at least βn members of $\mathcal{A} \cup \mathcal{B}$ are also linearly separable. Moreover, $\beta \geq 1 - (1-\alpha)^{1/(d+2)}$, and this lower bound on β is tight.*

Proof. The duality transform can be applied to $\mathcal{A} \cup \mathcal{B}$ to obtain the family of halfspaces $C = \mathcal{D}^-(\mathcal{A}) \cup \mathcal{D}^+(\mathcal{B})$. Thus, C contains n halfspaces of E^{d+1} that are in one to one correspondence with the points of $\mathcal{A} \cup \mathcal{B}$. Let f denote an arbitrary $(d+2)$-member subset of $\mathcal{A} \cup \mathcal{B}$. By Lemma 1, if f admits a strict linear separation, then $\bigcap \mathcal{D}^-(f \cap \mathcal{A}) \cap \mathcal{D}^+(f \cap \mathcal{B})$ is non-empty. If there are $\alpha\binom{n}{d+2}$ such $(d+2)$-member subsets of $\mathcal{A} \cup \mathcal{B}$, then there are $\alpha\binom{n}{d+2}$ intersecting $(d+2)$-tuples of C. It follows from the Fractional Helly Theorem, that there are at least βn halfspaces of C that share a common intersection. Then by Lemma 1, the dual of a point $h^* \in E^{d+1}$ from this intersection produces a hyperplane $h \in E^d$ that strictly separates at least βn members of $\mathcal{A} \cup \mathcal{B}$. ∎

Here, β (as in Theorem 2.1) is in fact an asymptotic bound that holds for all n. However, we may refine this bound when we wish to assess the fractional linear separability of $\mathcal{A} \cup \mathcal{B}$ for a specific value of n. As established in the proof of Theorem 1.3, linear separators correspond to intersection points among the halfspaces of $\mathcal{A} \cup \mathcal{B}$. Therefore, we can readily apply the bound on the greatest common intersection from Theorem 2.2 to the case of fractional linear separability. Once we determine the value of α, which signifies the fraction of strictly linearly separable $(d+2)$-tuples in E^d, we can conclude that there are αn intersecting dual halfspaces in E^{d+1}. Now, considering $d' = d+1$ and $k = d'+1$, Theorem 2.2 can be applied, resulting in:

$$\beta = \frac{r + d' + 1}{n}, \tag{3}$$

where r is determined as:

$$r = \arg\max_{r} \left\{ r \mid \sum_{i=0}^{d'} \binom{r}{k-i}\binom{n-r}{i} < \alpha \binom{n}{k} \right\}. \tag{4}$$

In this manner, we leverage the knowledge of n, the size of the dataset, to establish a precise lower bound on the fractional linear separability of $\mathcal{A} \cup \mathcal{B}$. To demonstrate the tightness of the lower bound provided by Theorem 1.3 in relation to the optimal linear separator, we conducted a series of experiments. In these experiments, we randomly placed points in a d-dimensional hypercube, assigning each point a random label of either $y_i = 1$ or $y_i = -1$ with equal probability. We repeated this process for $n = 20$ points across 5000 trials. For each trial, we performed and recorded the following computations:

1) We calculated α by examining each $(d{+}2)$-tuple of points to test for linear separability.
2) Using α, we derived the theoretical lower bound of β using Equations (3) and (4).
3) We determined the ground truth value of β by identifying an optimal linear separator minimizing the misclassification count. In Figure 3, the true value of β is referred to as β_{actual}.

The procedure was conducted separately for $d = 2, 3$, and Figure 3 presents the results.

As stated in Section II, if a hyperplane exists that places the points of $\mathcal{A}$ in the positive half-space and the points of $\mathcal{B}$ in the lower half-space, a refined lower bound can be obtained. The refinement is based on the proportion of $(d+1)$-tuples that are *consistently* linearly separable. This is stated precisely in the following "refined" Fractional Kirchberger Theorem:

