THESIS

GENERIC SUPPORT VECTOR MACHINES AND RADON'S THEOREM

Submitted by

Brittany M. Carr

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2019

Master's Committee:

Advisor: Dr. Henry Adams

Dr. Patrick Shipman
Dr. Anders Fremstad

ABSTRACT

GENERIC SUPPORT VECTOR MACHINES AND RADON'S THEOREM

A support vector machine, (SVM), is an algorithm which finds a hyperplane that optimally separates labeled data points in $\mathbb{R}^n$ into positive and negative classes. The data points on the margin of this separating hyperplane are called *support vectors*. We study the possible configurations of support vectors for points in general position. In particular, we connect the possible configurations to Radon's theorem, which provides guarantees for when a set of points can be divided into two classes (positive and negative) whose convex hulls intersect. If the positive and negative support vectors in a generic SVM configuration are projected to the separating hyperplane, then these projected points will form a Radon configuration.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Henry Adams for his help with this project from its conception all the way through to its final form. I would also like to thank Dr. Elly Farnell for her insight, understanding, and perspective. Further, I would like to thank Michael Kirby, Chris Peterson, and Simon Rubinstein–Salzedo for helpful conversations. Finally, I would like to thank everyone who has supported me on my Master's journey.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

Support vector machines (SVM) are an algorithm which, when given a set of linearly separable points in $\mathbb{R}^n$, will find the separating hyperplane with the widest margin of separation between the two classes. This distance is called the margin of error. The vectors from either the positive or negative class that minimize the distance to the separating hyperplane are called the *supporting vectors*; their positions define the location of the optimal separating hyperplane. SVM has several different incarnations but this paper will focus on the most classical type of SVM, hard margin SVM. This type of SVM does not allow for any misclassified points and it is restricted to linearly separable data. For an example of hard margin SVM in 2-dimensions, see Figure 2.1. We are interested in the theoretical properties of these support vectors in the hard margin case. To describe these properties, we need to borrow ideas from geometry and topology, and in particular a result called Radon's theorem.

Radon's theorem is a classical result in geometry and topology that looks at $n + 2$ points in n-dimensional space. Essentially, Radon's theorem states that given a set T of k points in Euclidean n-dimensional space $\mathbb{R}^n$ with $k \geq n + 2$, then there are disjoint sets T_1 and T_2 with $T = T_1 \cup T_2$ where the intersection of the convex hulls of T_1 and T_2 is nonempty. In 2-dimensional space, the convex hull of a set of points can be visualized by putting a rubber band around all of the points in that set. Radon's theorem, along with other topological tools, can shed light on the properties of support vectors in SVM.

In this thesis, we explore the possible configurations of the SVM support vectors when given a set of points in general position. General position in this context essentially means the points are randomly spread out and there are no configurations where, for example, three points fall on a line. Using Radon's theorem, we show that the projection of the support vectors onto the separating hyperplane intersect. After establishing a stronger notion of points in general position, and studying its properties using classical algebraic geometry, we show that linearly separable points in strong

general position in $\mathbb{R}^n$ can have anywhere from 2 to at most $n + 1$ support vectors. Further, in this generic case, we show that the projections of our support vectors onto the separating hyperplane have the convex hulls intersecting at a unique point. Finally, we conjecture that perturbing the support vectors by an arbitrarily small amount leaves the set of support vectors unchanged.

Chapter 2

Background on support vector machines

Support vector machines are a popular supervised learning technique that have seen success in a wide variety of applications; see [1] for a small sampling of examples. First we establish our definition of $\mathbb{R}^n$ and general position. Then we provide a brief introduction to support vector machines (SVM)s as background to this work, but we refer the interested reader to one of several resources for further information: [2–5].

2.1 Preliminaries

For our work with support vector machines, our focus will be restricted to $\mathbb{R}^n$. Let $\mathbb{R}^n$ denote Euclidean space of dimension n. For $\mathbf{x} \in \mathbb{R}^n$ we let $\|\mathbf{x}\| = \sqrt{x_1^2 + \ldots + x_n^2}$ denote its length, and for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ we let $\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + \ldots + x_n y_n$ denote their inner product. Note $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$, where the superscript T denotes the matrix transpose. Another name for the inner product is the dot product, and indeed we also denote $\langle \mathbf{x}, \mathbf{y} \rangle$ by $\mathbf{x} \cdot \mathbf{y}$.

A *linear subspace* of $\mathbb{R}^n$ is a vector space subset of $\mathbb{R}^n$; any linear subspace $V \subseteq \mathbb{R}^n$ can be written as $V = \{\sum_{i=1}^k c_i \mathbf{v}_i \mid c_i \in \mathbb{R}\}$ for some collection of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ with $0 \leq k \leq n$; the subspace has dimension k if the vectors are linearly independent[1]. In this setting we may also denote the linear subspace V by $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$, or by $\operatorname{span}(S)$ where $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. An *affine subspace* of $\mathbb{R}^n$ is a translation of a linear subspace. More explicitly, an affine subspace is of the form $\mathbf{x} + V := \{\mathbf{x} + \mathbf{v} \mid \mathbf{v} \in V\}$, where V is a linear subspace of $\mathbb{R}^n$. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ spans the linear subspace V, and if $S = \{x, x + \mathbf{v}_1, \ldots, x + \mathbf{v}_k\}$, then we may write $\operatorname{aff\,span}(S) := x + V$ to denote the affine span of the vectors in S.

Any linear subspace of dimension $n - 1$ in $\mathbb{R}^n$ can be written as $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{w}^T \mathbf{x} = 0\}$ for some normal vector $\mathbf{w} \in \mathbb{R}^n$. Similarly, any affine subspace of dimension $n - 1$ in $\mathbb{R}^n$ can be

[1] When $k = 0$, then V is a single point, the origin in $\mathbb{R}^n$.

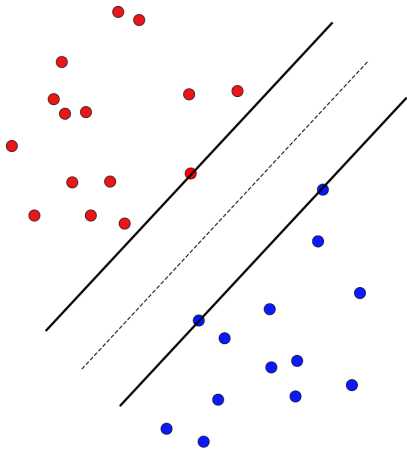


Figure 2.1: Linearly separable two-class data, along with a linear classifier (a separating hyperplane) with the maximal margin of separation.

written as $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{w}^T\mathbf{x} = b\}$ for some normal vector $\mathbf{w} \in \mathbb{R}^n$ and offset $b \in \mathbb{R}$. We refer to $(n-1)$-dimensional affine subsets of $\mathbb{R}^n$ as *hyperplanes*.

Further, we want to examine datasets with generic configurations of points. Thus a collection of points $X \subseteq \mathbb{R}^n$ is typically said to be in "general position" (for a particular property) if there exists some $\varepsilon > 0$ such that for any ε-perturbation of the points, the particular property remains unchanged. Some of the most common notions of general position can be rephrased in terms of the number of points lying in an affine subspace, or alternatively affine subspaces and spheres. For example, in [6] a finite set $S \subseteq \mathbb{R}^n$ is in *general position* if, for any $k < n$, no $(k+2)$-subset of S lies in a k-dimensional affine subspace.

2.2 SVMs in the linearly separable case

We begin with the most basic definition of support vector machines, and discuss more general versions later in the paper. Consider a set of training data $\{\mathbf{x}_i\}_{i=1}^m$ drawn from a space X and associated classes $\{y_i\}_{i=1}^m$; for now, we assume $X \subseteq \mathbb{R}^n$ and each $y_i \in \{-1, 1\}$. From this data, we train a support vector machine, which will be a linear classifier maximizing the separation between the two classes of training data. For the moment, we assume the data classes are linearly

separable — that is, we assume there exists a hyperplane[2] in $\mathbb{R}^n$ such that each data point $\mathbf{x}_i$ with $y_i = 1$ is on one side of the hyperplane, and each data point $\mathbf{x}_i$ with $y_i = -1$ is on the opposite side of the hyperplane. See Figure 2.1 for an example of linearly separable two-class data.

We construct the support vector machine as a nonlinear optimization problem with constraints in the following way. First, define the classifier as $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w}$ and b are the parameters that must be 'learned.' Note that $f(\mathbf{x}) = 0$ defines a linear decision boundary and that the sign of $f(\mathbf{x})$, namely $\text{sign}(f(x)) \in \{-1, 1\}$, determines which of the two classes the classifier predicts as the true membership class of data point $\mathbf{x}$.

Note that for any constant $c \neq 0$, the parameters $c\mathbf{w}$ and cb would define the same hyperplane as $\mathbf{w}$ and b. Thus, we introduce the notion of a canonical hyperplane as in [3]: Given a set of data $\mathbf{x}_i \in X$, a hyperplane representation is *canonical* if $\min_{\mathbf{x}_i} |\mathbf{w}^T\mathbf{x}_i + b| = 1$. We define *support vectors* for such a hyperplane to be precisely those $\mathbf{x}_i$ for which this minimum is achieved.

