FUNDAMENTAL ISSUES IN SUPPORT VECTOR MACHINES

Samuel P. McWhorter

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2014

APPROVED:

Robert R.Kallman, Major Professor
Douglas Brozovic, Committee Member
Neal Brand, Committee Member
Su Gao, Chair of the Department of Mathematics
Mark Wardell, Dean of the Toulouse Graduate School

McWhorter, Samuel P. *Fundamental Issues in Support Vector Machines.* Doctor of Philosophy (Mathematics), May 2014, 48 pp., bibliography, 4 titles.

This dissertation considers certain issues in support vector machines (SVMs), including a description of their construction, aspects of certain exponential kernels used in some SVMs, and a presentation of an algorithm that computes the necessary elements of their operation with proof of convergence.

In its first section, this dissertation provides a reasonably complete description of SVMs and their theoretical basis, along with a few motivating examples and counter-examples. This section may be used as an accessible, stand-alone introduction to the subject of SVMs for the advanced undergraduate.

Its second section provides a proof of the positive-definiteness of a certain useful function here called E and defined as follows: Let V be a complex inner product space. Let N be a function that maps a vector v from V to its norm. Let p be a real number between 0 and 2 inclusive and for any v in V, let $f(v)$ be $N(v)$ raised to the p-th power. Finally, let a be a positive real number. Then $E(v)$ is $\exp(-af(v))$. Although the result is not new (other proofs are known but involve deep properties of stochastic processes) this proof is accessible to advanced undergraduates with a decent grasp of linear algebra.

Its final section presents an algorithm by Dr. Kallman (preprint), based on earlier Russian work by B.F. Mitchell, V.F Demyanov, and V.N. Malozemov, and proves its convergence. The section also discusses briefly architectural features of the algorithm expected to result in practical speed increases.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

1.1. Justification for Study: General Aspects of Support Vector Machines

A “support vector machine” is a variety of binary classifier, a procedure that, when given appropriate inputs, decides whether a case under consideration belongs in a certain category or a separate disjoint category. Such problems are virtually innumerable and include as examples: determining whether a potential loan recipient is a “good” or a “bad” credit risk depending on known credit information, determining whether a patient’s lab results indicate the presence or absence of a particular disease, and determining whether radar and visual sensor information indicates the presence of an enemy tank or not. Since any sort of discrete finite categorization scheme can be cast as a sequence of binary classifications, much as in the children’s game of “20 Questions,” the variety of “real world” problems potentially addressed by a support vector machine is vast. They are capable of basing their decisions on any information that can be expressed as a vector of real numbers which, by using elementary techniques from computer science, includes virtually any sort of information. Because of the breadth of their input and the generality of their results, it is clear that support vector machines are worthy of study.

We begin this study with a recognition of a debt: this dissertation owes much to Dr. Kallman’s article [2]. Though the proofs presented here are independent of the proofs there, except where explicitly noted, the fundamentals (separating suitable images of two sets by hyperplane, the “kernel trick,” the realization that the transform-plus-geometry and the kernel-based method of approach are identically expressive) are taken from there and the references therein. I deeply appreciate the support I received while developing the theory using the “Moore Method” and I hope the alternative approache shown here can both serve as an independent verification of what has been shown before and also provide a useful alternative method of presentation.

1.2. Geometric Classification: Obstacles to be Overcome

Support vector machines classify individuals by mapping them via their traits into a geometrical space. The mapping is designed so that the disjoint categories can be separated from each other by means of a hyperplane in the space. As real Hilbert spaces are the logical extension of Euclidean spaces to high- or infinite-dimensional settings, the mapping pairs individuals (by means of their traits) and points in a real Hilbert space. It is a strength of the techniques used that the mapping does not need to be calculated explicitly.

1.3. Lemmata and Proofs

As real Hilbert spaces and convex sets are important to the discussion to follow, a few important lemmata are here developed:

Lemma 1.1. *Let V be a real Hilbert space and D a non-empty convex subset of that space. There exists a unique vector v such that all minimizing sequences for $\|\cdot\|$ from D converge to it. Further, $\|v\| = \inf_{d\in D} \|d\|$, and for any $d \in D$, $\langle d - v, v\rangle \geq 0$.*

Proof. (after Kallman [2]:) Let $\eta = \inf_{d\in D} \|d\|$. Let $S = \{s_n\}_{n=1}^{\infty}$ be an arbitrary minimizing sequence for $\|\cdot\|$ from D. Since $\lim_{n\to\infty} \|s_n\| = \eta$ and squaring is a continuous function from $\mathbf{R}$ to $\mathbf{R}$, we know that $\lim_{n\to\infty} \|s_n\|^2 = \eta^2$, Note that for any s_m, s_n from the sequence, we have

$$\begin{aligned}
\left\|\frac{s_m + s_n}{2}\right\|^2 + \left\|\frac{s_m - s_n}{2}\right\|^2 &= \frac{1}{4}\left(\|s_m + s_n\|^2 + \|s_m - s_n\|^2\right) \\
&= \frac{1}{4}\left(\|s_m\|^2 + \|s_n\|^2 + 2\langle s_m, s_n\rangle + \|s_m\|^2 + \|s_n\|^2 - 2\langle s_m, s_n\rangle\right) \\
&= \frac{1}{2}\left(\|s_m\|^2 + \|s_n\|^2\right)
\end{aligned}$$

Since D is convex, $\frac{s_m+s_n}{2} \in D$, as are s_m and s_n. So $\left\|\frac{s_m+s_n}{2}\right\| \geq \eta$ and $-2\left\|\frac{s_m+s_n}{2}\right\|^2 \leq -2\eta^2$.

Continuing, we see:

$$\begin{aligned}
\left\| \frac{s_m + s_n}{2} \right\|^2 + \left\| \frac{s_m - s_n}{2} \right\|^2 &= \frac{1}{2}\left(\|s_m\|^2 + \|s_n\|^2\right) \\
\left\| \frac{s_m - s_n}{2} \right\|^2 &= \frac{1}{2}\left(\|s_m\|^2 + \|s_n\|^2 - 2\left\| \frac{s_m + s_n}{2} \right\|^2\right) \\
\|s_m - s_n\|^2 &= 2\left(\|s_m\|^2 + \|s_n\|^2 - 2\left\| \frac{s_m + s_n}{2} \right\|^2\right)
\end{aligned}$$

and

$$-2\left\| \frac{s_m + s_n}{2} \right\|^2 \leq -2\eta^2$$

$$\|s_m\|^2 + \|s_n\|^2 - 2\left\| \frac{s_m + s_n}{2} \right\|^2 \leq \left(\|s_m\|^2 - \eta^2\right) + \left(\|s_n\|^2 - \eta^2\right)$$

$$\|s_m - s_n\|^2 = 2\left(\|s_m\|^2 + \|s_n\|^2 - 2\left\| \frac{s_m + s_n}{2} \right\|^2\right) \leq 2\left(\left(\|s_m\|^2 - \eta^2\right) + \left(\|s_n\|^2 - \eta^2\right)\right)$$

Now let ϵ be an arbitrary positive number. Let N be such that for any $n > N$, $\left|\|s_n\|^2 - \eta^2\right| < \epsilon^2/4$, which number N exists because $\lim_{n\to\infty} \|s_n\|^2 = \eta^2$. Since $\|s_n\| \geq \eta$ for any valid index n, $\|s_n\|^2 \geq \eta^2$, so $0 \leq \|s_n\|^2 - \eta^2 < \epsilon^2/2$ for any $n > N$. This implies

$$\|s_m - s_n\|^2 \leq 2\left(\left(\|s_m\|^2 - \eta^2\right) + \left(\|s_n\|^2 - \eta^2\right)\right) < 2\left(\left(\frac{\epsilon^2}{4}\right) + \left(\frac{\epsilon^2}{4}\right)\right) = \epsilon^2$$

Therefore

$$\|s_m - s_n\| < \epsilon$$

Since ϵ was chosen arbitrarily, it is the case that for any $\epsilon > 0$, there exists N such that for any $m, n > N$, $\|s_m - s_n\| < \epsilon$. So the minimizing sequence is a Cauchy sequence, and V is a Hilbert space (hence complete), so the sequence converges to a vector v. Since $\|\cdot\|$ is continuous, $\|v\| = \|\lim_{n\to\infty} s_n\| = \lim_{n\to\infty} \|s_n\| = \eta = \inf_{d\in D} \|d\|$.