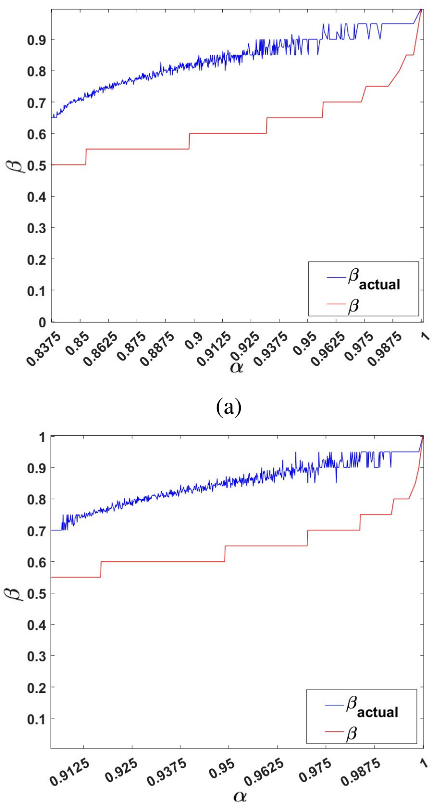
Theorem 3.1. ***(Refined Fractional Kirchberger)*** *Let $\alpha \in (0,1]$ and $\mathcal{A}, \mathcal{B} \subset E^d$ denote disjoint, non-empty point sets, with $|\mathcal{A} \cup \mathcal{B}| = n$, and let $\mathcal{F}$ denote the family of all $(d+1)$-subsets of $\mathcal{A} \cup \mathcal{B}$. If $\alpha|\mathcal{F}|$ of the $(d+1)$-member subsets can be strictly linearly separated by a hyperplane h, such that for $f \in \mathcal{F}$, $f \cap \mathcal{A} \subset h^+$ and $f \cap \mathcal{B} \subset h^-$(or vice versa), then there exists a constant $\beta(\alpha, n) = 1 - (1-\alpha)^{1/(d+1)}$, such that at least βn members of $\mathcal{A} \cup \mathcal{B}$ are strictly separable by a hyperplane.*

To prove the refined version of the fractional Kirchberger, theorem we introduce another duality transformation that maps points and hyperplanes to their duals in a space of the *same* dimension. We will distinguish this second duality transformation with the notation $\mathcal{D}_2$.

Definition 3.1. ***(Duality Transform 2)*** *The duality transformation $\mathcal{D}_2 : E^d \mapsto E^d$ is defined in the following manner for point $p \in E^d$, and hyperplane h:*

$$p = (p_1, p_2, \cdots, p_d) \mapsto p^* = \{x \in E^d \mid \langle x, p \rangle = 1\}$$

$$h = \{x \in E^d \mid a_1x_1 + a_2x_2 + \cdots + a_dx_d = 1\} \mapsto$$



(a)

(b)

Fig. 3: (a) β comparison in E^2. (b) β comparison in E^3.

$$h^* = (a_1, a_2, \cdots, a_d)$$

Note that $\mathcal{D}_2$ preserves incidences in the same manner as $\mathcal{D}$. We will use the same notation for signed halfspaces :

1) $p \in h \iff h^* \in p^*$
2) $p \in h^+ \iff h^* \in p^{*+}$ and $p \in h^- \iff h^* \in p^{*-}$

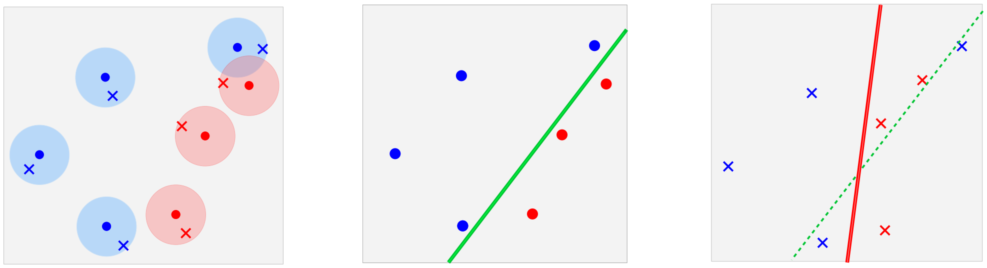
We now give the formal proof of Theorem 3.1.

Proof. Without loss of generality we can assume that each $(d{+}1)$-member subset f that has a strict separating hyperplane h is such that $f \cap \mathcal{A} \in h^+$ and $f \cap \mathcal{B} \in h^-$. Then for each of the points $a \in f \cap \mathcal{A}$ and $b \in f \cap \mathcal{B}$, and hyperplane h, we may consider may consider their duality transforms as follows:

1) $a^{*+} = \mathcal{D}_2^+(a)$
2) $b^{*-} = \mathcal{D}_2^-(b)$
3) $h^* = \mathcal{D}_2(h)$

It follows from definition 15 that $h^* \in a^{*+} \cap b^{*-}$. By observing that each f, a^{*+} and b^{*-} are convex sets with non-empty intersection, it follows from the fractional Helly theorem there are at least $\beta(\alpha, n) = 1 - (1-\alpha)^{1/(d+1)}$ members of $\mathcal{D}_2^+(\mathcal{A}) \cup \mathcal{D}_2^-(\mathcal{B})$ with nonempty intersection. Let v^* be a point in this intersection. Then v is a hyperplane that strictly separates βn points of $\mathcal{A} \cup \mathcal{B}$. ■

The primary practical advantage of this refined bound with respect to assessing the classification potential of a dataset via local computations is to reduce the number of computations from $\binom{n}{d+2}$ to $\binom{n}{d+1}$, as well as marginally reducing the complexity of each computation (testing linear separability).



(a) (b) (c)

Fig. 4: (a). Red points and blue points surrounded by bounded areas representing the set of possible positions after perturbation. '$\times$'s represents one possible configuration after perturbation. (b) The original point set is linearly separable with margin w_0. (c) The perturbed set is not linearly separable. The solid line represents the decision boundary for the perturbed points and the dashed line represents the decision boundary for the original set.