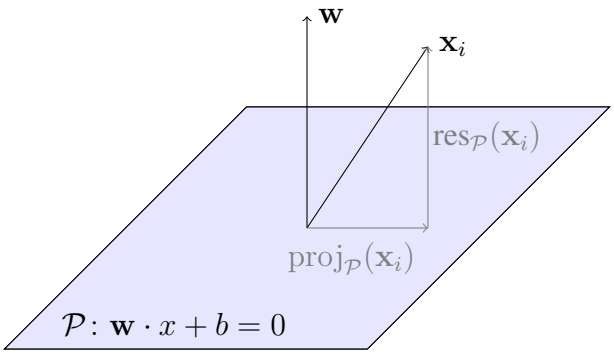


Figure 2.2: Decision boundary hyperplane along with support vector $\mathbf{x}_i$ and decomposition.

Recall that the support vector machine defined by $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ should be a linear classifier with maximal margin of separation between two data classes. Therefore, by characterizing the margin of separation between the classes, we can determine the form the appropriate optimization problem should take. Consider the canonical hyperplane $\mathcal{P}$ defined by

$$\mathcal{P} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{w}^T\mathbf{x} + b = 0\},$$

[2]Recall a hyperplane is an $(n-1)$-dimensional affine subspace of $\mathbb{R}^n$

and let $\mathbf{x}_i$ be a support vector with angle less than $\pi/2$ with $\mathbf{w}$ (thus, $\mathbf{x}_i$ is on the same side of the hyperplane as $\mathbf{w}$). Since our hyperplane representation is canonical, this gives $\mathbf{w} \cdot \mathbf{x}_i = 1 - b$. As in Figure 2.2, note that $\mathbf{x}_i$ can be decomposed orthogonally into a component in $\mathcal{P}$ and its residual: $\mathbf{x}_i = \text{proj}_{\mathcal{P}}(\mathbf{x}_i) + \text{res}_{\mathcal{P}}(\mathbf{x}_i)$. The minimum distance d from the hyperplane $\mathcal{P}$ to X is then $d = \|\text{res}_{\mathcal{P}}(\mathbf{x}_i)\|$. Note that by construction, we have $\text{res}_{\mathcal{P}}(\mathbf{x}_i) = d\frac{\mathbf{w}}{\|\mathbf{w}\|}$, where $\frac{\mathbf{w}}{\|\mathbf{w}\|}$ is the unit vector in the direction of $\mathbf{w}$. Taking the dot product of the decomposition of $\mathbf{x}_i$ with $\mathbf{w}$ and using the fact that $\mathbf{x}_i$ is a support vector, we have

$$\begin{aligned} 1 - b &= \mathbf{w} \cdot \mathbf{x}_i \\ &= \mathbf{w} \cdot \text{proj}_{\mathcal{P}}(\mathbf{x}_i) + \mathbf{w} \cdot \text{res}_{\mathcal{P}}(\mathbf{x}_i) \\ &= -b + \mathbf{w} \cdot \left(d\frac{\mathbf{w}}{\|\mathbf{w}\|} \right) && \text{since } \text{proj}_{\mathcal{P}}(\mathbf{x}_i) \in \mathcal{P} \\ &= -b + d\|\mathbf{w}\|. \end{aligned}$$

Adding b to both sides gives $1 = d\|\mathbf{w}\|$, i.e., $d = \frac{1}{\|\mathbf{w}\|}$.

A similar argument produces the same conclusion when $\mathbf{x}_i$ is assumed to be a support vector with angle $\theta \in (\pi/2, \pi]$ with $\mathbf{w}$. Thus, we have that the minimum distance from the linear decision boundary to any point $\mathbf{x}_i$ in X is $\frac{1}{\|\mathbf{w}\|}$. This means that the margin of separation between the classes is $\frac{2}{\|\mathbf{w}\|}$; consequently, in order to maximize the margin of separation, we seek to minimize the norm of $\mathbf{w}$. For convenience, we will instead minimize $\frac{1}{2}\|\mathbf{w}\|^2$.

To characterize correct classification of all $\mathbf{x}_i \in X$, we observe that if $y_i = 1$, a support vector machine with perfect classification must have $f(\mathbf{x}_i) > 0$ and if $y_i = -1$, we require $f(\mathbf{x}_i) < 0$. But our definition of canonical hyperplane requires that $|f(\mathbf{x}_i)| \geq 1$ for all i. So classification of all points is correct if $y_i f(\mathbf{x}_i) \geq 1$ for all i. That is, classification of all points is correct if $y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) \geq 1$ for all i.

Therefore, the support vector machine optimization problem is

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) \geq 1 \text{ for all } i.$$

2.3 Solving the SVM optimization problem

Now that we have an optimization problem, we need some additional tools to solve it. The Karush-Kuhn-Tucker (KKT) theorem can be applied to the SVM optimization problem and provide an optimal solution, provided the original problem satisfies certain conditions. The KKT theorem works in a similar fashion to Lagrange multipliers. Under the right conditions, it takes a bounded lower-dimensional problem and turns it into a higher-dimensional, unbounded problem which can be solved using calculus. The development of the dual problem will follow the argument in [7], which follows the paper [8].

Theorem 2.3.1 (Karush-Kuhn-Tucker)**.** *Consider an optimization problem in $\mathbb{R}^n$ of the form*

$$\min(f(x)) \textit{ subject to } g_i(x) \leq 0 \textit{ for all } i = 1, \ldots, m,$$

where $f(x)$ is a differentiable function of input variables x, and the $g_i(x)$ are affine degree one polynomials. Suppose z is a local minimum of f. Then, there exist constants $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}$ such that

(1) $-\nabla f(z) = \sum_{i=1}^{m} \alpha_i \nabla g_i(z)$

(2) $g_i(z) \leq 0$ *for all* i*,*

(3) $\alpha_i \geq 0$ *for all* i*, and*

(4) $\alpha_i g_i(z) = 0$ *for all* i*.*

The four results of the KKT theorem apply directly to the SVM optimization problem. Condition (1) gives that the Lagrangian is zero. Conditions (2) and (3) guarantee that the original and dual constraints are satisfied. The final condition guarantees that the support vectors must have margin exactly 1. Each original constraint corresponds to a Lagrange multiplier and a new constraint for the dual problem. Instead of optimizing the original SVM problem, we will instead be optimizing the dual.

To optimize the dual problem, we first need to define it. Because of some additional constraints on the SVM problem, solving the dual will directly solve the original problem. Further, the dual will be a maximization problem as opposed to the minimization problem. The entries of the dual are $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, with one α_i for each constraint of the SVM problem. By the KKT theorem, we know that $\alpha_j \geq 0$ for all j. Thus, our generalized Lagrangian (dual for our SVM problem) is

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{m} y_j \alpha_j (\langle \mathbf{w}, \mathbf{x}_j \rangle + b) \\ &= \frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{m} \alpha_j - \sum_{j=1}^{m} y_j \alpha_j (\langle \mathbf{w}, \mathbf{x}_j \rangle) - \sum_{j=1}^{m} \alpha_j y_j b. \end{aligned}$$

By condition (1) of the KKT theorem, our Lagrangian must be 0. Thus, $\frac{\partial L}{\partial b} = \sum_{j=1}^{m} \alpha_j y_j = 0$. Now for each $\mathbf{w}_i$, we have $\frac{\partial L}{\partial \mathbf{w}_i} = \mathbf{w}_i - \sum_{j=1}^{m} y_j \alpha_j x_{j,i}$, where $x_{j,i}$ is the i-th coordinate of vector $\mathbf{x}_j$. Again by condition (1), setting this to 0 gives us $\mathbf{w} = \sum_{j=1}^{m} y_j \alpha_j \mathbf{x}_j$. We want to write our optimal solution in terms of Lagrange multipliers since our support vector criteria, condition (4), dictates that several of our α_i terms will be 0. Using our new definition of $\mathbf{w}$, we construct our dual function as,

$$L(\alpha) = \sum_{j=1}^{m} \alpha_j - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle.$$

From here, we can solve the unbounded optimization problem using a variety of methods. The non-zero α_j's are the solutions to the dual, and they correspond to the $\mathbf{x}_j$'s which serve as the support vectors. This defines the desired linear hyperplane which best divides the two classes of data.

2.4 SVMs in the general case (soft margins)

The most common versions of support vector machines do not require the hypothesis that the data be linearly separable — indeed, the optimization problem is edited to optimize over two preferences: maximizing the margin of separation, i.e. the "width of the road", while also minimizing

the amount to which points lie within the margin (or are misclassified). To do so, we look at the hinge loss function, $\max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i - b))$ which yields 0 if the point is on the correct side of the margin, and if the point is on the incorrect side, the function provides the loss which is proportional to how far away the point is from the margin. As discussed in [2], the soft-margin optimization problem is to minimize

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i - b))\right] + \lambda||\mathbf{w}||^2,$$

where λ is a variable which compensates for the compromise of a wide margin as opposed to classifying points incorrectly.