To see that any two minimizing sequences for $\|\cdot\|$ from D converge to the same value, merely note that interleaving two such minimizing sequences (taking $\{d_n\}_{n=1}^{\infty}$ and $\{e_n\}_{n=1}^{\infty}$

to generate a sequence $\{g_n\}_{n=1}^{\infty}$ by

$$g_n = \begin{cases} d_{\frac{n+1}{2}} & n \text{ odd} \\ e_{\frac{n}{2}} & n \text{ even} \end{cases}$$

) is itself such a minimizing sequence, hence converges to some value. Since each (infinite) subsequence of a convergent sequence converges to the same limit, it follows that d and e converge to the same value. Since d and e were arbitrary, it follows that any such minimizing sequences all converge to the same value v.

Finally, let d be an arbitrary element of D and let $\{s_n\}_{n=1}^{\infty}$ be a minimizing sequence for $\|\cdot\|$ from D. For any $\lambda \in [0,1]$ and any natural number n, since D is convex, $\lambda d + (1-\lambda)s_n \in D$. So we must have:

$$\|s_n + \lambda(d - s_n)\|^2 = \|s_n\|^2 + 2\lambda \langle d - s_n, s_n\rangle + \lambda^2 \|d\|^2 \geq \eta^2$$

Since this is true for all natural numbers n, and since $\langle \cdot, \cdot \rangle$ is a continuous function on V (as are addition and scalar multiplication), it follows that

$$\begin{aligned} \lim_{n\to\infty} \left(\|s_n\|^2 + 2\lambda \langle d - s_n, s_n\rangle + \lambda^2 \|d\|^2\right) &\geq \eta^2 \\ \left\|\lim_{n\to\infty} s_n\right\|^2 + 2\lambda \left\langle d - \lim_{n\to\infty} s_n, \lim_{n\to\infty} s_n\right\rangle + \lambda^2 \|d\|^2 &\geq \eta^2 \\ \|v\|^2 + 2\lambda \langle d - v, v\rangle + \lambda^2 \|d\|^2 &\geq \eta^2 \end{aligned}$$

Since $\eta = \|v\|$,

$$\begin{aligned} \eta^2 + 2\lambda \langle d - v, v\rangle + \lambda^2 \|d\|^2 &\geq \eta^2 \\ 2\lambda \langle d - v, v\rangle + \lambda^2 \|d\|^2 &\geq 0 \\ \langle d - v, v\rangle &\geq -\frac{\lambda}{2} \|d\|^2 \end{aligned}$$

where the last equation is true for $\lambda \in (0,1]$. Since the final inequality is true for all $\lambda \in (0,1]$, it follows that:

$$\begin{aligned} \langle d - v, v\rangle &\geq \lim_{\lambda\to 0} -\frac{\lambda}{2} \|d\|^2 \\ \langle d - v, v\rangle &\geq -\frac{\lim_{\lambda\to 0} \lambda}{2} \|d\|^2 = 0 \\ \langle d - v, v\rangle &\geq 0 \end{aligned}$$

where the second-to-last step follows by the continuity of scalar multiplication. □

LEMMA 1.2. *Let V be a Hilbert space, and C and W non-empty subsets of V, and H_C and H_W be their respective convex hulls. Let $D = \{w - c \mid w \in H_W \wedge c \in H_C\}$. Let E be the convex hull of $F = \{w - c \mid w \in W \wedge c \in C\}$. Then $D = E$ (and D is therefore a convex set).*

PROOF. Let d be an arbitrary element in D. Then d is equal to $u - v$ where $u \in H_W$ and $v \in H_C$. Then there exist natural numbers m and n, non-negative scalars $\alpha_1, \ldots, \alpha_m$ where $\sum_{i=1}^m \alpha_i = 1$, non-negative scalars $\beta_1, \ldots, \beta_n$ where $\sum_{j=1}^n \beta_j = 1$, vectors $w_1, \ldots, w_m$ from W such that $u = \sum_{i=1}^m \alpha_i w_i$, and vectors $c_1, \ldots, c_n$ from C such that $v = \sum_{j=1}^n \beta_j c_j$. Expressing d in terms of these elements gives us:

$$\begin{aligned}
d &= u - v \\
&= \left(\sum_{i=1}^m \alpha_i w_i\right) - \left(\sum_{j=1}^n \beta_j c_j\right) \\
&= \left(\sum_{i=1}^m \left(\sum_{j=1}^n \beta_j\right) \alpha_i w_i\right) - \left(\sum_{j=1}^n \left(\sum_{i=1}^m \alpha_i\right) \beta_j c_j\right) \\
&= \left(\sum_{i=1}^m \left(\sum_{j=1}^n \beta_j \alpha_i w_i\right)\right) - \left(\sum_{j=1}^n \left(\sum_{i=1}^m \alpha_i \beta_j c_j\right)\right) \\
&= \left(\sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j w_i\right) - \left(\sum_{j=1}^n \sum_{i=1}^m \alpha_i \beta_j c_j\right) \\
&= \left(\sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j w_i\right) - \left(\sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j c_j\right) \\
&= \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j (w_i - c_j)
\end{aligned}$$

For each $(i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$ the scalar given by $\alpha_i \beta_j$ is non-negative and vector given by $w_i - c_j$ is in F. The sum of $\alpha_i \beta_j$ over all such pairs (i, j) is $\sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j = \left(\sum_{i=1}^m \alpha_i\right) \left(\sum_{j=1}^n \beta_j\right) = 1$. Consequently, $d = \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j (w_i - c_j)$ is also in E. Since d was an arbitrary element of D, it follows that $D \subseteq E$.

Now let e be an arbitrary element of E. There exists a natural number n, non-negative scalars $\gamma_1, \ldots, \gamma_n$ that sum to 1, elements $w_1, \ldots, w_n$ of W, and elements $c_1, \ldots, c_n$ of C, such that $e = \sum_{i=1}^{n} \gamma_i(w_i - c_i)$. Let $u = \sum_{i=1}^{n} \gamma_i w_i$ and $v = \sum_{i=1}^{n} \gamma_i c_i$ and note that $e = u - v$. Note also that $u \in H_W$ and $v \in H_C$, so $e \in D$. Since e was an arbitrary element of E, it follows that $E \subseteq D$. Since $D \subseteq E$ and $E \subseteq D$, $D = E$. □

Lemma 1.3. *Let P be a hyperplane in a real Hilbert space V. Then there exists $v \neq 0$ and a scalar c such that $P = \{x \mid x \in V \wedge \langle x, v \rangle = q\}$. Further: if there are two vector/scalar pairs (v_1, q_1) and (v_2, q_2) that give rise to P in this way (i.e. $P = \{x \mid x \in V \wedge \langle x, v_1 \rangle = q_1\}$ and $P = \{x \mid x \in V \wedge \langle x, v_2 \rangle = q_2\}$, then $v_2 = \alpha v_1$ where α is a non-zero scalar, and neither v_1 nor v_2 are 0.*