B. Hard-Margin and Soft-Margin SVM Classification

A dataset may contain points that have been perturbed or contain errors, causing them to deviate slightly from their true values. Such errors can arise from factors like faulty hardware, transmission losses, or other sources of noise, leading to bounded deviations from the original data points. As illustrated in Figure 4, training an SVM on a dataset with bounded errors can produce a decision boundary that differs significantly from the one obtained on the unperturbed dataset. This highlights the importance of developing a measure of classification performance that is robust to noise. By accounting for errors bounded by w_0, data points can be interpreted as *d-dimensional balls* with radius w_0. Under this interpretation, the results presented in this work provide a robust measure of classification performance that accommodates such noise.

Here we present a result that establishes a crucial equivalence: the local conditions required for ensuring perfect linear separation with a specified margin, specifically the distance between a separating hyperplane and the closest data points to that hyperplane, remain identical to those articulated in Kirchberger's theorem. The key distinction lies in replacing 'linearly separable' with 'linearly separable with a margin of w_0, as illustrated in Figure 5. This leads us to a significant conclusion: Kirchberger's theorem applies not only to strict linear separation but also to strict linear separation with a margin, making it directly applicable to support vector machines. We state this result below (also mentioned in [12] without a proof), since it will also be instrumental in proving Theorem 1.4.

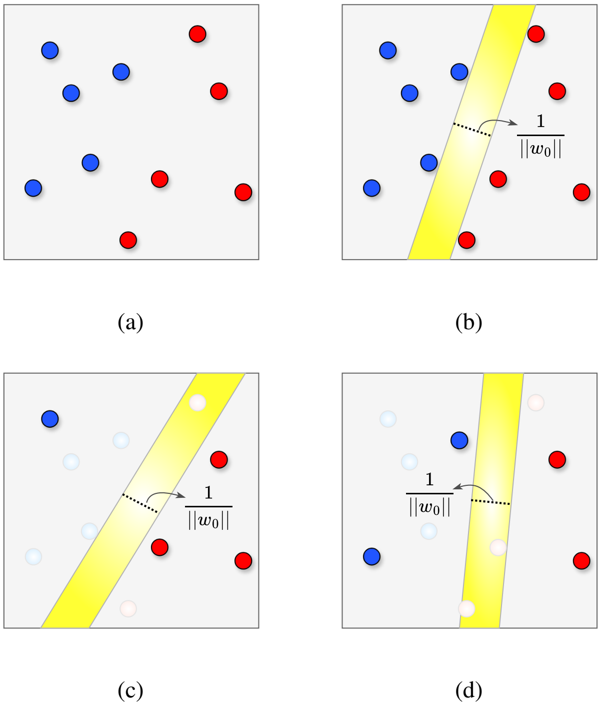
Theorem 1.2. ***(SVM-Kirchberger)*** *Let $\mathcal{A}$ and $\mathcal{B}$ be disjoint, non-empty compact sets of E^d. Then $\mathcal{A} \cup \mathcal{B}$ is strictly linearly separable with margin w_0 if and only if for each subset $T \subset \mathcal{A} \cup \mathcal{B}$ of $d+2$ or fewer points, there exists a linear SVM of margin w_0 that strictly separates $T \cap \mathcal{A}$ and $T \cap \mathcal{B}$.*

Proof. We will prove the non-trivial implication of the theorem: If there exists a linear SVM of margin w_0 for every $(d+2)$-point subset $U \subset \mathcal{A} \cup \mathcal{B}$, then there is a linear SVM of margin w_0 for $\mathcal{A} \cup \mathcal{B}$. Since each U is linearly classifiable with margin w_0 then for all such U there is a hyperplane $\mathbf{w}$, such that $\forall a \in U \cap \mathcal{A}, \langle \mathbf{w}, a \rangle \leq 1$ and $\forall b \in U \cap \mathcal{B}, \langle \mathbf{w}, b \rangle \geq 1$, with $\frac{1}{\|\mathbf{w}\|} \geq w_0$. This condition is equivalent to the following: $\forall u \in U, \forall x \in \mathbf{w}$, $\min(d(u,x)) \geq \frac{w_0}{2}$ and $\forall a \in U \cap \mathcal{A}$, $a \in \mathbf{w}^-, \forall b \in U \cap \mathcal{B}, b \in \mathbf{w}^+$, where $\mathbf{w}^{+/-}$ is the positive/negative halfspace bounded by $\mathbf{w}$. Consider the modified sets $\mathcal{A}'$ and $\mathcal{B}'$ such that $\mathcal{A}' = \mathcal{A} + \{x \in E^d : \|x\| < \frac{w_0}{2}\}$ and $\mathcal{B}' = \mathcal{B} + \{x \in E^d : \|x\| < \frac{w_0}{2}\}$. In other words, the modified sets consist of open balls of radius $\frac{w_0}{2}$ centered at the points of the original set. For each U', where U' is a $(d+2)$-member subset of $\mathcal{A}' \cup \mathcal{B}'$, $(\min(d(u',x)) \geq \min(d(u,x)) - \frac{w_0}{2} > 0$ for $u' \in U', x \in \mathbf{w}$. Furthermore, since the original points of U are correctly classified by $\mathbf{w}$ with minimum distance $\frac{w_0}{2}$ from $\mathbf{w}$, then $U' \cap \mathcal{A}' \subset \mathbf{w}^-, U' \cap \mathcal{B}' \subset \mathbf{w}^+$, and therefore $\mathbf{w}$ strictly separates U'. Since this is true for all U', then by Kirchberger's Theorem there exists hyperplane $\mathbf{w'}$ that strictly separates $\mathcal{A}' \cup \mathcal{B}'$. It is clear that that $\mathbf{w'}$ also strictly separates $\mathcal{A} \cup \mathcal{B}$. Additionally, we have that $\min(d(a,x)) \geq \frac{w_0}{2}$, and $\min(d(b,x)) \geq \frac{w_0}{2}$ for all $a \in \mathcal{A}, b \in \mathcal{B}, x \in \mathbf{w'}$. ∎