Given this new optimization problem, we could ask what support vector configurations are possible in the soft-margin case, and whether or not those configurations produce a Radon point. Further, we could compare the soft margin support configurations to those found in the hard margin case. But this paper will focus on the hard margin support vector machines, restricted to linearly separable data, which is the mathematically simpler case.

2.5 SVMs and VC dimension

The Vapnik–Chervonenkis (VC) dimension, first introduced by Vapnik and Chervonenkis in 1971 [9], is a measure of the complexity of a classification model. Given a set of points X, note that there are $2^{|X|}$ ways to label the points with labels $+1$ or -1. We say that a classification model can *shatter* a set of points X if no matter how the labels $+1$ or -1 are assigned to X, we can find a classifier from that model that correctly recovers the assigned labels. The VC dimension of a classification model is the cardinality of the largest set of points that model can shatter [10].

As an example, consider the VC dimension of affine separators (binary classifiers that are defined by which side of an affine hyperplane a data point lives in) in $\mathbb{R}^2$. Note that support vector machines are a particular type of affine separators in which the affine hyperplane is chosen in a particular way. Consider first a set $X \subseteq \mathbb{R}^2$ of three points, namely $|X| = 3$, that do not lie on a

single line. Note there are $2^3 = 8$ ways to label these points with values in $\{-1, 1\}$. No matter how those three points are labelled with values in $\{-1, 1\}$, there exists an affine line that correctly separates the two classes ($+1$ and -1). Thus, the VC dimension of affine separators in $\mathbb{R}^2$ is at least three.

Figure 2.3: Examples of two sets of size 4 that cannot be shattered. Consider the labeled classes as drawn in red and blue. In each example, the convex hulls of the two classes intersect at the origin. As we explain below, this implies that no affine separator can correctly classify according to these labels.

We now argue that the VC dimension of affine separators in $\mathbb{R}^2$ exactly three [11]. To do this, we must show that for any set $X \subseteq \mathbb{R}^2$ with $|X| = 4$, the set X cannot be shattered. Radon's theorem, which is discussed in Chapter 3, states that since X is a set of 4 points in $\mathbb{R}^2$, there must be disjoint sets X_+ and X_- with $X = X_+ \cup X_-$ and $\text{conv}(X_+) \cap \text{conv}(X_-) \neq \emptyset$ (see Figure 2.3). If we label all the points in X_+ as belonging to the positive class $+1$, and all the points in X_- as belonging to the negative class -1, then no affine separator will be able to correctly classify these labels. Indeed, any affine separator assigning each point in X_+ the label $+1$ must also assign all of $\text{conv}(X_+)$ the label $+1$, and any affine separator assigning each point in X_- the label -1 must also assign all of $\text{conv}(X_-)$ the label -1. This contradicts the fact that $\text{conv}(X_+) \cap \text{conv}(X_-)$ is nonempty! Hence no set of four points in $\mathbb{R}^2$ can be shattered by an affine separator, and so the VC dimension of affine separators in $\mathbb{R}^2$ is three.

More generally, Radon's theorem states if X is a set of k points in Euclidean n-dimensional space $\mathbb{R}^n$ with $k \geq n + 2$, then there are disjoint sets X_+ and X_- with $X = X_+ \cup X_-$ and $\text{conv}(X_+) \cap \text{conv}(X_-) \neq \emptyset$. So, if we label the points in X with X_+ as the positive class and

with X_- as the negative class, then since their convex hulls intersect, this set of points cannot be shattered by an affine separator. Thus, no configuration of $n + 2$ or more points in $\mathbb{R}^n$ can be shattered. It follows that $n + 1$ is an upper bound for the VC dimension of affine separators in $\mathbb{R}^n$. It is also true that so long as $n + 1$ points in $\mathbb{R}^n$ do not lie in a $(n - 1)$-dimensional affine plane (for example, if those points live at the vertices of a regular n-simplex), then those points can be shattered by an affine separator, whose VC dimension is therefore exactly $n + 1$.

Chapter 3

Background on Radon's theorem

Radon's theorem is a classical result in topology that has applications across a variety of fields. We proceed with some definitions, Radon's theorem, and additional corollaries that pertain to classifying support vectors in $\mathbb{R}^n$.

In Euclidean space $\mathbb{R}^n$ of dimension n, a *hyperplane* is a $(n-1)$-dimensional affine subspace. A set $S \subseteq \mathbb{R}^n$ is said to be *convex* if for all $\mathbf{x}, \mathbf{y} \in S$ and $t \in (0,1)$, we also have that $t\mathbf{x}+(1-t)\mathbf{y} \in S$. Let $\mathrm{conv}(T)$ denote the *convex hull* of a set of points $T \subseteq \mathbb{R}^n$, which is the minimal convex set in $\mathbb{R}^n$ containing T. Equivalently, $\mathrm{conv}(T)$ is the intersection of all convex sets in $\mathbb{R}^n$ containing T. Furthermore, let $\mathrm{conv}(T_1) \cap \mathrm{conv}(T_2)$ denote the intersection of two convex hulls. We begin with a lemma that allows us to partition our set to separate any two points.

Lemma 3.0.1 (Separation Lemma [6]). *If $T = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{n+1}\}$ is a subset of $n+2$ points in $\mathbb{R}^n$ that does not lie in a hyperplane, then there is a hyperplane passing through n of the points of T which separates the remaining two points.*

With this in mind, we state and prove Radon's theorem. The original reference is [12], and modern references include, for example, [13].

Theorem 3.0.2 (Radon's Theorem). *If T is a set of k points in Euclidean n-dimensional space $\mathbb{R}^n$ with $k \geq n+2$, then there are disjoint sets T_1 and T_2 with $T = T_1 \cup T_2$ and $\mathrm{conv}(T_1) \cap \mathrm{conv}(T_2) \neq \emptyset$.*

Proof. It suffices to prove the case $k = n+2$. Let $T = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{n+1}\}$ be a set of $n+2$ points in $\mathbb{R}^n$. There exists coefficients, $(a_0, ..., a_{n+1})$, such that

$$\sum_{i=0}^{n+1} a_i \mathbf{x}_i = \vec{0} \quad \text{and} \quad \sum_{i=0}^{n+1} a_i = 0, \tag{3.1}$$

and for at least one i, $a_i \neq 0$. Indeed, this is because (3.1) is a collection of $n+1$ homogeneous linear equations with $n+2$ unknowns. Consider a specific non-trivial solution, $(a_0, ..., a_{n+1})$ and

Figure 3.1: Two Radon configurations in $\mathbb{R}^2$.

define the set $S = \{a_0, ..., a_{n+1}\}$. We partition S such that S_1 contains all positive coefficients and S_2 contains all zero-valued and negative coefficients. Since the sum of the coefficients must be 0, there exists a point $\mathbf{v} \in \mathbb{R}^n$ such that

$$\mathbf{v} = \sum_{i \in S_1} \frac{a_i}{A} \mathbf{x}_i = -\sum_{j \in S_2} \frac{a_j}{A} \mathbf{x}_j,$$

where $A = \sum_{i \in S_1} a_i = -\sum_{j \in S_2} a_j$. Thus $\mathbf{v}$ is an intersection point of the two convex hulls, since each sum is the representation of $\mathbf{v}$ as a convex combination for points in T_1 and T_2, respectively. Therefore, $\text{conv}(T_1) \cap \text{conv}(T_2) \neq \emptyset$. □

In other words, given a set of points containing at least $n + 2$ points, there exists a partition of the set such that the convex hulls intersect. Additional proofs of Radon's theorem can be found in a variety of sources; for example a geometric proof can be found in a paper by Peterson [6]. The points in the intersection of the convex hulls will be relevant throughout the rest of the paper. We call them *Radon points* since they are guaranteed by Radon's theorem.

Definition 3.0.3. Given a finite set $T \subseteq \mathbb{R}^n$ with disjoint labeled subsets T_1 and T_2, a *Radon point* is any point $\mathbf{v} \in \text{conv}(T_1) \cap \text{conv}(T_2)$.

Some applicable corollaries follow from the proof of Radon's theorem. For this section, we use the following definition of general position.

Definition 3.0.4. A finite set $X \subseteq \mathbb{R}^n$ is in *general position* if, for any $k < n$, no $(k+2)$-subset of X lies in a k-dimensional affine subspace, also known as a *k-flat*.

Theorem 3.0.5 ([6]). *Let T be a $(n+2)$-set in $\mathbb{R}^n$. Then T is in general position if and only if the partition $\{T_1, T_2\}$ guaranteed by Radon's Theorem is unique.*

This uniqueness goes even further. Not only is the partition unique, general position also implies that the intersection of the convex hulls is a unique single point.

Theorem 3.0.6 ([6]). *Let $\{T_1, T_2\}$ be the Radon partition of a set of $n+2$ points in general position in $\mathbb{R}^n$. Then $\mathrm{conv}(T_1) \cap \mathrm{conv}(T_2)$ is a single point.*

In the context of support vector machines, in the following chapter we will show that if our support vectors are in general position, then upon projecting them onto the $(n-1)$-dimensional separating hyperplane, their convex hulls intersect at a single Radon point.