Proof. Since P is a hyperplane, P is closed. Since P is an affine subspace of V, there exists H such that H is a vector subspace of V and at least one vector $c \in V$ such that $P = \{h + c \mid h \in H\}$. H is then a parallel hyperplane that contains the origin. Since P has co-dimension 1, H does as well, so there exists a vector d such that $d \notin H$. Consider the set $T = \{h + d \mid h \in H\}$. Since $d \notin H$ and H is a group under addition, $-d \notin H$ so $0 \notin T$. Since H was closed, T is closed as well, and since T is closed and does not contain 0, there's an open ball of non-zero radius separating 0 from T. Let $\eta = \inf_{t \in T} \|t\|$ and note that η must therefore be greater than 0. Since T is convex and non-empty ($0 \in H$, so $0 + d = d \in T$), by lemma 1.1, there exists a vector v such that any minimizing sequence for $\|\cdot\|$ from T converges to v, where $\|v\| = \eta > 0$. Since v is the limit of a sequence of elements of T, and since T is closed, $v \in T$. So there exists $h' \in H$ such that $h' + d = v$. So the set $\{h + v \mid h \in H\} = \{h + h' + d \mid h \in H\}$. Since $H + h' = H$, this is just $\{h + d \mid h \in H\}$, i.e. T. From the referenced lemma, we know that $\langle t - v, v \rangle \geq 0$ for all $t \in T$, so $\langle h + v - v, v \rangle = \langle h, v \rangle \geq 0$ for all $h \in H$. If $\langle h, v \rangle \neq 0$ for any $h \in H$, then $-\frac{1}{\langle h, v \rangle} h \in H$ as well, since H is closed under scalar multiplication, so $\left\langle -\frac{1}{\langle h, v \rangle} h, v \right\rangle = -\frac{1}{\langle h, v \rangle} \langle h, v \rangle = -1 < 0$. Since this contradicts $\langle h, v \rangle \geq 0$, it follows that there can exist no elements $h \in H$ such that $\langle h, v \rangle \neq 0$. So $\langle h, v \rangle = 0$ for all $h \in H$. Now consider the set $E = \{x \mid x \in V \wedge \langle x, v \rangle = \langle c, v \rangle\}$

(where c is as defined earlier). Since $v \notin H$ and $v \neq 0$ and H has co-dimension 1, it follows that any element $x \in V$ can be expressed in the form $\alpha v + h$, for some $\alpha \in \mathbf{R}$ and $h \in H$. So $c = \alpha_c v + h_c$ for some scalar α_c and some $h_c \in H$, and we can express E in the form $\{\alpha v + h \mid \alpha \in \mathbf{R} \wedge h \in H \wedge \langle \alpha v + h, v\rangle = \langle c, v\rangle\}$. $\langle \alpha v + h, v\rangle = \alpha \langle v, v\rangle + \langle h, v\rangle = \alpha \langle v, v\rangle$ and $\langle c, v\rangle = \langle \alpha_c v + h_c, v\rangle = \alpha_c \langle v, v\rangle + \langle h_c, v\rangle = \alpha_c \langle v, v\rangle$. So $\langle \alpha v + h, v\rangle = \langle c, v\rangle$ when and only when $\alpha \langle v, v\rangle = \alpha_c \langle v, v\rangle$. Since $v \neq 0$, this is true when and only when $\alpha = \alpha_c$. So $E = \{\alpha_c v + h \mid h \in H\}$. Since $h_c \in H$ and H is a group under addition, the sets $\{h \mid h \in H\}$ and $\{h + h_c \mid h \in H\}$ are identical, so $E = \{\alpha_c v + h_c + h \mid h \in H\} = \{h + c \mid h \in H\}$, which is exactly the original hyperplane P. So the hyperplane P can be expressed in the form $P = \{x \mid x \in V \wedge \langle x, v\rangle = q\}$ where $v \neq 0$ and $q = \langle c, v\rangle$.

Finally, let there be two vectors v_1 and v_2 and scalars q_1 and q_2 such that $P = \{x \mid x \in V \wedge \langle x, v_1\rangle = q_1\} = \{x \mid x \in V \wedge \langle x, v_2\rangle = q_2\}$. If v_j were 0 for some $j \in \{1, 2\}$, then for all $x \in V$, $\langle x, v_j\rangle = 0$, so $P = V$ or $P = \emptyset$ as $q_j = 0$ or $q_j \neq 0$. Since neither V nor $\emptyset$ are affine subspaces of V of co-dimension 1, it follows that $v_j \neq 0$ for any $j \in \{1, 2\}$. Let c be a fixed arbitrary element of P and consider the set $H = \{v - c \mid v \in P\}$. Let $j \in \{1, 2\}$ be arbitrary. H may also be re-written:

$$H = \{v - c \mid v \in V \wedge \langle v, v_j\rangle = q_j\}$$

$$= \{w \mid w + c \in V \wedge \langle w + c, v_j\rangle = q_j\}$$

and since V is a group under addition,

$$H = \{w \mid w \in V \wedge \langle w + c, v_j\rangle = q_j\}$$

$$= \{w \mid w \in V \wedge \langle w, v_j\rangle + \langle c, v_j\rangle = q_j\}$$

and since $c \in P$,

$$H = \{w \mid w \in V \wedge \langle w, v_j\rangle + q_j = q_j\}$$

$$= \{w \mid w \in V \wedge \langle w, v_j\rangle = 0\}$$

so H is the kernel of the map $l_j \colon V \to \mathbf{R}$ given by $l_j(w) = \langle w, v_j\rangle$. Since $\langle v_j, v_j\rangle \neq 0$, $l_j(v_j) \neq 0$, so $v_j \notin H$. This is true for any $j \in \{1, 2\}$, so for arbitrary $j \in \{1, 2\}$, $l_1(v_j) \neq 0$ and $l_2(v_j) \neq 0$. Consider the map $p \colon V \to V$ given by $p(x) = \frac{l_1(x)}{l_1(v_1)} v_1$. For any $x \in V$, consider $l_1(x-p(x))$: $l_1(x-p(x)) = l_1(x)-l_1(p(x)) = l_1(x)-l_1(\frac{l_1(x)}{l_1(v_1)}v_1) = l_1(x)-\frac{l_1(x)}{l_1(v_1)}l_1(v_1) = l_1(x) - l_1(x) = 0$. Therefore, $x - p(x) \in H$ for any $i \in \{1, 2\}$. Thus, any vector $x \in V$ can be expressed in the form $x = h + \alpha v_1$ for some $h \in H$ and scalar α. So $v_2 = h + \alpha v_1$ for some $h \in H$ and $\alpha \in \mathbf{R}$. Since $0 = \langle h, v_2\rangle = \langle h, v_1\rangle$, we have: $0 = \langle h, v_2\rangle = \langle h, h + \alpha v_1\rangle = \langle h, h\rangle + \alpha \langle h, v_1\rangle = \langle h, h\rangle + \alpha \cdot 0 = \langle h, h\rangle$, Since $\|h\|^2 = 0$, $\|h\| = 0$ and $h = 0$. So $v_2 = \alpha v_1$. Since $v_2 \neq 0$, $\alpha \neq 0$. □

1.4. Motivating Examples

Now that our introductory lemmata have been proven, let us consider a few illustrative examples of situations that can be (and some that cannot easily be) represented by support vector machines.

Example 1.4. Assume for the moment that we're separating the sheep from the goats in a flock composed of both. We might guess that the features that distinguish sheep from goats might be some combination of: number of horns, straightness of hair (perhaps on a scale of 0 to 10 with 0 being straightest and 10 being curliest), weight in pounds, height at the shoulder in inches, and average number of tin cans eaten per day. So we start with the Cartesian product $\mathbf{Z} \times \mathbf{R}^3 \times \mathbf{Q}$. In this space, a flock member might be represented by the tuple $(2, 0, 50, 36, 7)$ if he or she had two horns, perfectly straight hair, weighed 50 lbs, stood 36 inches tall at the shoulder, and ate an average of 7 tin cans per day. There are certain points in this space that cannot represent any goat (for example, $(-1, 10, 50, 36, 7)$ as negative horn counts make no sense), so we consider only the subset of that space that can represent an actual flock member. We call this subset the "feature space" and represent it by $\mathcal{F}$.