Next, we utilize the Fractional Kirchberger Theorem to establish a fractional counterpart to Theorem 1.2. Essentially, we provide a lower bound on the performance of a soft-margin SVM when not all samples from the dataset $\{(X_i, y_i)\}$ can be linearly classified with a margin of w_0. As previously, we maintain the notations $\mathcal{A}$ and $\mathcal{B}$ to denote the two classes within data.

Theorem 1.4. ***(SVM Fractional Kirchberger)*** *Let $\mathcal{A}$ and $\mathcal{B}$ be disjoint, non-empty compact sets of E^d, with $|\mathcal{A} \cup \mathcal{B}| = n$. For $\alpha \in (0,1]$, assume an α fraction of the $(d+2)$-member subsets of $\mathcal{A} \cup \mathcal{B}$ can be linearly classified by an SVM algorithm with a margin of w_0. Then, there exists a constant $\beta(\alpha, d) > 0$ such that βn members of $\mathcal{A} \cup \mathcal{B}$ can be accurately classified by a soft-margin SVM with a margin of w_0.*

Proof. Since each of the $\alpha\binom{n}{d+2}$ of the $(d+2)$-member sets of $\mathcal{A} \cup \mathcal{B}$ admit a linear separator with margin w_0, they are linearly separable, then Theorem 2 implies the existence of β, for which there exists some $U \subset \mathcal{A} \cup \mathcal{B}$, with $|U| \geq \beta n$ such that $conv(U \cap \mathcal{A}) \cap conv(U \cap \mathcal{B}) = \emptyset$. Here, $conv(*)$ denotes the convex hull. To simplify the following argument, we may assume that each $(d+2)$-member subset of $\mathcal{A} \cup \mathcal{B}$ that are not separable with margin w_0 are also not linearly separable. Observe that the lower bound of β would remain valid with or without our simplifying assumption, which is to say that the $(d+2)$-tuples of $\mathcal{A} \cup \mathcal{B}$ that are separable



(a) (b)

(c) (d)

Fig. 5: (a) Original set. (b) Original set, linearly separable with margin w_0. (c) and (d) $(d+2)$-tuples linearly separable with margin w_0.

with margin w_0 but are possibly linearly separable have no representation in U. By this assumption, if a set of points is linearly separable, it is linearly separable with margin w_0. Since U is linearly separable, then by Kirchberger's Theorem, every $(d+2)$-member subset of U must be linearly separable and therefore by our assumption must be separable with margin w_0. Then we observe that there are some closest pair of points $(a_{min}, b_{min}) : \{a_{min} \in U \cap \mathcal{A}, b_{min} \in U \cap \mathcal{B}\}$ such that $d(a_{min}, b_{min}) > \delta$, and the remainder of the proof follows identically to that of Theorem 1.2. ■

Theorems 1.2, 1.3, and 1.4 primarily establish that the statistic of separable $(d+2)$-sized subsets gives a lower bound on the performance of an optimal classifier. Thus we have demonstrated "local-to-global" implications for linear classifiers that provide insight into the linear separability of a particular dataset. Furthermore, the bound is obtained in an *efficient* manner, that is in terms of the complexity of individual computations. In Section IV we will show that the number of local computations required to arrive at an accurate global lower bound will generally be much smaller than $\binom{n}{d+2}$ (or even $\binom{n}{d+1}$, as in Theorem 3.1). Before addressing the tradeoff between the computational efficiency and accuracy of the proposed fractional bounds we establish an extension of this bound to the performance of non-linear classifiers.