Chapter 4

Radon's theorem and SVMs

Radon's theorem can be used as a tool to identify and classify support vector configurations. The support vectors in SVM are points in general position as defined in Definition 3.0.4. Further, the support vectors are already partitioned into two different classes, the positive class and the negative class. As such, we can say a few new things about their configurations and their projections onto the separating hyperplane. For example, Lemma 4.0.1 shows the projection of the convex hulls of the support vectors from the two classes intersect. In $\mathbb{R}^3$, the support vector configurations will look like the figures found in Figure 4.1, and when we project those configurations onto the separating hyperplane we get a Radon point. To say more about the properties of the Radon points (such as uniqueness), we will need some additional definitions and concepts which will be discussed in Chapters 5 and 6.

Lemma 4.0.1. *If $X \subset \mathbb{R}^n$ is a set of linearly separable labeled points, then the projections of the convex hulls of the positive and negative support vectors onto the separating hyperplane intersect.*

Proof. By the KKT conditions we have (1) $\mathbf{w} = \sum_j y_j \alpha_j \mathbf{x}_j$ and (2) $\sum_j \alpha_j y_j = 0$. Let $P = \{j \mid y_j = 1\}$ be the set of indices in the positive class, and let $N = \{j \mid y_j = -1\}$ be the set of indices in the negative class. Reorganizing equation (2), we may define $C := \sum_{j \in P} \alpha_j = \sum_{j \in N} \alpha_j$. Since the intersection patterns of projections of the convex hulls are unchanged by translation, we may asume in our SVM model that the translation vector $\mathbf{b}$ is the zero vector. Let $\rho \colon \mathbb{R}^n \to \mathbb{R}^n$ be the orthogonal projection onto the orthogonal complement of $\mathbf{w}$, that is, the orthogonal projection onto the separating hyperplane. Then equation (1) gives that $\vec{0} = \rho(\mathbf{w}) = \rho(\sum_j y_j \alpha_j \mathbf{x}_j)$, which we may reorganize using the linearity of ρ to get

$$\rho\left(\sum_{j \in P} \alpha_j \mathbf{x}_j\right) = \rho\left(\sum_{j \in N} \alpha_j \mathbf{x}_j\right).$$

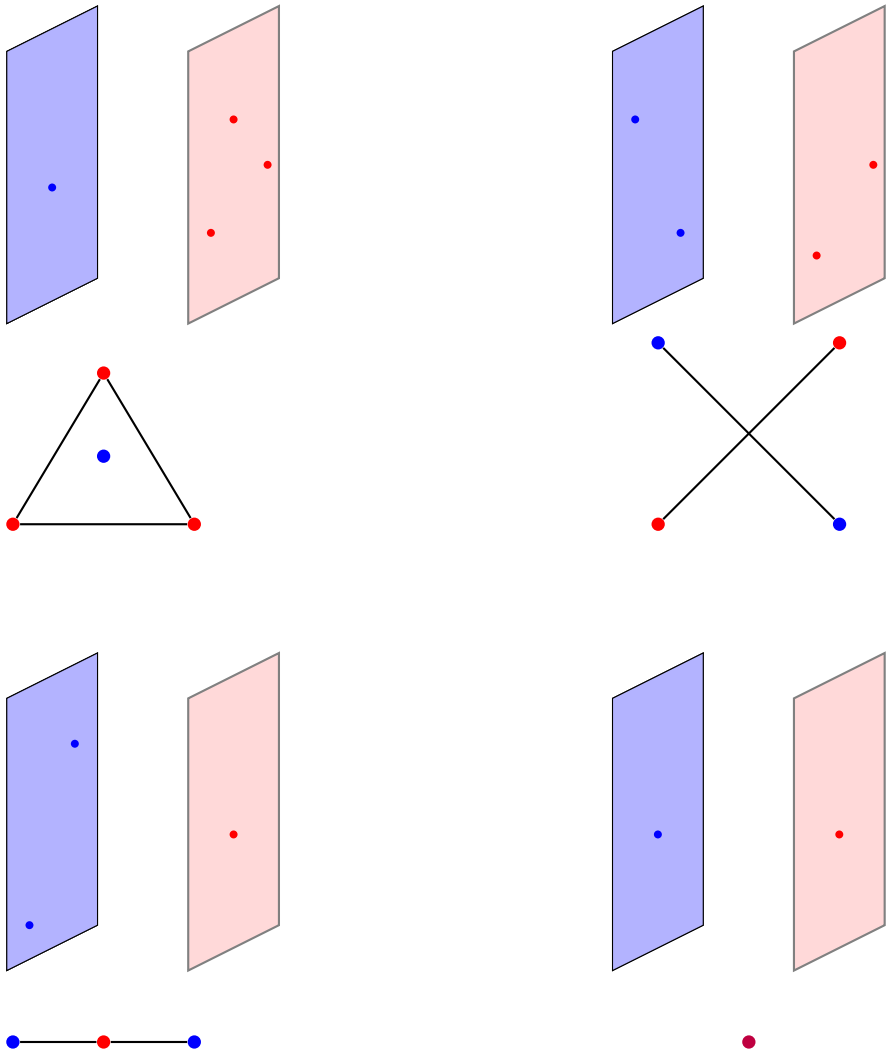


Figure 4.1: The four possible generic support vector configurations in $\mathbb{R}^3$. For each configuration, the supporting hyperplanes are drawn on top with the corresponding support vectors. Below the hyperplanes are the projections of the support vectors (forming a Radon configuration) onto the separating hyperplane.

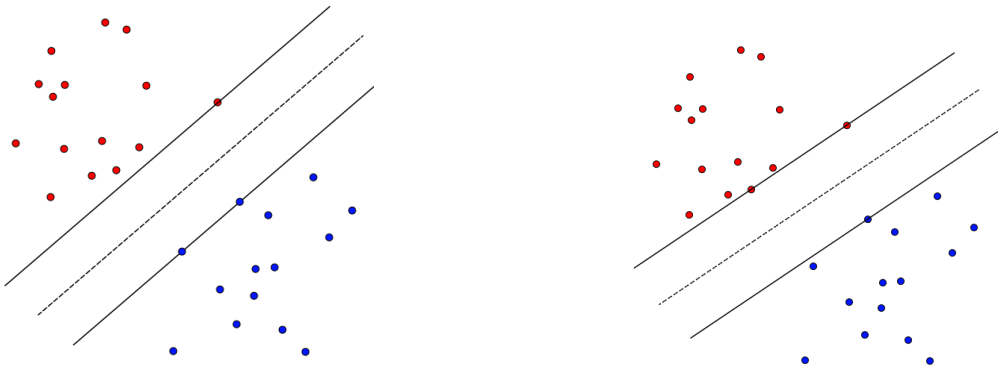
We may rescale by $\frac{1}{C}$ to obtain the same equality for convex combinations

$$\rho\left(\sum_{j\in P}\frac{\alpha_j}{C}\mathbf{x}_j\right)=\rho\left(\sum_{j\in N}\frac{\alpha_j}{C}\mathbf{x}_j\right),$$

where note that these combinations are convex since $1=\sum_{j\in P}\frac{\alpha_j}{C}=\sum_{j\in N}\frac{\alpha_j}{C}$. Hence we have shown that the projections of the convex hulls of the positive and negative support vectors onto the separating hyperplane intersect. □

Similar to the definition of a Radon point in Chapter 3, we define a Radon point in the SVM setting.

Definition 4.0.2. Given a linearly separable set of labeled points, $X\in\mathbb{R}^n$, where $X=X_-\cup X_+$, a *Radon point* is a point $\mathbf{v}\in\mathbb{R}^n$ such that $\mathbf{v}\in\rho(\text{conv}(X_-))\cap\rho(\text{conv}(X_+))$, where ρ is the projection onto the separating hyperplane.



(a) Incorrect support vectors

(b) Correct support vectors

Figure 4.2: The same dataset with two different separating hyperplanes. Note that the one on the left does not contain a Radon point in the projection of the convex hulls whereas the one on the right does. Further, the margin on the right example is larger than the one on the left.

Thus if we have a "claimed" support vector configuration where the projection of the convex hulls does not result in at least one Radon point, then we know the choice of support vectors was incorrect. Consider Figure 4.2. The two datasets in (a) and (b) are the same, but Figure 4.2a has support vectors where if one projects the convex hulls, they do not intersect. By Lemma 4.0.1, these must not be the correct support vectors. Indeed, Figure 4.2b does have a Radon point in the projection of the convex hulls and the margin between the two datasets is wider. Furthermore, Figure 4.2b has the optimal separating hyperplane.

To get stronger results, namely that in SVM the intersection of the projected convex hulls of the support vectors is a single Radon point and no more, we will need to add genericity assumptions. Indeed, we will need our set of data points to be in not only in general position, but also in *strong general position*. We introduce these genericity assumptions in Chapter 6 after giving some preliminaries involving algebraic geometry.

Chapter 5

Background on algebraic geometry and the Zariski topology

Our further exploration of general position and support vector machines relies on different aspects of algebraic geometry. All of the content in this section follows from [14–16].