Notionally, there is a function Φ that serves to map elements from $\mathcal{F}$ (the data from our flock) into a geometric setting in such a way that the "sheep" are mapped "away from" the "goats". Since real Hilbert spaces are, in some sense, the natural generalization of finite-dimensional Euclidean space, we let Φ map $\mathcal{F}$ into a real Hilbert space V (with possibly infinite, perhaps even uncountable, dimension). With careful construction of the notional mapping function Φ, one can almost literally draw a “line in the sand” by determining a hyperplane in V (an affine subspace of co-dimension 1) that separates the categories. If a member of the flock, represented by an element x of $\mathcal{F}$, is a sheep, then $\Phi(x)$ will be on one side of the separating hyperplane. If the member of the flock is a goat, then $\Phi(x)$ will be on the other side of the separating hyperplane.

Separation by drawing “lines in the sand” may be an appealing rhetorical device but is often difficult to achieve.

Example 1.5. Consider a set S_0 in $\mathbf{R}^2$ given by $\{(x, y) \mid x^2 + y^2 \leq 1\}$ and a set S_1 given by $\{(x, y) \mid r^2 \leq x^2 + y^2 \leq (r + 1)^2\}$, where $r > 1$. Let r be as large as necessary for S_0 to be considered “separate from” S_1. Since S_0 is the closed unit disc and S_1 is a closed ring of inner radius r (and outer radius $r + 1$) centered on the origin, it is intuitively clear that there exists no separating hyperplane (separating line in this case) of the desired sort.

In some sense, this is a failure of convexity: S_0 is clearly in the convex hull of S_1, but if two non-empty sets can be separated by hyperplane, their convex hulls are separated by the same hyperplane.

Another failure of separation is of a more subtle sort and involves difficulty in resolving on what side of a separating hyperplane an individual's image may lie.

Example 1.6. Let S_0 be the subset of $\mathbf{R}^2$ defined by $S_0 = \{(x, y) \mid x > 0 \wedge y > 1/x\}$ and let $S_1 = \{(x, y) \mid x > 0 \wedge y < 0\}$. S_0 and S_1 are clearly disjoint. Separating hyperplanes in this case will again be straight lines. Any line other than $y = 0$ will have non-zero y-value for some positive value of x, hence will intersect either S_0 or S_1 and fail to separate the two sets. The line $y = 0$ does indeed separate the two sets S_0 and S_1, as all points in S_0 have positive y-cöordinate and all points in S_1 have negative y-cöordinate. Consider however the case of an individual mapped by our hypothetical Φ function to a point with great x-value but negligible y-value. *Any uncertainty in the value of y, whatever the cause, may very well result in the mis-classification of this point and similar points.* The result is that a classifier based on this function Φ would not be robust. This problem can exist in principle whenever the classification values (here, simply the y cöordinates of the images under Φ of individuals) for each category (here S_0 and S_1) can be arbitrarily close together.

The sets S_0 and S_1 are separated in the sense that they have disjoint convex hulls, unlike in example 1.5, but they're not separated well enough to support a robust classifier based on them, as there are cases where even small errors can result in a point from S_0 being categorized with points from S_1, and vice-versa.

1.5. Separation by Hyperplane: Separating Functions.

Are these the only significant problems appearing when trying to separate sets by hyperplane? In other words, given two non-empty sets C and W with convex hulls at a non-zero distance from each other, can they be definitively separated by hyperplane?

The answer to this question is, unsurprisingly, yes. Let us state the result as a theorem for later reference:

Theorem 1.7. *Let V be a real Hilbert space. Let C and W be non-empty sets with convex hulls H_C and H_W. Let $\eta = \inf_{c \in H_C \wedge w \in H_W} \|w - c\|$. Then H_C and H_W can be well-separated by an affine hyperplane H if and only if $\eta > 0$. When this is true, there then exists an affine hyperplane $P = \{x \mid x \in V \wedge \langle x, v \rangle = a\}$ that separates H_C and H_W well, where $\|v\| = \eta > 0$, $\langle (w - c) - v, v \rangle \geq 0$ for all $c \in H_C$ and $w \in H_W$, and v is the limit of any sequence $\{s_n\}_{n=1}^{\infty}$ where there exists for each natural number n $c \in H_C$ and $w \in H_W$ such that $s_n = w - c$ and where $\lim_{n \to \infty} \|s_n\| = \eta$.*

Proof. Let us consider in a real Hilbert space V two non-empty sets C and W where H_C and H_W are well-separated by a hyperplane H. By lemma 1.3, H can be expressed as $\{x \mid x \in V \wedge \langle x, v \rangle = p\}$ where v is a fixed non-zero vector and p a scalar. Let $l(x) = \langle x, v \rangle$. Now consider the images of C, W, and H under l. By our selection of v, $l(H)$ contains the single point p. For p to separate $l(C)$ from $l(W)$, either $l(C) \subset (-\infty, p]$ and $l(W) \subset [p, \infty)$, or $l(C) \subset [p, \infty)$ and $l(W) \subset (-\infty, p]$. Without loss of generality (either through relabeling the sets C and W, or choosing $-v$ and $-p$ instead of v and p for the form of our expression for the set H), we assume $l(C) \subset (-\infty, p]$ and $l(W) \subset [p, \infty)$. Let $c_* = \sup_{c \in H_C} l(c)$ and

$w_* = \inf_{w \in H_W} l(w)$. Since H_C and H_W are well-separated by this hyperplane, there must be a gap between c_* and w_*: $|w_* - c_*| > 0$, and p must fall within this gap: $p \in (w_*, c_*)$. H separates the closure of their convex hulls as well: the closure of H_C is a subset of the closed set $l^{-1}((-\infty, c_*])$ and the closure of H_W is a subset of the closed set $l^{-1}([w_*, \infty))$. To see that the closed convex hulls are separated by a non-zero distance (hence H_C and H_W themselves), note that l is uniformly continuous. So for $\epsilon = w_* - c_*$, there exists $\delta > 0$ such that for any two points x_1 and x_2 of the Hilbert space v, if $\|x_1 - x_2\| < \delta$, $|l(x_1) - l(x_2)| < w_* - c_*$. Since for any two points $c \in \overline{H_C}$ and $w \in \overline{H_W}$ $|l(c) - l(w)| \geq w_* - c_*$, it follows that $\|c - w\| \geq \delta > 0$. So $\overline{H_C}$ and $\overline{H_W}$ necessarily have non-zero separation, as do H_C and H_W consequently.

To show the sufficiency of these conditions, assume that C and W are non-empty subsets of the Hilbert space V, let H_C and H_W be their respective convex hulls, and $\eta > 0$ be their distance from each other ($\eta = \inf_{c \in H_C \wedge w \in H_W} \|w - c\|$). Let $D = \{w - c \mid c \in H_C \wedge w \in H_W\}$. Lemma 1.2 tells us that D is convex and obviously $\inf_{d \in D} \|d\| = \eta$. Lemma 1.1 tells us that there exists a unique vector v in the closure of D that is the limit of any minimizing sequence for $\|\cdot\|$ from D with $\|v\|$ consequently being $\inf_{d \in D} \|d\| = \eta$, and for any $d \in D$, $\langle d - v, v \rangle \geq 0$, so for any $c \in H_C$ and $w \in H_W$, $\langle (w - c) - v, v \rangle \geq 0$. Let $c \in H_C$ and $w \in H_W$. Since $\langle (w - c) - v, v \rangle \geq 0$, we have:

$$\langle (w - c) - v, v \rangle \geq 0$$

$$\langle w, v \rangle - \langle c, v \rangle - \langle v, v \rangle \geq 0$$

$$\langle w, v \rangle - \langle c, v \rangle \geq \langle v, v \rangle$$

Temporarily fix c and note that since this is true for all $w \in H_W$, the set $\{\langle w, v \rangle \mid w \in H_W\}$ is bounded from below by $\langle c, v \rangle + \|v\|^2$. So $w_* = \inf_{w \in H_W} \langle w, v \rangle$ exists and $w_* - \langle c, v \rangle \geq \|v\|^2$. Since $w_* - \langle c, v \rangle \geq \|v\|^2$ is true for any $c \in H_C$, the set $\{\langle c, v \rangle \mid c \in H_C\}$ is bounded from above by $w_* - \|v\|^2$, so it follows that $c_* = \sup_{c \in H_C} \langle c, v \rangle$ exists and $w_* - c_* \geq \|v\|^2$. So the linear functional $l(x) = \langle x, v \rangle$ expresses a clear separation between the sets H_C and H_W. Let $a = \frac{c_* + w_*}{2}$. Consider the affine hyperplane defined by $P = \{x \mid x \in V \wedge \langle x, v \rangle = a\}$. Since