C. Non-linear Classification

Thusfar, we have focused on linear classifiers under the framework of the "local-to-global" implications characteristic of Helly-type problems. However, this framework also applies to non-linear separators. We will now present the proof of Theorem 1.5, which uses this framework to provide lower bounds on the classification performance of a canonical non-linear example, the spherical classifier, on a given dataset in terms of the spherical separability of its local subsets. Specifically, the lower bound is based on a spherical classifier's performance with samples of size $d+3$. As with linear separation, the local conditions for achieving perfect (no misclassiciations) spherical separability of the *entire* dataset are well-known in the literature [12], [31], [32]. In this work, we modify these conditions to apply to the *fractional* setting. We begin by providing a formal definition of spherical separability.

Definition 3.2. ***(Spherical Separability)*** *Given point sets $\mathcal{A}$, $\mathcal{B}$ in E^d. $\mathcal{A} \cup \mathcal{B}$ are strictly separable by hypersphere $h_s = \{x \in E^d : \|x - p\| = \gamma\}$ if for $a \in \mathcal{A}$, $\|a - p\| < \gamma$ and for $b \in \mathcal{B}$, $\|b - p\| > \gamma$, or vice versa.*

A simple proof involves stereographically projecting E^d onto a tangent hypersphere of E^{d+1} (as Figure 6 illustrates), effectively transforming the problem into one of linear separation in E^{d+1} (hence the $d+3$ requirement). We consider the point set $\mathcal{A} \cup \mathcal{B}$, where not all of its $(d+3)$-member subsets exhibit strict spherical separability. To extend this result to the fractional case, we demonstrate that when only $\alpha\binom{n}{d+3}$ of $(d+3)$-point samples can be correctly classified by a hypersphere, we can apply the Fractional Helly Theorem to the dual of the projected points. This yields a lower bound on the size of the largest subset with a common intersection (βn). Through the duality transformation, this lower bound on the intersection number in the dual space corresponds to a lower bound on the number of points in the original dataset that are accurately classified by a hypersphere in the primal space.

Theorem 1.5. ***(Fractional Hypersphere Separation)*** *For any $\alpha \in (0, 1]$, and finite disjoint point sets $\mathcal{A}, \mathcal{B} \subset E^d$, if $\alpha\binom{n}{d+3}$ of the distinct $(d+3)$-member subsets of $\mathcal{A} \cup \mathcal{B}$ are strictly spherically separable, then there exists a constant $\beta \in (0, 1]$ such that βn points of $\mathcal{A} \cup \mathcal{B}$ can be strictly separated by a hypersphere.*

Proof. Let T be a subset of $d+3$ points in E^d. Consider the embedding of E^d into a hyperplane h of E^{d+1}. Let S be a $(d+1)$-dimensional hypersphere tangent to h at an arbitrary point p. Let p_0 denote the antipodal point to p. Then there is a bijective map, π, that sends each point $z \in E^d$ to $S \setminus \{p_0\}$, where r is the ray originating from z and passing through p_0. If T is strictly spherically separable in E^d, then a d-dimensional hypersphere h_s exists such that all points in $T \cap \mathcal{A}$ lie in h_s^{int} and all points in $T \cap \mathcal{B}$ lie in h_s^{ext}, where h_s^{int} and h_s^{ext} denote the interior and exterior of h_s, respectively. Considering the projections of T and h_s onto S, it follows that $\pi(h_s)$ is contained in the intersection of S with a hyperplane $H_s \in E^{d+1}$ such that $\pi(h_s^{int})$ and $\pi(h_s^{ext})$ are strictly linearly separated by H_s. Observe that the projections of a set of $d+3$ points in E^d corresponds to a set of $d'+2$ points in $E^{d'}$ for $d' = d+1$. Since there are $\alpha\binom{n}{d'+2}$ such subsets of projected points that are strictly linearly separable in $E^{d'}$, then by Theorem 1.3 there exists a hyperplane $H'_s \subset E^{d'}$ that strictly linearly separates the projections of βn members of

$\mathcal{A} \cup \mathcal{B}$. Thus, $\pi^{-1}(H'_s)$ is a hypersphere in E^d that satisfies the claim. ■

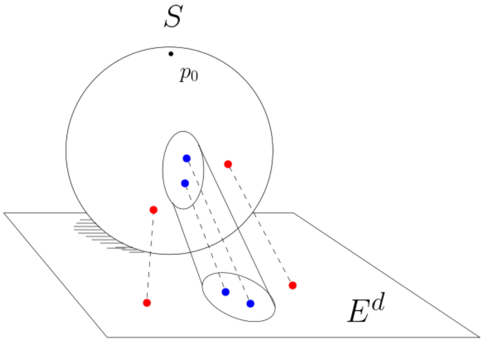


Fig. 6: Stereographic projection of E^d onto S.

We now move on to address how an approximate lower bound on a dataset's classification potential may be efficiently obtained while retaining a high degree of accuracy.