The focus of this paper is restricted to the field $\mathbb{R}$, and all following definitions will reflect that focus. Nevertheless, certain definitions can be extended to other fields, such as $\mathbb{C}$ and $\mathbb{Q}$. For this reason, for some definitions we let $\mathbb{A}$ denote an arbitrary field, even though we will later restrict attention to the case $\mathbb{A} = \mathbb{R}$. The n-fold product space $\mathbb{A}^n$ is called *affine space*.

Definition 5.0.1. An *affine variety* is the set of common zeros of a finite family of polynomials. Given a set $S \subseteq \mathbb{A}[x_1, x_2, \cdots x_n]$ of polynomials in some affine space $\mathbb{A}^n$, the affine variety defined by S is the set

$$\mathcal{V}(S) := \{a \in \mathbb{A}^n \mid f(a) = 0 \text{ for all } f \in S\}.$$

For the applications in this paper, $\mathbb{A}^n$ will be $\mathbb{R}^n$.

It is important to note that we defined affine varieties in terms of sets, but we can also recast them in terms of ideals. If $I = \langle S \rangle$ is the ideal generated by the polynomials in S, then $V(I) = V(S)$.

Some of the nice properties of closed and open sets that hold in Euclidean topology also hold for algebraic varieties equipped with the *Zariski topology*. The Zariski topology on $\mathbb{A}^n$ is the topology whose closed sets are the affine varieties, called *Zariski closed sets*, in $\mathbb{A}^n$.

Lemma 5.0.2. *Given two ideals I and J, we have $\mathcal{V}(I \cdot J) = \mathcal{V}(I \cap J)$*

Proof. Since $I \cdot J \subseteq I \cap J$, it follows that $\mathcal{V}(I \cap J) \subseteq \mathcal{V}(I \cdot J)$.

Let $p \in \mathcal{V}(I \cdot J) \subseteq \mathbb{A}^n$ be a point in $\mathcal{V}(I \cdot J)$. We show that for any polynomial $h \in I \cap J$, we have $h(p) = 0$. Note if $f \in I$ and $g \in J$, i.e. if $f \cdot g \in I \cdot J$, then $f(p)g(p) = 0$. In particular we

have $h \cdot h \in I \cdot J$. Thus $h(p) \cdot h(p) = 0$, which implies that $h(p) = 0$ and $p \in \mathcal{V}(I \cap J)$. Therefore $\mathcal{V}(I \cdot J) \subseteq \mathcal{V}(I \cap J)$.

Together these two containments imply $\mathcal{V}(I \cdot J) = \mathcal{V}(I \cap J)$. □

Theorem 5.0.3. *The intersection of any collection of affine varieties is an affine variety. The union of any finite collection of affine varieties is an affine variety.*

Proof. Let $\{I_s\}_{s \in S}$ be a collection of ideals in $\mathbb{A}^n[x_1, x_2, \cdots, x_n]$, indexed by the elements $s \in S$. It is not hard to see that

$$\bigcap_{s \in S} \mathcal{V}(I_s) = \mathcal{V}\left(\bigcup_{s \in S} I_s\right).$$

Hence, the intersection of any collection of affine varieties is an affine variety.

Recall that by Lemma 5.0.2, we have

$$V(I_{s_1} \cdot I_{s_2}) = \mathcal{V}(I_{s_1} \cap I_{s_2}) = \mathcal{V}(I_{s_1}) \cup \mathcal{V}(I_{s_2}).$$

Proceeding via induction, assume $\mathcal{V}(I_{s_1}) \cup \ldots \cup \mathcal{V}(I_{s_x}) = \mathcal{V}(I_{s_1} \cap \ldots \cap I_{s_x})$. Now, consider $\mathcal{V}(I_{s_1}) \cup \ldots \cup \mathcal{V}(I_{s_x}) \cup \mathcal{V}(I_{s_{x+1}})$. Thus,

$$\begin{aligned}
\mathcal{V}(I_{s_1}) \cup \ldots \cup \mathcal{V}(I_{s_x}) \cup \mathcal{V}(I_{s_{x+1}}) &= \mathcal{V}(I_{s_1} \cap \ldots \cap I_{s_x}) \cup \mathcal{V}(I_{s_{x+1}}) \\
&= \mathcal{V}(I_{s_1} \cdot \ldots \cdot I_{s_x}) \cup \mathcal{V}(I_{s_{x+1}}) \\
&= \mathcal{V}(I_{s_1} \cdot \ldots \cdot I_{s_x} \cdot I_{s_{x+1}}) \\
&= \mathcal{V}(I_{s_1} \cap \ldots \cap I_{s_x} \cap I_{s_{x+1}}).
\end{aligned}$$

Therefore the union of any finite collection of affine varieties is an affine variety. □

We now restrict attention to the field $\mathbb{A} = \mathbb{R}$ for the remainder of the paper.

To show why the union of infinite affine varieties is not always a variety, consider

$$X := \cup_{a \in \mathbb{N}} \mathcal{V}(x - a) = \mathbb{N}.$$

On one hand, $\mathcal{V}$ is an infinite union of affine varieties, but $\mathbb{N}$ is not an affine variety. To see this, note that affine varieties over $\mathbb{R}$ are a set of solutions to a set of polynomial equations. In $\mathbb{R}$, a nonzero polynomial can have at most n solutions. Thus we have two cases, either the set of solutions must be finite or if all polynomials are the zero polynomial, the affine variety must be $\mathbb{R}$. Hence, $\mathbb{N}$ cannot be an affine variety.

Algebraic varieties are defined by the zero sets of polynomials. One way to construct a polynomial is to take the determinant of an $n \times n$ matrix with variable entries. Connecting determinants to algebraic varieties requires one to find where the determinant is equal to 0. This provides a polynomial with a zero set, which in turn provides an algebraic variety.

Definition 5.0.4. The *determinant* of an $n \times n$ matrix A is

$$det(A) = \sum_{\sigma \in S_n} \left(sgn(\sigma) \prod a_{i,\sigma_i} \right),$$

where σ is an element in the symmetric group on n elements, and a_{i,σ_i} represents the ith row and σ_ith column entry of A (see for example [17, Page 437]).

For an example, we show how this works on a 3×3 matrix. Let $A = [a_{i,j}]$ where $1 \leq i \leq 3$ and $1 \leq j \leq 3$. Thus,

$$\begin{aligned}
det(A) &= \sum_{\sigma \in S_3} \left(sgn(\sigma) \prod a_{i,\sigma_i} \right) \\
&= sgn(e) \prod a_{i,(e)_i} + sgn(123) \prod a_{i,(123)_i} + sgn(132) \prod a_{i,(132)_i} \\
&\quad + sgn(13) \prod a_{i,(13)_i} + sgn(12) \prod a_{i,(12)_i} + sgn(23) \prod a_{i,(23)_i} \\
&= \prod a_{i,(e)_i} + \prod a_{i,(123)_i} + \prod a_{i,(132)_i} - \prod a_{i,(13)_i} - \prod a_{i,(12)_i} - \prod a_{i,(23)_i} \\
&= a_{1,1}a_{2,2}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2} - a_{1,3}a_{2,2}a_{3,1} - a_{1,2}a_{2,1}a_{3,3} - a_{1,1}a_{2,3}a_{3,2}.
\end{aligned}$$

Note, the determinant is zero if and only if the rows and columns are linearly dependent, which occurs if and only if the matrix A has *deficient rank*. Setting the determinant of an arbitrary matrix

A in $\mathbb{R}^{n\times n}$ to 0, we can construct an algebraic variety from $det(A) = \sum_{\sigma\in S_n} (sgn(\sigma)\prod a_{i,\sigma_i}) = 0$, which is a polynomial in the entries of A.

Now let M be a matrix of size $m \times n$, where $m \geq n$. We say that matrix M has *deficient rank* if the rank of M is less than n, i.e. if the columns of M are linearly dependent. Furthermore, the columns of M are linearly dependent if and only if all of the $n \times n$ minors (determinants of the corresponding $n \times n$ submatrix) of M are zero. See [18] for a full proof of this fact. The set of all $\binom{m}{n}$ different $n \times n$ minors of an $m \times n$ matrix M with $m \geq n$ thus defines an algebraic variety whose members are the matrices M of deficient rank.

Thus we have proved the variety defined below contains precisely those $m \times n$ matrices M with deficient rank.

Definition 5.0.5. Let $\mathcal{V}_{\text{rd}}(m, n) \subseteq \mathbb{R}^{mn}$ for $m \geq n$ be the algebraic variety generated by the set of polynomials $\{N_i\}_{i\in I}$, where I is the set containing all $\binom{m}{n}$ choices of n rows, and N_i is the minor of the submatrix consisting of those rows.