$c_* < w_*$, $c_* < a < w_*$, so $w_* - a = w_* - \frac{c_*+w_*}{2} = \frac{w_*-c_*}{2}$ and $a - c_* = \frac{c_*+w_*}{2} - c_* = \frac{w_*-c_*}{2}$. Thus, this hyperplane cleanly (and symmetrically!) separates the images of H_C and H_W. □

So for robust separation of non-empty sets by hyperplane it is necessary and sufficient to have their convex hulls be at a strictly positive distance from each other.

In general, depending on the nature of Φ, there may be many hyperplanes that suffice to divide the images of the two categories. Let C_1 and C_2 be subsets of $\mathcal{F}$ containing the representatives of the two categories. (In example 1.4 above, C_1 would consist of the images of each "sheep" in $\mathcal{F}$ and C_2 consist of the images of each "goat" in $\mathcal{F}$.) Assume that Φ has been selected in such a way that the convex hulls of $C = \Phi(C_1)$ and $W = \Phi(C_2)$ have non-zero separation. We select for our hyperplane the hyperplane P whose existence was proven in theorem 1.7 above, the associated vector v, and associated scalar a.

Given the function Φ, the process of constructing a support vector machine is then to determine this vector v, associated linear functional $l(x) = \langle x, v \rangle$, and associated hyperplane $H = l^{-1}(\{a\})$. From this, we then construct a function $f \colon \mathcal{F} \to \mathbf{R}$ given by $f = l \circ \Phi$. We can call this function "f" a "separating function." (One algorithm for generating such a vector v will be given in section 3.) By our selection of f, $f(C_1)$ either lies entirely below a or above a, and $f(C_2)$ lies entirely above or below a respectively. Without loss of generality, assume $f(C_1)$ lies entirely below a and $f(C_2)$ above a. f is our sought-after classifying function: Any individual u represented in $\mathcal{F}$ can be classified as to membership by checking whether $f(u) >$, $<$, or $= a$. If $f(u) < a$, we consider u to be in the category C_1. If $f(u) > a$, we consider it to belong to category C_2. If $f(u) = a$, it may be with equal justice be assigned to C_1 or C_2 — an appropriate tie-breaking procedure will need to be determined.

In some sense, the Hilbert space V is likely larger than it needs to be. Consider: since individuals are chosen from our feature space $\mathcal{F}$, we only need V to be a Hilbert space large enough to contain $\operatorname{span} \Phi(\mathcal{F})$. Since $C_1 \subseteq \mathcal{F}$ and $C_2 \subseteq \mathcal{F}$, $\Phi(C_1) \subseteq \Phi(\mathcal{F})$ and $\Phi(C_2) \subseteq \Phi(\mathcal{F})$. Let $C = \Phi(C_1)$ and $W = \Phi(C_2)$ and note that both $C \subseteq \Phi(\mathcal{F})$ and $W \subseteq \Phi(\mathcal{F})$. Let H_C be the convex hull of C and H_W be the convex hull of W and note that both are subsets of $\operatorname{span} \Phi(\mathcal{F})$. Since v is the limit of differences of elements from H_C and H_W, v is the limit of elements from $\operatorname{span}\Phi(\mathcal{F})$, hence $v \in \overline{\operatorname{span}\Phi(\mathcal{F})}$. Since $\overline{\operatorname{span}\Phi(\mathcal{F})}$ is the smallest Hilbert space containing $\Phi(\mathcal{F})$, and since our separating function has domain only of $\mathcal{F}$, we can without loss of generality assume that $V = \overline{\operatorname{span}\Phi(\mathcal{F})}$ or, equivalently, $\operatorname{span}\Phi(\mathcal{F})$ is dense in V.

It will be useful in further developments to express our separating function f more directly in terms of Φ and the inner product relation. Recall that v is the limit of a sequence of vectors from the set $D = \{w - c \mid c \in H_C \wedge w \in H_W\}$, which by lemma 1.2 is just the convex hull of the set $F = \{w - c \mid w \in W \wedge c \in C\}$. Choose one such sequence $\{s_n\}_{n=1}^{\infty}$ with elements in D converging to v. For each natural number n, let $s_n = \sum_{i=1}^{q_n} \alpha_{n,i}(w_{n,i} - c_{n,i})$ where q_n is a natural number and additionally, for each integer i between 1 and q_n inclusive, let $\alpha_{n,i}$ be a non-negative scalar, $w_{n,i}$ be an element of W, and $c_{n,i}$ be an element of C. Let the $\alpha_{n,i}$ satisfy the additional restriction that for each natural number n, $\sum_{i=1}^{q_n} \alpha_{n,i} = 1$. (As D is the convex hull of F, each s_n can be so written.) For each natural number n and integer i between 1 and q_n inclusive, because $W = \Phi(C_2)$, we can express each element $w_{n,i}$ as the image under Φ of an element $x_{n,i}$ from C_2, and because $C = \Phi(C_1)$, we can express each element $c_{n,i}$ as the image under Φ of an element $y_{n,i}$ from C_1. We can now express f in the following fashion:

$$\begin{aligned} f(x) = \langle \Phi(x), v \rangle &= \lim_{n\to\infty} \left\langle \Phi(x), \sum_{i=1}^{q_n} \alpha_{n,i}(w_{n,i} - c_{n,i}) \right\rangle \\ &= \lim_{n\to\infty} \sum_{i=1}^{q_n} \alpha_{n,i} \left(\langle \Phi(x), \Phi(x_{n,i}) \rangle - \langle \Phi(x), \Phi(y_{n,i}) \rangle \right) \end{aligned} \tag{1}$$

This equation:

$$f(x) = \lim_{n\to\infty} \sum_{i=1}^{q_n} \alpha_{n,i} \left(\langle \Phi(x), \Phi(x_{n,i}) \rangle - \langle \Phi(x), \Phi(y_{n,i}) \rangle \right) \tag{2}$$

will be of considerable use in later development of the theory.

Although our original example refers to sheep and goats in an imaginary flock, the arbitrariness of the choice of $\mathcal{F}$ and of Φ shows that the general method can be adapted to any non-trivial binary classification situation. Indeed, the situation can be reduced to an absurdity: any binary classification scheme can be expressed as a set of individuals

U, a subset C_1, and a subset C_2 where $\{C_1, C_2\}$ is a partition of U. Define a mapping $m\colon U \to \{-1, 1\}$ as $m(u) = -1$ when $u \in C_1$ and $m(u) = 1$ otherwise (when $u \in C_2$). The feature space $\mathcal{F}$ would then be $\{-1, 1\}$. The Hilbert space V would just be $\mathbf{R}$ and Φ would be the natural embedding e of $\mathcal{F}$ into $\mathbf{R}$. The associated vector "v" could be any non-zero real number, and the corresponding separating hyperplane would simply be the null vector.