IV. Computationally Efficient Approximations of Linear Separability

In the preceding sections, we explored the problem of linear and nonlinear separability under various conditions, demonstrating that the separability of the small subsets of a dataset correlates strongly with the overall separability of the dataset. We also established that this correlation is theoretically optimal in the sense that it provides a tight lower bound on the number of correctly classified points of the dataset. From an algorithmic standpoint, applying these bounds in practice to determine the linear separability of large, high-dimensional datasets would necessitate examining the separability of all $\binom{n}{d+1}$ subsets, where n is the number of data points in d dimensions. Often, this requirement imposes prohibitively high computational costs for achieving an optimal measure of linear separability, denoted as α. In this section, we demonstrate that it is possible to approximate the optimal value of α at a significantly lower computational cost. Additionally, we provide a theoretical framework to evaluate the quality of such approximations, offering a trade-off between computational resources and approximation accuracy. Formally, we state the following result.

Theorem 4.1. *For a set S of n data points in d-dimensional space, each belonging to one of two different classes, let α be the fraction of $\binom{n}{d+1}$ subsets in which points from one class are linearly separable from the points of the other class. For any two given positive constants, ϵ and δ, only $\phi \geq \frac{2+\epsilon}{\epsilon^2} \ln(2/\delta)$ uniformly chosen samples out of all $(d+1)$-tuples are sufficient to compute an ϵ approximation of α with probability at least δ, i.e.,*

$$\mathbb{P}(|\alpha_\phi - \alpha| \geq \epsilon) \leq 2 \exp\left(-\frac{\phi\epsilon^2}{2+\epsilon}\right),$$

where α_ϕ is the fraction of ϕ uniformly selected $d+1$ subsets that are can be linearly separated.

Proof. Let $\epsilon, \delta, \phi, \alpha$, and α_ϕ be as defined in the statement of the theorem. We observe that linear separability of a randomly chosen $d+1$-tuple T can be represented by a Bernoulli random variable γ_T, where, γ_T is set to one when a tuple is linearly separable and zero otherwise. We have,

$$\sum_{T \subseteq S, |T|=d+1} \gamma_T = \alpha.$$

Therefore, α_ϕ, is a random variable with Binomial distribution. Further, α_ϕ is an unbiased estimator, i.e.,

$$\mathbb{E}(\alpha_\phi) = \alpha.$$

Further note that the Chernoff upper bound [33] on the ϵ tail of a Binomial random variable α_ϕ is given by,

$$\mathbb{P}(|\alpha_\phi - \mathbb{E}(\alpha_\phi)| \geq \epsilon) \leq \exp\left(-\frac{2\phi\epsilon^2}{2+\epsilon}\right),$$

where ϕ is the number of Bernoulli samples. We have,

$$\mathbb{P}(|\alpha_\phi - \alpha| \geq \epsilon) \leq 2 \exp\left(-\frac{\phi\epsilon^2}{2+\epsilon}\right),$$

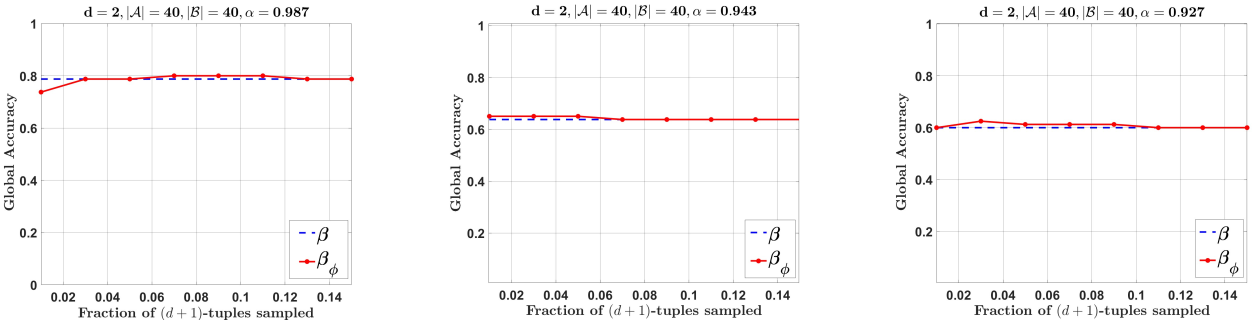
as required. ■

Theorem 4.1 provides us with the crucial ability to accurately estimate α based on the practical value of the proposed bound and to determine a sufficient sample size to ensure that α is estimated within a factor of ±ϵ with probability $1-\delta$. For any required ϵ and δ, the sample size is given by,