Lemma 5.0.6. *Let $M(y)$ be an $m \times n$ matrix with $m \geq n$, depending on $y \in \mathbb{R}^k$. Suppose the entries of $M(y)$ are linear (or even polynomial) functions in the entries of y. Then $\mathcal{V}_{M(y)} := \{y \in \mathbb{R}^k \mid M(y)$ is rank deficient$\}$ is an algebraic variety.*

Proof. Let $M(y) \in \mathbb{R}^{m\times n}$ with $m \geq n$ be a matrix with entries of polynomial functions in terms of y. Suppose $M(y)$ has deficient rank. Thus all $n \times n$ submatrices A of $M(y)$ have $\det(A) = 0$, which gives a system of $\binom{m}{n}$ polynomials in y, each polynomial being set equal to 0. Thus the set of $\binom{m}{n}$ different $n\times n$ minors defines an algebraic variety. Thus, $\mathcal{V}_{M(y)} = \{y \mid M(y)$ is rank deficient$\}$ is an algebraic variety. □

In the following section we will be studying configurations of points in general position, which is connected to algebraic geometry for example via the following theorem.

Theorem 5.0.7 (Proposition 1 of Appendix A of [19], or Theorem 1.3.2 of [16], for example)**.** *The nonempty complement of a algebraic variety in $\mathbb{R}^n$ is open and dense in the Euclidean topology on $\mathbb{R}^n$.*

Recall that an algebraic variety is a closed set in the Zariski topology, and therefore the complement of of an algebraic variety is an open set in the Zariski topology. The above theorem says that a nonempty open set in the Zariski topology is both open and dense in the Euclidean topology on $\mathbb{R}^n$. In the next chapter we will use this theorem to show that collections of points in general position (or strong general position) in $\mathbb{R}^n$ form an open and dense subset of all possible collections of points in $\mathbb{R}^n$.

Chapter 6

SVMs for points in general position

There are a wide variety of notions of general position; in Definition 3.0.4 we gave only one such notion. We begin in Subsection 6.1 by defining and studying a stronger notion of general position that will be useful for studying SVMs. In Subsection 6.2, we give the basic properties of SVMs that hold in this slightly more restrictive setting of strong general position. Finally, in Subsection 6.3, we conjecture that for points in strong general position, an arbitrarily small perturbation cannot change the set of support vectors.

6.1 General position and a stronger notion

Roughly speaking, a certain property is *generic* if it holds on a countable intersection of open dense sets. We begin this section by showing that the notion of general position in Definition 3.0.4, as well as with a stronger notion in Definition 6.1.2, both hold on an open dense set.

By a small abuse of notation, we identify each subset $X = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^n$ of cardinality m with a point $X \in \mathbb{R}^{mn}$, and also with a $m \times n$ matrix $X \in \mathbb{R}^{m \times n}$. Hence we identify $\{X \subseteq \mathbb{R}^n \mid |X| = m\}$ with $\mathbb{R}^{mn}$.

Lemma 6.1.1. *The collection of all subsets $X \subseteq \mathbb{R}^n$ in general position (Definition 3.0.4) and with $|X| = m$ is an open and dense subset of $\mathbb{R}^{mn}$. Further, if $X \subseteq \mathbb{R}^n$ is in general position, then there exists an $\varepsilon > 0$ such that perturbing any point by at most ε does not remove X from general position.*

Proof. We will show that the collection of all configurations not in general position forms an algebraic variety.

By definition, X is not in general position if and only if $k+2$ of the points in X lie in a k-flat (for some $k < n$). So X is not in general position if and only if there exists a set $S = \{x_{i_1}, \ldots, x_{i_{k+1}}\}$ and a point $x \in X \setminus S$ such that $x \in \text{aff span}(S)$, for some $k < n$. This can be recast algebraically

as follows. Consider the $n \times (k+1)$ matrix $M_{S,x} = \left[x_{i_1} - x, x_{i_2} - x, \ldots, x_{i_{k+1}} - x\right]$ whose columns are the vectors $x_{i_j} - x$ for each $1 \leq j \leq k+1$. It follows, for example from [20, Lemma 1.3.2], that $x \in \text{aff span}(S)$ if and only if the matrix $M_{S,x}$ has deficient rank, i.e., if and only if $M_{S,x} \in \mathcal{V}_{\text{rd}}(k+1, n)$ (recall Definition 5.0.5). Therefore, X is not in general position if and only if some such matrix $M_{S,x}$ belongs to the algebraic variety $\mathcal{V}_{\text{rd}}(k+1, n)$.

We will define $\mathcal{V}_{\text{ng}}(m, n) \subseteq \mathbb{R}^{mn}$ to be the algebraic variety of all matrices $X \in \mathbb{R}^{m \times n}$ such that the points determined by X are not in general position. This definition will require some additional notation which we will provide below. If a $(k+2) \times n$ matrix N describes a set of $k+2$ points in $\mathbb{R}^n$, then those points lie in a k-flat if and only if for any choice of row, subtracting that row from all others and then deleting that row gives a matrix in $\mathcal{V}_{\text{rd}}(k+1, n)$. Let $\widetilde{N}$ be the $(k+1) \times n$ matrix obtained from N by subtracting the last row from all others, and then deleting the last row. Next, let U_{k+2} be the collection of all subsets I of $\{1, \ldots, m\}$ of size $|I| = k+2$. Given an $m \times n$ matrix M and $I \in U_{k+2}$, let M_I be the $(k+2) \times n$ submatrix of M formed by selecting out the rows given by I. For $I \in U_{k+2}$, let

$$\mathcal{V}(m, n)_I := \{M \in \mathbb{R}^{m \times n} \mid \widetilde{M_I} \in \mathcal{V}_{\text{rd}}(k+1, n)\}.$$

We define $\mathcal{V}_{\text{rd}}(m, n)_{k+2} = \cup_{I \in U_{k+2}} \mathcal{V}(m, n)_I$. Since $\mathcal{V}_{\text{rd}}(m, n)_{k+2}$ is a finite union of algebraic varieties, by Theorem 5.0.3, it is also an algebraic variety. Now define $\mathcal{V}_{\text{ng}}(m, n) := \cup_{k<n} \mathcal{V}_{\text{rd}}(m, n)_{k+2}$. Similarly $\mathcal{V}_{\text{ng}}(m, n)$ is the finite union of algebraic varieties, and thus is also an algebraic variety. A matrix $X \in \mathbb{R}^{m \times n}$ belongs to $\mathcal{V}_{\text{ng}}(m, n)$ precisely when the points determined by X are not in general position.

Now, since $M_{S,x} \in \mathcal{V}(m, n)_I$, taking the union over all choices of k and selection of k subsets implies $X \in \mathcal{V}_{\text{ng}}(m, n)$. Since $\mathcal{V}_{\text{ng}}(m, n)$ is an algebraic variety this implies that $\mathcal{V}_{\text{ng}}(m, n)$ is a *Zariski closed set*.

We next consider the complement of $\mathcal{V}_{\text{ng}}(m, n)$, i.e., all subsets of X in general position, First we show that the Zariski open set of all configurations of points not in general position is nonempty. Indeed, we can find one set of points in general position by choosing points on the moment curve:

$\gamma(t) = \{t, t^2, \ldots, t^d \mid t \in \mathbb{R}\} \in \mathbb{R}^{mn}$ [21]. Theorem 5.0.7 then implies that the set of all configurations of points in general position are open and dense in the Euclidean topology on $\mathbb{R}^{mn}$. Therefore since X is open in the Euclidean topology, there exists $\epsilon > 0$ such that perturbing any point by ϵ does not remove X from general position. □

We now provide the following definition for a stronger form of general position.

Definition 6.1.2. A set of points $X \subseteq \mathbb{R}^n$ is in *strong general position* if

(i) for $k < n$, no $k + 2$ subset of X lies in a k-flat

(ii) for any $k + 1$ points in X (determining a k-flat), the orthogonal projection of any other point in X to that k-flat does not hit the affine span of k of those points

(iii) for $k + l \leq n$, no disjoint k-flats and l-flats contain parallel vectors.

Lemma 6.1.3. *The collection of all subsets $X \subseteq \mathbb{R}^n$ with $|X| = m$ in strong general position is an open and dense subset of $\mathbb{R}^{mn}$. It follows that if $X \subseteq \mathbb{R}^n$ is in strong general position, then there exists an $\varepsilon > 0$ such that perturbing any point by at most ε does not remove X from strong general position.*

Proof. We will construct an algebraic variety for each of the three conditions in Definition 6.1.2. The union of these three algebraic varieties will be an algebraic variety of all configurations not in general position. After showing that the complement is nonempty, the result will follow from Theorem 5.0.7.

For the first condition (i), the algebraic variety is the one provided in the proof of Lemma 6.1.1.

For condition (ii), denote $S = \{x_{i_1}, \ldots, x_{i_{k+1}}\} \subseteq X$, and let $\pi_S \colon \mathbb{R}^n \to \text{aff span}\,(x_{i_1}, \ldots, x_{i_{k+1}})$ be the orthogonal projection onto the affine span. Consider the set of all points $x \in X \setminus S$ such that $\pi_S(x)$ lives in $\text{aff span}(x_{i_1}, \ldots, \hat{x}_{i_j}, \ldots, x_{i_{k+1}})$ for some $1 \leq j \leq k + 1$, where $\hat{x}_{i_j}$ denotes the deletion of the $x_{i_j}^{\text{th}}$ term. Since π_S is a linear operator and $\text{aff span}(x_{i_1}, \ldots, \hat{x}_{i_j}, \ldots, x_{i_{k+1}})$ is an affine subspace, it follows that $\pi_S^{-1}(\text{aff span}(x_{i_1}, \ldots, \hat{x}_{i_j}, \ldots, x_{i_{k+1}}))$ is an affine subspace [22].