While this technique can therefore accommodate any binary classification scheme, it appears to do so at a high price. Although a feature space $\mathcal{F}$ can be constructed simply by including any information that might conceivably be relevant, construction of the function Φ would seem to require deep understanding of the particular classification scheme to be implemented. As an example, support vector machines have been used to recognize pictures of characters, including the character "a." Implementation of a function to recognize the letter "a" would seem to require detailed knowledge of just how to recognize that letter at any size, with any acceptable shape of glyph, possibly in the presence of other characters, in some variety of possible orientations.

1.6. Kernel Functions

Let $\mathcal{F}$, Φ, V, C_1, C_2, v, f be as in our discussion in the previous section. In that section, we've presumed the existence of a function Φ. Notice that in equation 2 for our separating function f, however, Φ appears only in combination with the inner product $\langle \cdot, \cdot \rangle$ on V.

Is it possible that we can somehow specify this combined function without having to specify both explicitly? In other words, is there some function $\kappa\colon \mathcal{F}^2 \to \mathbf{R}$ such that $\kappa(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$ for each $(x_1, x_2) \in \mathcal{F}^2$?

This is indeed possible. As far as determining a separating function goes, specifying a unified function with certain properties has the same expressive power as separately specifying a Φ and a Hilbert space V, as we will now show. (The fundamental notion here, the "kernel trick", is a well-known process.)

Assume a feature space $\mathcal{F}$, a Hilbert space V with inner product $\langle \cdot, \cdot \rangle)$, and a function $\Phi\colon \mathcal{F} \to V$ where $\operatorname{span} \Phi(\mathcal{F})$ is dense in V. Consider a function $\kappa\colon \mathcal{F}^2 \to \mathbf{R}$ defined as

$\kappa(x,y) = \langle \Phi(x), \Phi(y) \rangle$. Note that $\kappa(p,q) = \langle \Phi(p), \Phi(q) \rangle = \langle \Phi(q), \Phi(p) \rangle = \kappa(q,p)$, so κ is a symmetric function. Note also that for any natural number n, real constants $\{c_1, \ldots, c_n\}$, and elements of $\mathcal{F}$ $\{x_1, \ldots, x_n\}$,

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \kappa(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \left\langle \Phi(x_i), \Phi(x_j) \right\rangle \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle c_i \Phi(x_i), c_j \Phi(x_j) \right\rangle \\
&= \sum_{i=1}^{n} \left\langle c_i \Phi(x_i), \sum_{j=1}^{n} c_j \Phi(x_j) \right\rangle \\
&= \left\langle \sum_{i=1}^{n} c_i \Phi(x_i), \sum_{j=1}^{n} c_j \Phi(x_j) \right\rangle \\
&= \left\| \sum_{i=1}^{n} c_i \Phi(x_i) \right\|^2 \geq 0
\end{aligned}$$

So κ as specified is also a positive-definite function. Our separating function (call it f) can then be specified in terms of κ: we re-cast equation 2 as

$$\begin{aligned}
f(x) &= \lim_{n \to \infty} \sum_{i=1}^{q_n} \alpha_{n,i} \left(\langle \Phi(x), \Phi(x_{n,i}) \rangle - \langle \Phi(x), \Phi(y_{n,i}) \rangle \right) \\
&= \lim_{n \to \infty} \sum_{i=1}^{q_n} \alpha_{n,i} \left(\kappa(x, x_{n,i}) - \kappa(x, y_{n,i}) \right)
\end{aligned}$$

So a positive-definite function directly from the feature space $\mathcal{F}$ to $\mathbf{R}$ can summarize any Φ and Hilbert space V we require. Is there any other salient restriction that κ must satisfy? The answer turns out to be "no," as we see below. We will show that given any positive-definite function $\kappa \colon \mathcal{F}^2 \to \mathbf{R}$, we can create a Hilbert space $(V, \langle \cdot, \cdot \rangle_V)$ and a map $\Phi \colon \mathcal{F} \to V$ such that $\operatorname{span} \Phi(\mathcal{F})$ is dense in V and for any vectors $v_1 = \Phi(x_1)$, $v_2 = \Phi(x_2)$, $\langle v_1, v_2 \rangle = \kappa(x_1, x_2)$.

Let $\mathcal{F}$ be a non-empty set (presumed a feature space) and let $\kappa \colon \mathcal{F}^2 \to \mathbf{R}$ be a symmetric positive-definite function. From κ, we can construct a real Hilbert space V with inner product $\langle \cdot, \cdot \rangle$ and feature space map $\Phi \colon \mathcal{F} \to V$ such that $\kappa(x,y) = \langle \Phi(x), \Phi(y) \rangle$.

We begin the construction by letting U be the set of real-valued functions on $\mathcal{F}$ with finite support. Addition on U is point-wise addition: if $f, g \in U$, $(f+g)(x) = f(x) + g(x)$ for each $x \in \mathcal{F}$. (The support of $f+g$ is a subset of $\operatorname{supp}(f) \cup \operatorname{supp}(g)$, hence finite.) The field of scalars will be $\mathbf{R}$ and scalar multiplication will be standard multiplication of a function by a number: if $f \in U$ and $c \in \mathbf{R}$, $(c \cdot f)(x) = cf(x)$ for each $x \in \mathcal{F}$ (and $(c \cdot f)$'s support will either be f's support or the empty set as the scalar c is not or is zero). It is clear that under these conditions, U is a real vector space. For each $x \in \mathcal{F}$, let δ_x be the function that is 1 when its argument is x and 0 otherwise. δ_x clearly has finite support. We embed Φ in U by $x \mapsto \delta_x$ for each $x \in \mathcal{F}$, and this map is clearly injective. It will be useful to note that $U = \operatorname{span} \Phi(\mathcal{F})$: any element $f \in U$ can be expressed as a linear combination of the functions δ_x. Since f has finite support, let $\operatorname{supp}(f) = \{x_1, \ldots, x_n\}$. Then consider the function $w(x) = \sum_{i=1}^{n} f(x_i)\delta_{x_i}$: its set of support is exactly that of f by construction: for each $x_j \in \operatorname{supp}(f)$, $w(x_j) = \sum_{i=1}^{n} f(x_i)\delta_{x_i}(x_j) = f(x_j)\delta_{x_j}(x_j) = f(x_j) \neq 0$, and if $x \notin \operatorname{supp}((f))$, $\delta_{x_i}(x) = 0$ for each $i \in \{1, \ldots, n\}$, so $w(x) = 0$. Since w is zero when and only when f is, and for elements of x where f is not zero, w agrees with f, $w = f$. So any arbitrary element of U can be expressed as the linear combination of elements of $\Phi(\mathcal{F})$, so $U = \operatorname{span} \Phi(\mathcal{F})$.

Define a function $l \colon U \times U \to \mathbf{R}$ by $l(f,g) = \sum_{i \in \operatorname{supp}(f)} \sum_{j \in \operatorname{supp}(g)} f(i)g(j)\kappa(i,j)$ for any functions $f, g \in U$. L is clearly symmetric since κ is. If α is a non-zero scalar,

$$\begin{aligned}
l(\alpha f, g) &= \sum_{i \in \operatorname{supp}(\alpha f)} \sum_{j \in \operatorname{supp}(g)} (\alpha f)(i) g(j) \kappa(i,j) \\
&= \sum_{i \in \operatorname{supp}(f)} \sum_{j \in \operatorname{supp}(g)} \alpha f(i) g(j) \kappa(i,j) \\
&= \alpha \sum_{i \in \operatorname{supp}(f)} \sum_{j \in \operatorname{supp}(g)} \alpha f(i) g(j) \kappa(i,j) \\
&= \alpha l(f,g)
\end{aligned}$$

Now consider the situation when $\alpha = 0$. When $\alpha = 0$, $\operatorname{supp}(\alpha f) = \{\}$ and the sum defining $l(\alpha f, g)$ is empty, hence zero by definition. Since $\alpha l(f,g) = 0$ as well, $l(\alpha f, g) = \alpha l(f,g)$ when $\alpha = 0$. So for any scalar α, zero or non-zero, $l(\alpha f, g) = \alpha l(f,g)$.