$$\phi \geq \frac{2+\epsilon}{\epsilon^2} \ln(2/\delta) \tag{5}$$

To illustrate the effectiveness of our approach, we conducted an experiment with $n = 80$ points in a two-dimensional plane ($d = 2$), where $|\mathcal{A}| = |\mathcal{B}| = 40$. The positions of the 80 points were randomly assigned within a unit square and class labels were also randomly assigned. For various values of ϕ, we randomly selected ϕ of the $(d+1)$-tuples from the total of $\binom{n}{d+1}$ such tuples. We then computed α_ϕ, defined as the proportion of the selected ϕ samples that were linearly separable. Using α_ϕ and d, we determined β_ϕ by substituting these values into Equation (4). Additionally, we computed the true values of α and β by evaluating all $\binom{n}{d+1}$ tuples for linear separability. The results are presented in Figure 7. The key observation from these experiments is that β_ϕ closely matches β across all tested cases, even when ϕ is a small fraction of the total $\binom{n}{d+1}$ tuples. For example, in Figure 7(a), we compute α by checking the linear separability of all $\binom{n}{d+1} = 82,160$ tuples of size $d+1$, and subsequently determine β using Equation (4). Next, we randomly sample ϕ of these $\binom{n}{d+1}$ tuples to compute α_ϕ and then β_ϕ. Remarkably, β_ϕ that is obtained by sampling only 13% of the total $\binom{n}{d+1}$ tuples (i.e., $\phi = 0.13 \times \binom{n}{d+1}$), is identical to β. This result highlights the efficiency of the proposed method, as it achieves accurate approximations of β while using a significantly reduced number of $(d+1)$-tuples.

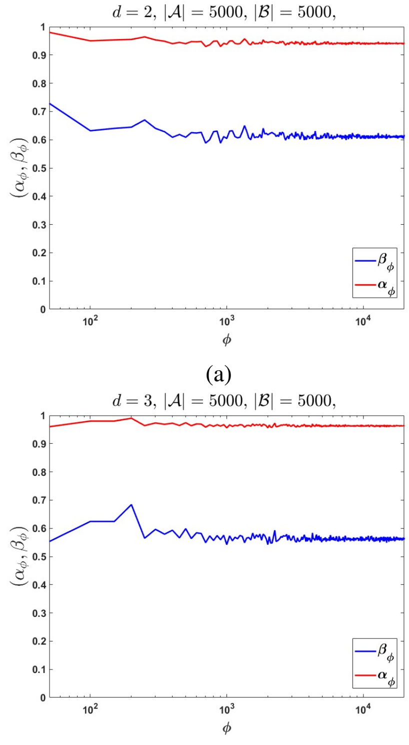
To experimentally demonstrate that β_ϕ converges as a function of ϕ alone, we conducted an experiment with $n = 10,000$ ($|\mathcal{A}| = |\mathcal{B}| = 5,000$), first for $d = 2$ and then for $d = 3$. The positions and labels of each point were generated randomly, as described earlier. We then computed the sample estimate α_ϕ and the corresponding bound estimate β_ϕ. The results,



(a) (b) (c)

Fig. 7: β vs. β_ϕ for (a) $\alpha = 0.987$, (b) $\alpha = 0.943$, (c) $\alpha = 0.927$. The x-axis represents $\frac{\phi}{\binom{n}{d+1}}$, the fraction of $(d+1)$-tuples that are sampled. The y-axis shows the sampled bound (β_ϕ) and the true bound (β) on the global accuracy of an optimal classifier. Here, global accuracy refers to the fraction of points in the dataset that are correctly classified.

presented in Figure 8, show that both α_ϕ and β_ϕ converge rapidly to stable values as ϕ increases. This demonstrates that, although the number of $(d+1)$-tuples grows factorially with n, the information gained from testing additional samples diminishes beyond a certain point. For a fixed degree of confidence, the required number of samples remains constant regardless of n.



(a)

(b)

Fig. 8: (a) and (b) show the sampled $\alpha_\phi(\phi)$ and $\beta_\phi(\phi)$ for ϕ in the range $[50, 20000]$, for $d = 2$ and $d = 3$, respectively.

Although presented primarily for illustrative purposes, these plots highlight that the practical utility of the results in this paper is not necessarily constrained by the $\binom{n}{d+1}$ linear separability tests. Instead, the number of required tests becomes a matter of achieving the desired confidence level in the bound, regardless of the dataset's size or dimension. This emphasizes the flexibility and applicability of our approach across a range of scenarios and applications.

V. Conclusion

Through the lens of duality, we have unveiled the remarkable potential of the combinatorial geometry to establish lower bounds on fractional linear separability, fractional separability with margin, and fractional separation by hypersphere. Notably, the bounds deduced from Theorem 1.2 and Theorem 1.4 find direct, real-world applications in hard-margin and soft-margin SVMs, respectively. One important aspect of our work is the deterministic nature of these lower bounds. Each bound is obtained as a direct combinatorial implication of the given classifier's performance on small subsets of the dataset $\{(X_i, y_i)\}$, offering a practical and qualitatively distinct alternative to VC-dimension-based performance analysis. Moreover, this approach takes into account dataset distribution while maintaining manageable local computation requirements with fixed sample cardinality. The fusion of combinatorial methods with machine learning approaches will afford a fresh perspective on evaluating classification potential. The practical implementation of these bounds and their integration into real-world machine learning systems opens exciting avenues for further research and innovation.