Thus, if $\pi_S(x)$ lives in aff $\text{span}(x_{i_1}, \ldots, \hat{x}_{i_j}, \ldots, x_{i_{k+1}})$ for some $1 \leq j \leq k+1$, then

$$x \in \pi_S^{-1}(\text{aff span}(x_{i_1}, \ldots, \hat{x}_{i_j}, \ldots, x_{i_{k+1}})).$$

This space forms an algebraic variety, $\mathcal{V}_S = \cup_{j=1}^{k+1} \pi_S^{-1}(\text{aff span}(x_{i_1}, \ldots, \hat{x}_{i_j}, \ldots, x_{i_{k+1}}))$. Thus, the finite union $\mathcal{V}_X = \cup_{S \subseteq X \text{ with } |S|=k+1} \mathcal{V}_S$ is an algebraic variety containing all the sets of points that do not satisfy condition (ii).

Finally, for condition (iii) consider two disjoint sets of points $T_1 = \{x_{i_1}, \ldots, x_{i_{k+1}}\} \subseteq X$ and $T_2 = \{x_{j_1}, \ldots, x_{j_{l+1}}\} \subseteq X$ such that $k + l \leq n$. We define $v_1 = x_{i_1} - x_{i_{k+1}}$, $v_2 = x_{i_2} - x_{i_{k+1}}$, $\ldots$, $v_k = x_{i_k} - x_{i_{k+1}}$ to be k difference vectors from T_1, and similarly we define $u_1 = x_{j_1} - x_{j_{l+1}}$, $\ldots$, $u_l = x_{j_l} - x_{j_{l+1}}$ to be l difference vectors for T_2. Geometrically, we are shifting aff $\text{span}(T_1)$ by $-x_{i_{k+1}}$ and shifting aff $\text{span}(T_2)$ by $-x_{j_{l+1}}$, thus aligning both to contain the origin. From a more algebraic perspective, consider $\text{span}(v_1, \ldots, v_k, u_1, \ldots, u_l)$. If the span is not $(k+l)$-dimensional, then condition (iii) is not satisfied. This is equivalent to constructing the matrix $M_T = [v_1, \ldots, v_k, u_1, \ldots, u_l]$ and examining its rank. If the rank of M_T is deficient, then (by an argument analogous to that in Lemma 6.1.1) we can use Lemma 5.0.6 to construct an algebraic variety $\mathcal{V}(k,l)_T$ of all subsets of X that do not satisfy condition (iii).

Thus, since the finite union of varieties is a variety, $\mathcal{V}(m,n)_I \cup \mathcal{V}_X \cup \mathcal{V}(k,l)_T$ is an algebraic variety, and hence a Zariski closed set, of all the configurations not in strong general position. After observing that the complement is non-empty, it follows from Theorem 5.0.7 that the set of all configurations of points in strong general position in $\mathbb{R}^{mn}$ is open and dense in the Euclidean topology on $\mathbb{R}^{mn}$. Therefore since X is open in the Euclidean topology, there exists $\epsilon > 0$ such that perturbing any point by ϵ does not remove X from strong general position. $\square$

6.2 Properties of support vector machines for points in strong general position

Now that we have a stronger notion of general position, we can say more about the possible configurations of support vectors. In this section we will show that we can have anywhere from 2 to at most $n + 1$ support vectors.

Lemma 6.2.1. *If $X \subset \mathbb{R}^n$ is a set of linearly separable labeled points in strong general position, then the projections of the convex hulls of the positive and negative support vectors onto the separating hyperplane intersect at a single Radon point.*

Proof. As shown in Lemma 4.0.1, the projections of the convex hulls of the positive and negative support vectors onto the separating hyperplane intersect in at least one point. We will now show that, given strong general position, the intersection is exactly one point. In search of contradiction, suppose the intersection of the projections of the convex hulls contains at least two distinct points, $x \neq x'$. Note that the intersection of the two projections of convex hulls is convex, and hence the intersection contains the entire line segment between x and x'. It follows that the affine span of the positive support vectors contains a 1-flat parallel to a 1-flat in the affine span of the negative support vectors. This is a contradiction since X is in strong general position; see Definition 6.1.2 (iii). Hence there is a unique Radon point. □

We will use the following lemma to show that there are at most $n + 1$ support vectors for linearly separable points in strong general position.

Lemma 6.2.2. *Let $X \in \mathbb{R}^n$ be a set of points, with $|X| \geq n + 2$, and such that X is a subset of two parallel $(n-1)$-dimensional hyperplanes. Then X cannot be in strong general position.*

Proof. Let $k = |X|$. Suppose first that we have $k = n + 2$ points contained in two parallel $(n-1)$-flats.

We proceed with two cases. First, suppose we have either no points or only one point contained in one of the two parallel hyperplanes. Then we have either $n + 2$ or $n + 1$ points in the other $(n-1)$-flat. This violates the first condition of strong general position, Definition 6.1.2(i).

Now suppose we have $k+1 \geq 2$ points in one flat and $l+1 = (n+2)-(k+1) \geq 2$ points in the other flat. Projecting the l-flat onto the *other* $(n-1)$-flat gives a k flat and an l flat living in an $(n-1)$-flat. Since $(k+1)+(l+1) = n+2$, this implies that $k+l = n > n-1$, which means that the projected k and l-flats intersect in at least a line, contradicting Definition 6.1.2(iii).

The proof above in the setting of $k = n+2$ also holds for any $k \geq n+2$. Hence, X cannot be in strong general position. □

Theorem 6.2.3. *Suppose $X \subseteq \mathbb{R}^n$ is in strong general position, and that X is equipped with linearly-separable labels. Then there are at most $n+1$ supporting vectors.*

Proof. Suppose we have $k \geq n+2$ support vectors. Then we have at least $n+2$ points in two parallel $(n-1)$-flats. By Lemma 6.2.2, this is a contradiction since X is in strong general position. Therefore we can have at most $n+1$ support vectors. □

The following lemma shows that any number of support vectors, from 2 up to $n+1$, can generically occur.

Lemma 6.2.4. *Let $2 \leq k \leq n+1$ and $1 \leq i \leq k-1$. Then there is a labeled subset $X \subseteq \mathbb{R}^n$ in strong general position with i supporting vectors from the positive class and with $k-i$ supporting vectors from the negative class.*

Proof. Since $k \leq n+1$, it is possible to find a $(k-1)$-simplex linearly embedded in $\mathbb{R}^n$ so that all of the vertex locations are equidistant. Assign i labels to the positive class and $k-i$ labels to the negative class arbitrarily. □

6.3 Stability of the support vectors

Now that the possible configurations of support vectors have been established, it remains to show that these configurations are stable. We begin by showing that points which are linearly separable with a positive margin remain linearly separable under any arbitrarily small perturbation. We define a notion of what it means for a support vector to be *stable* under arbitrarily small perturbations. We show that general position is not enough in order to guarantee that the support vectors

are stable, and we conjecture that strong general position will guarantee that the set of support vectors is stable.

Lemma 6.3.1. *If $X \subseteq \mathbb{R}^n$ is a set of linearly separable labeled points with positive margin, then there exists an $\varepsilon > 0$ such that upon perturbing any point by at most ε, X remains linearly separable.*

Proof. Let $X \subseteq \mathbb{R}^n$ be a set of linearly separable, labeled points, and let H denote the separating hyperplane. Let $\varepsilon_1 = min_{x \in X} d(x, H)$, where the distance between point x and set H is defined as $d(x, H) = \inf_{h \in H} d(x, h)$, where $d(x, h) = \|x - h\|$. By the assumption that X is linearly separable with positive margin, we have $\varepsilon_1 > 0$. Thus, we choose $\varepsilon = \frac{\varepsilon_1}{2}$ to be the maximum distance that we will allow the points to be perturbed. Upon perturbing by at most ε, note that the perturbed points are still linearly separable (and still with positive margin) by the same separating hyperplane H. Therefore, there exists an $\varepsilon > 0$ such that if every point is perturbed by at most ε, the points remain linearly separable. □

Definition 6.3.2. A support vector v is called *stable* if there exists an $\epsilon > 0$ such that any ϵ-perturbation (of all the points) leaves v as a support vector.

We propose to use this notion of stability to address the issue of numerically finding (too) many support vectors in software when, in reality, there should be a small number. Software programs such as Matlab often return more than $n+1$ support vectors, but we showed in Lemma 6.2.3 that we can have at most $n+1$ support vectors when given points in strong general position. So to identify the true support vectors, we would need to increase our precision of our algorithms. Furthermore, we can probably look at which vectors are needed for a Radon configuration in order to identify which support vectors v are stable and which are extraneous.

Conjecture 6.3.3. Let $X \subseteq \mathbb{R}^n$ be a set of linearly separable labeled points in strong general position. Let $\varepsilon_0 > 0$ be the minimum distance between any two distinct points in X. Then there exists $0 < \varepsilon < \frac{\varepsilon_0}{2}$ such that if each point is perturbed by at most ε, then the set of supporting vectors remains unchanged.