If $f, g \in U$, then let $h \in U$ and consider $\mathrm{supp}(f+h)$. It can be seen readily that $\mathrm{supp}(f+h) = (\mathrm{supp}(f) \cup \mathrm{supp}(h)) \setminus \{x \mid x \in \mathrm{supp}(f) \cap \mathrm{supp}(h) \wedge f(x) + h(x) = 0\}$. This includes all of $\mathrm{supp}(f)$, except that part of $\mathrm{supp}(f) \cap \mathrm{supp}(h)$ such that $f(x)+h(x) = 0$ for any x in that part. So $\mathrm{supp}(f) \subseteq \mathrm{supp}(f+h) \cup \{x \mid x \in \mathrm{supp}(f) \cap \mathrm{supp}(h) \wedge f(x)+h(x) = 0\}$ and a similar statement holds for $\mathrm{supp}(h)$. Let $Z = \{x \mid x \in \mathrm{supp}(f) \cap \mathrm{supp}(h) \wedge f(x)+h(x) = 0\}$.

$$
\begin{aligned}
&l(f+h, g)\\
&\quad = \sum_{i \in \mathrm{supp}(f+h)} \sum_{j \in \mathrm{supp}(g)} (f+h)(i) g(j) \kappa(i,j)\\
&\quad = \sum_{i \in \mathrm{supp}(f+h)} \sum_{j \in \mathrm{supp}(g)} (f(i)+h(i)) g(j) \kappa(i,j)\\
&\quad = \sum_{i \in \mathrm{supp}(f+h) \cup Z} \sum_{j \in \mathrm{supp}(g)} (f(i)+h(i)) g(j) \kappa(i,j)\\
&\quad = \left(\sum_{i \in \mathrm{supp}(f+h) \cup Z} \sum_{j \in \mathrm{supp}(g)} f(i) g(j) \kappa(i,j) \right) + \left(\sum_{i \in \mathrm{supp}(f+h) \cup Z} \sum_{j \in \mathrm{supp}(g)} h(i) g(j) \kappa(i,j) \right)\\
&\quad = \left(\sum_{i \in \mathrm{supp}(f)} \sum_{j \in \mathrm{supp}(g)} f(i) g(j) \kappa(i,j) \right) + \left(\sum_{i \in \mathrm{supp}(h)} \sum_{j \in \mathrm{supp}(g)} h(i) g(j) \kappa(i,j) \right)\\
&\quad = l(f,g) + l(h,g)
\end{aligned}
$$

so l is linear in its first argument. By symmetry, therefore, l is linear in its second argument as well, hence l is a bilinear function. Since κ is positive-definite, $l(x,x) \geq 0$ for any $x \in U$. U is generally not a Hilbert space, nor is l an inner product as specified, because there is potentially a non-zero element x of U such that $l(x,x) = 0$. However, by taking an appropriate quotient, and forming its completion, we can recover V. But first we show that if $l(x,x) = 0$ for $x \in U$, for any $y \in U$, $l(y,x) = 0$. To see this, consider $l(x+ay, x+ay)$. As shown earlier, this must be non-negative. So we can, by expanding and using bilinearity, show that $l(x,x) + 2al(x,y) + a^2 l(y,y) \geq 0$. Since $l(x,x) = 0$ by hypothesis, we have that $2al(x,y) + a^2 l(y,y) \geq 0$. If $l(y,y) = 0$, then consider $a = -l(x,y)$. We then have $-2l(x,y)^2 \geq 0$, implying that $l(x,y) = 0$. Otherwise, if $l(y,y) \neq 0$, consider $a = \frac{-l(x,y)}{l(y,y)}$: $2\frac{-l(x,y)}{l(y,y)} l(x,y) + \left(\frac{-l(x,y)}{l(y,y)}\right)^2 l(y,y) = \frac{-2l(x,y)^2}{l(y,y)} + \frac{(-l(x,y))^2}{l(y,y)} = -\frac{-l(x,y)}{l(y,y)} \geq 0$. Since $l(y,y) > 0$, it follows that $l(x, y)$ must be 0.

So now consider the set $Z = \{x \mid x \in U \wedge l(x, x) = 0\}$. Since $0 \in Z$, Z is non-empty. If $x, y \in Z$, then for any scalar a, $l(x+ay, x+ay) = l(x, x)+2al(x, y)+a^2l(y, y) = 2al(x, y)$. By the previous paragraph, $l(x, y) = 0$, so $l(x + ay, x + ay) = 2al(x, y) = 0$. This suffices to prove that Z is a subspace of U. Let $W = U/Z$, the quotient of U by Z. For sake of convenience, let $\pi_Z \colon U \to U/Z$ be the standard quotient map. Let $L \colon W \times W \to \mathbf{R}$ be given by $L(x, y) = l(p, q)$ for any $x, y \in W$ and for any p in the coset x and q in the coset y. This is well-defined: if p, p' are each in the coset x, then $p' - p \in Z$ and $p' = p + z_p$ where $z_p \in Z$ and if q, q' are each in the coset y, $q' - q \in Z$ and $q' = q + z_q$ where $z_q \in Z$. So $l(p', q') = l(p + z_p, q + z_q) = l(p, q) + l(p, z_q) + l(z_p, q) + l(z_p, z_q)$. Since $l(y, z) = 0$ for any $y \in U$ and $z \in Z$, the last three terms of this expression are 0, so $l(p', q') = l(p, q)$. Since x and y were arbitrary elements of W, this shows that L is well-defined. Since l was bilinear, L is bilinear and, since l is positive semi-definite, L is (at least) positive semi-definite. If $L(a, a) = 0$, then for any element a' in the coset a, $l(a', a') = 0$. So $a' \in Z$. Every element in W is of the form $x + Z$ for some element $x \in U$, so $a = a' + Z = Z$ and $a = 0$. Since $L(a, a) = 0$ implies $a = 0$, and since $L(a, a) \geq 0$ as proven before for any $a \in W$, L is a positive-definite bilinear function. So W with the function L can be considered a real inner-product space.

Finally, it is well-known that for any real inner-product space W and inner product L, there is essentially a unique completion of W, which is a Hilbert space in which we can consider W embedded densely and whose inner product restricted to $W \times W$ is exactly L. Call this space V, let $\langle \cdot, \cdot \rangle$ be its inner product, and let $i \colon W \to V$ be the dense embedding. For any x in the feature space $\mathcal{F}$, let $\delta_x \colon \mathcal{F} \to \mathbf{R}$ be given by

$$\delta_x(y) = \begin{cases} 1, & y = x \\ 0, & \text{otherwise} \end{cases}$$

Then $\Phi \colon \mathcal{F} \to V$ is given succinctly by $\Phi(x) = i(\pi_Z(\delta_x))$. To see this, let x and y be arbitrary elements of $\mathcal{F}$. Then

$$\langle \Phi(x), \Phi(y) \rangle = L(\pi_Z(\delta_x), \pi_Z(\delta_y)) = l(\delta_x, \delta_y) =$$

$$\sum_{i \in \operatorname{supp}(\delta_x)} \sum_{j \in \operatorname{supp}(\delta_y)} \delta_x(i)\delta_y(j)\kappa(i,j) = \delta_x(x)\delta_y(y)\kappa(x,y) = \kappa(x,y)$$

Thus, given a feature-space $\mathcal{F}$ and positive-definite function $\kappa \colon \mathcal{F} \to \mathbf{R}$, we can construct a Hilbert space V and a mapping function $\Phi \colon \mathcal{F} \to V$ such that $\operatorname{span} \Phi(\mathcal{F})$ is dense in V and such that for any $x_1, x_2 \in \mathcal{F}$, $\kappa(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$, In this way, we can see that in the construction of support vector machines, selection of a kernel is essentially equivalent to selection of a map Φ from a feature space to a particular Hilbert space.