References

[1] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer: New York, 2000.

[2] N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension bounds for piecewise linear neural networks," in *Conference on Learning Theory*. PMLR, 2017, pp. 1064–1068.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *Journal of the ACM (JACM)*, vol. 36, no. 4, pp. 929–965, 1989.

[4] T. Steinke and L. Zakynthinou, "Open problem: Information complexity of vc learning," in *Conference on Learning Theory*. PMLR, 2020, pp. 3857–3863.

[5] S. B. Holden and M. Niranjan, "On the practical applicability of VC dimension bounds," *Neural Computation*, vol. 7, no. 6, pp. 1265–1288, 1995.

[6] A. Kowalczyk and H. Ferrá, "MLP can provably generalize much better than VC-bounds indicate," *Advances in Neural Information Processing Systems*, vol. 9, 1996.

[7] A. C. Lorena, L. P. Garcia, J. Lehmann, M. C. Souto, and T. K. Ho, "How complex is your classification problem? a survey on measuring classification complexity," *ACM Computing Surveys*, vol. 52, no. 5, pp. 1–34, 2019.

[8] E. Mossel and C. Umans, "On the complexity of approximating the VC dimension," *Journal of Computer and System Sciences*, vol. 65, no. 4, pp. 660–671, 2002.

[9] X. Hu, L. Chu, J. Pei, W. Liu, and J. Bian, "Model complexity of deep learning: A survey," *Knowledge and Information Systems*, vol. 63, pp. 2585–2619, 2021.

[10] R. J. Webster, "Another simple proof of kirchberger's theorem," *Journal of Mathematical Analysis and Applications*, vol. 92, no. 1, pp. 299–300, 1983.

[11] M. E. Houle, "Theorems on the existence of separating surfaces," *Discrete & Computational Geometry*, vol. 6, pp. 49–56, 1991.

[12] S. R. Lay, *Convex sets and their applications*. Dover Publications, 2007.

[13] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[14] C. Campbell and Y. Ying, *Learning with support vector machines*. Springer Nature, 2022.

[15] Y. Tan and J. Wang, "A support vector machine with a hybrid kernel and minimal vapnik-chervonenkis dimension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 385–395, 2004.

[16] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning*. Springer, 2013, vol. 112.

[17] A. L. Chau, X. Li, and W. Yu, "Convex and concave hulls for classification with support vector machine," *Neurocomputing*, vol. 122, pp. 198–209, 2013.

[18] Y. Ma and G. Guo, *Support Vector Machines Applications*. Springer, 2014, vol. 649.

[19] S. Salcedo-Sanz, J. L. Rojo-Álvarez, M. Martínez-Ramón, and G. Camps-Valls, "Support vector machines in engineering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 3, pp. 234–267, 2014.

[20] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *Journal of Machine Learning Research*, vol. 2, pp. 125–137, 2001.

[21] P. K. Agarwal and M. Sharir, "Pseudo-line arrangements: Duality, algorithms, and applications," *SIAM Journal on Computing*, vol. 34, no. 3, pp. 526–552, 2005.

[22] K. P. Bennett and E. J. Bredensteiner, "Duality and geometry in SVM classifiers," in *ICML*, vol. 2000. Citeseer, 2000, pp. 57–64.

[23] J. Matousek, *Lectures on Discrete Geometry*. Berlin, Heidelberg: Springer-Verlag, 2002.

[24] J. Eckhoff, "Helly, Radon, and Carathéodory type theorems," in *Handbook of Convex Geometry*. Elsevier, 1993, pp. 389–448.

[25] I. Bárány, M. Katchalski, and J. Pach, "Quantitative helly-type theorems," *Proceedings of the American Mathematical Society*, vol. 86, no. 1, pp. 109–114, 1982.

[26] I. Bárány and G. Kalai, "Helly-type problems," *Bulletin of the American Mathematical Society*, vol. 59, no. 4, pp. 471–502, 2022.

[27] M. Katchalski and A. C. Liu, "A problem of geometry in $\mathbb{R}^n$," 1979.

[28] G.Kalai, "Intersection patterns of convex sets," *Israel J. Math*, vol. 48, pp. 161–174, 1984.

[29] D. Watson, "A refinement of theorems of kirchberger and carathéodory," *Journal of the Australian Mathematical Society*, vol. 15, no. 2, p. 190–192, 1973.

[30] N. Amenta, J. A. De Loera, and P. Soberón, "Helly's theorem: new variations and applications," *Algebraic and Geometric Methods in Discrete Mathematics*, vol. 685, pp. 55–95, 2017.

[31] S. R. Lay, "On separation by spherical surfaces," *The American Mathematical Monthly*, vol. 78, no. 10, pp. 1112–1113, 1971.

[32] W. Simons and G. Trapp, "Separating points by spheres," *Discrete Mathematics*, vol. 10, no. 1, pp. 163–166, 1974.

[33] H. Chernoff, "A Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493 – 507, 1952.