Defining what it means for the set of supporting vectors to “remain unchanged" requires a bit of care, since the vectors are being moved. What we mean is the following. Since the points in X are perturbed by less than $\frac{\varepsilon_0}{2}$, it follows that each ball of radius $\frac{\varepsilon_0}{2}$ centered at each point in X contains exactly one point in the perturbed collection of points $\tilde{X}$. This gives a bijection between X and the perturbed points $\tilde{X}$. Conjecture 6.3.3 is then saying that a point $x \in X$ is a support vector for X if and only if its corresponding point $\tilde{x} \in \tilde{X}$ is a support vector for $\tilde{X}$.

Remark 6.3.4. In this remark we give two different types of examples of why general position (as opposed to strong general position) is not sufficient for the conclusion of Conjecture 6.3.3 to hold.

(We’ll refer to (i)–(iii) in the definition of strong general position, Definition 6.1.2.)

As the first example, consider a linearly separable set of five labeled points in $\mathbb{R}^3$, $(0,0,0)$, $(0,2,0)$, $(0,0,3)$ in the positive class and $(2,1,1)$, $(2,0.5,0.5)$ in the negative class. All five of the points are support vectors since the line spanned by the two points in the negative class is parallel to some vector inside the span of the three points in the negative class. Even though these points are in general position, perturbing any single point in the x direction will change the number of support vectors, leaving only four support vectors.

For a second example, consider a set of 4 points where the single support vector of the positive class, $(0,0,0)$ lies on the boundary the triangle formed by the support vectors of the negative class, $(-1,2,0)$, $(1,2,0)$, and $(0,2,1)$. Perturbing the point $(0,0,0)$ to anywhere outside the projected boundary of the triangle formed by the three negative support vectors changes the set of support vectors. Indeed, this follows from Lemma 4.0.1, which says that the projections of the convex hulls of the support vectors must intersect. Hence there is no $\varepsilon > 0$ for which this set of support vectors can be stabily perturbed.

Chapter 7

Conclusion

In this paper, we examined support vector machines and some of their theoretical properties. After establishing background on SVM, Radon's theorem, and some additional algebraic tools in Chapters 2, 3,and 5, we were able to prove some new properties about support vectors and their possible configurations in Chapters 4 and 6. Given a set of linearly separable data, the projection of the convex hulls of the support vectors to the separating hyperplane intersect. Furthermore, if the points are in strong general position, then there is a unique point of intersection, the Radon point. Also for points in strong general position, we show there can be at most $n + 1$ support vectors in $\mathbb{R}^n$ dimensional space. Finally, we conjecture that given linearly separable data in strong general position, there exists an $\varepsilon > 0$ such that perturbing the data points by at most ε keeps the set of support vectors the same. These results lead to additional problems and questions that we can ask.

Support vector machines are a common machine learning algorithm used to look at separable classes of data. Understanding some properties of the support vectors allows us to take a closer look at the possible configurations of data. Several software programs, such as Matlab, pull too many support vectors from the data set. Our research shows that there is a maximum number of support vectors, for generic points, and that we are guaranteed to get certain (Radon) configurations. If we do not have a Radon configuration or we have too many support vectors, that would imply our points are not in general position. Thus since that happens with probability 0, our error estimation would have been incorrect and we would need to increase our precision.

We share a list of questions that are natural to ask following our results.

Question 7.0.1. What can be said for spherical or ellipsoidal SVM? See for example [23], especially Figure 3 within. In spherical SVM, we suppose that the two classes of data can be separated with one class inside a sphere, and with the second class outside. Ellipsoidal SVM is analogous, except that the sphere is generalized to also allow for separating ellipsoids. What is the maximal number of possible support vectors for separable points in general position for spherical and ellip-

soidal SVM, where “general position" will have to be carefully defined for this new context? Is there a version of Radon's theorem that relates to spherical or ellipsoidal SVM?

Question 7.0.2. Is there anything precise that can be said when the data is not linearly separable, and hence one uses soft margin SVM (allowing errors) instead of hard margins? There are several different ways one could define support vectors in this context.

Question 7.0.3. Which of the possible Radon configurations (say one support vector in each class, or one support vector in the positive class and two in the negative class, or two support vectors in each class, *etc.*) is the most likely? For this question to make sense, we need a random model of labeled data points that are linearly separable. What are reasonable such models? There is almost certainly no canonical such model, but instead a zoo of random models of linearly separable data that one could consider.

One class of probability models could be as follows. Select two different probability distributions in $\mathbb{R}^n$ with linearly separable supports. Sample positively labeled points from one distribution, and negatively labeled points from the second. The resulting points will be linearly separable.

A second class of probability models could to consider two arbitrary probability distributions in $\mathbb{R}^n$, one corresponding to each of the positive and negative classes. Upon sampling a point from a distribution, if that new point makes it so that the data are no longer linearly separable, then reject that point and sample again at random.

Under any such random model for linearly separable labeled points, we are interested in the following question. Which Radon configurations occur with positive probability, and what are those probabilities? The answers will of course depend on the random model selected.

The paper [24] considers the problem of computing the probability that two probabilistic point sets are linearly separable.

Question 7.0.4. If you toss two 6-sided die into the air and they collide, then (with probability one) they either collide vertex-to-face or edge-to-edge. The probability of a vertex-to-face collision is 46%, and the probability of an edge-to-edge collision 54%. The answer is computed using integral geometry. This is called Firey's dice problem; see [25–27] and [28, Pages 358–359].

When the two dice collide, it is analogous to a very special case of SVM where the support vectors from both classes all lie on the same hyperplane, i.e. the width of the margin is zero. A question that is perhaps more closely related to SVM is: what is the probability of the different types of “closest point configurations" when we specify that the dice are at exactly distance d apart? Once $d > 0$, the two closest points can both be vertices with positive probability, or the two closest points can be a vertex and a point on an edge. The vertex-and-face and edge-and-edge configurations still occur with positive probability. The Firey dice problem can perhaps be considered as a limit as $d \to 0$. This problem is also interesting for dice of any convex shape, for example tetrahedral dice.

There are many new directions to extend this research and we hope to pursue them at a later date.

Bibliography

[1] Yunqian Ma and Guodong Guo. *Support Vector Machines Applications*. Springer Publishing Company, Incorporated, 2014.

[2] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[3] Vojislav Kecman. Support vector machines—An introduction. In *Support vector machines: Theory and applications*, pages 1–47. Springer, 2005.

[4] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[5] K-R Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2):181–201, 2001.

[6] BB Peterson. The geometry of Radon's theorem. *American Mathematical Monthly*, pages 949–963, 1972.

[7] Jeremy Kun. Duality for the SVM. 2010.

[8] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft, April 1998.

[9] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[10] George E Sakr and Imad H Elhajj. VC-based confidence and credibility for support vector machines. *Soft Computing*, 20(1):133–147, 2016.

[11] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[12] Johann Radon. Mengen konvexer körper, die einen gemeinsamen punkt enthalten. *Mathematische Annalen*, 83(1-2):113–115, 1921.

[13] Kenneth L Clarkson, David Eppstein, Gary L Miller, Carl Sturtivant, and Shang-Hua Teng. Approximating center points with iterative radon points. *International Journal of Computational Geometry & Applications*, 6(03):357–377, 1996.

[14] Robin Hartshorne. *Algebraic geometry*, volume 52. Springer Science & Business Media, 2013.

[15] Gregor Kemper. *A course in commutative algebra*, volume 256. Springer Science & Business Media, 2010.

[16] Luis David Garcia Puente. Chapter 1: Varieties. http://www.shsu.edu/~ldg005/data/689/L1.pdf.

[17] D.S. Dummit and R.M. Foote. *Abstract Algebra*. Wiley, 2004.

[18] A.K.V. A. R. Vasishtha. *Matrices*. Krishna Prakashan, 1991.

[19] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale. *Complexity and real computation*. Springer Science & Business Media, 2012.

[20] Jiří Matoušek. *Using the Borsuk–Ulam theorem: Lectures on topological methods in combinatorics and geometry*. Springer, 2003.

[21] Jiří Matoušek. *Lectures on Discrete Geometry*. Springer-Verlag, Berlin, Heidelberg, 2002.

[22] Lynn Harold Loomis and Shlomo Sternberg. *Advanced calculus*. World Scientific, 1968.

[23] Chih-Chia Yao. Utilizing ellipsoid on support vector machines. In *Machine Learning and Cybernetics, 2008 International Conference on*, volume 6, pages 3373–3378. IEEE, 2008.

[24] Martin Fink, John Hershberger, Nirman Kumar, and Subhash Suri. Separability and convexity of probabilistic point sets.

[25] William J Firey. An integral-geometric meaning for lower order area functions of convex bodies. *Mathematika*, 19(2):205–212, 1972.

[26] William J Firey. Kinematic measures for sets of support figures. *Mathematika*, 21(2):270–281, 1974.

[27] P McMullen. A dice probability problem. *Mathematika*, 21(2):193–198, 1974.

[28] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Springer, 2008.