An aside: Although computing with W (rather than U) seems a technicality, this step is actually quite significant. It is the step where underlying linearities of the feature space, when viewed “through the lens” of κ are exposed. Say, for example, that the underlying feature space $\mathcal{F}$ is a finite-dimensional real vector space and that κ is a symmetric bilinear function. Let us say that $\{e_1, \ldots, e_q\}$ are a basis for $\mathcal{F}$. Consider an arbitrary element $v \in \mathcal{F}$ of the form $\sum_{i=1}^{q} c_i e_i$ where $v \notin \{e_1, \ldots, e_q\}$ and consider two different functions with finite support from $\mathcal{F} \to \mathbf{R}$ given by:

$$v_1(x) = \begin{cases} 1, & x = v \\ 0, & x \neq v \end{cases}$$

and

$$v_2(x) = \begin{cases} c_i, & x = e_i \text{ for some } i \in \{e_1, \ldots, e_q\} \text{ and } c_i \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

With some justice, both may be naïvely thought to represent v in W. (Without the assumption of κ’s bilinearity, however, this equivalence does not hold.) Let us compute $l(v_2 - v_1, v_2 - v_1)$. Note that since $v \notin \{e_1, \ldots, e_q\}$, $\operatorname{supp}(v_1) \cap \operatorname{supp}(v_2) = \emptyset$, so $\operatorname{supp}(v_2 - v_1) =$

$\operatorname{supp}(v_2) \cup \operatorname{supp}(v_1)$. We have:

$$\begin{aligned}
l(v_2 - v_1, v_2 - v_1) &= \sum_{i \in \operatorname{supp}(v_2 - v_1)} \sum_{j \in \operatorname{supp}(v_2 - v_1)} (v_2(i) - v_1(i))(v_2(j) - v_1(j))\kappa(i,j) \\
&= \sum_{i=1}^{q} \sum_{j=1}^{q} (v_2(e_i) - v_1(e_i))(v_2(e_j) - v_1(e_j))\kappa(e_i, e_j) + \\
&\quad \sum_{i \in \{v\}} \sum_{j=1}^{q} (v_2(i) - v_1(i))(v_2(e_j) - v_1(e_j))\kappa(i, e_j) + \\
&\quad \sum_{i=1}^{q} \sum_{j \in \{v\}} (v_2(e_i) - v_1(e_i))(v_2(j) - v_1(j))\kappa(e_i, j) + \\
&\quad \sum_{i \in \{v\}} \sum_{j \in \{v\}} (v_2(i) - v_1(i))(v_2(j) - v_1(j))\kappa(i, j) \\
&= \sum_{i=1}^{q} \sum_{j=1}^{q} (c_i - 0)(c_j - 0)\kappa(e_i, e_j) + \\
&\quad \sum_{i \in \{v\}} \sum_{j=1}^{q} (0 - 1)(c_j - 0)\kappa(i, e_j) + \\
&\quad \sum_{i=1}^{q} \sum_{j \in \{v\}} (c_i - 0)(0 - 1)\kappa(e_i, j) + \\
&\quad \sum_{i \in \{v\}} \sum_{j \in \{v\}} (0 - 1)(0 - 1)\kappa(i, j) \\
&= \sum_{i=1}^{q} \sum_{j=1}^{q} c_i c_j \kappa(e_i, e_j) + \sum_{j=1}^{q} -c_j \kappa(v, e_j) + \sum_{i=1}^{q} -c_i \kappa(e_i, v) + \kappa(v, v)
\end{aligned}$$

By κ's assumed bilinearity, this reduces to:

$$\begin{aligned}
l(v_2 - v_1, v_2 - v_1) &= \sum_{i \in \mathrm{supp}(v_2 - v_1)} \sum_{j \in \mathrm{supp}(v_2 - v_1)} (v_2(i) - v_1(i))(v_2(j) - v_1(j))\kappa(i, j) \\
&= \sum_{i=1}^{q}\sum_{j=1}^{q} c_i c_j \kappa(e_i, e_j) + \sum_{j=1}^{q} -c_j \kappa(v, e_j) + \sum_{i=1}^{q} -c_i \kappa(e_i, v) + \kappa(v, v) \\
&= \kappa(\sum_{i=1}^{q} c_i e_i, \sum_{j=1}^{q} c_j e_j) + \kappa(-v, \sum_{j=1}^{q} c_j e_j) + \kappa(\sum_{i=1}^{q} c_i e_i, -v) + \kappa(-v, -v) \\
&= \kappa(\left(\sum_{i=1}^{q} c_i e_i\right) - v, \sum_{j=1}^{q} c_j e_j) + \kappa(\left(\sum_{i=1}^{q} c_i e_i\right) - v, -v) \\
&= \kappa(0, \sum_{j=1}^{q} c_j e_j) + \kappa(0, -v) \\
&= 0
\end{aligned}$$

Thus, since $v_2 - v_1 \in Z$, $v_2 + Z = v_1 + Z$ and the equivalence class of v_2 and v_1 in W are the same. Since v was an arbitrary vector not in the set of basis vectors, the resulting inner product space is certainly spanned by the set of functions $\delta_{e_i}(x)$ (where $\delta_{e_i}(x) = 1$ when and only when $x = e_i$, 0 otherwise, for any $i \in \{1, \ldots, q\}$). It's easy to see that the requirement of finite-dimensionality for $\mathcal{F}$ isn't truly necessary to prove the conclusion and the resulting inner-product space has no greater dimension than the original space. Letting W be U/Z serves the same sort of purpose here as "modding out" the bilinearity relations from M^2 (where M is some left R-module for some ring R) to form the tensor product of M with itself.

In summary, then: support vector machines are essentially functions designed to separate sets. Their operation is perhaps best understood geometrically through real Hilbert spaces, though their implementation is almost always by means of limits of sums of positive-definite kernel functions from a feature space to $\mathbf{R}$. Construction of this function proceeds by selecting a family of kernel functions, computing a minimum-distance vector between the well-separated convex hulls of the categories in a related Hilbert space, then using this vector to construct a separating function as a sum of linear combinations of kernel functions with one variable free, and one evaluated at a particular element of the feature space as in equation 2. A practical method of finding such a vector is discussed in section 3.

CHAPTER 2

A USEFUL KERNEL

2.1. Introduction

Kernels are important in support vector machine operation. Kernels are constructed from positive-definite functions, usually from $\mathbf{R}^n$ for some natural number n. A typical example of a kernel function is $\exp(-\|x\|^2)$, where x varies over $\mathbf{R}^n$ for some fixed positive integer n. Less well known is that fact that $\exp(-\|x\|^\lambda)$ is positive definite if and only if $0 < \lambda \leq 2$. The proof of this fact for $0 < \lambda < 2$ appears to be quite difficult. The first person to prove this was Paul Lévy ([3]) in 1925, who used deep properties of stochastic processes as the foundation of his proof. This was followed by a second proof by Salomon Bochner ([1]) which supposedly is simpler but again used nontrivial properties of stochastic processes. The purpose of this chapter is to give a completely elementary proof that $\exp(-\|x\|^\lambda)$ is positive definite (when $0 < \lambda \leq 2$) using only concepts from undergraduate mathematics.

The least elementary fact from undergraduate matrix algebra that we use is that for each square, self-adjoint, positive semi-definite real- or complex-valued matrix M, there exists a square real- or complex-valued matrix N respectively such that $M = N^*N$, where A^* is defined as the transpose (respectively, conjugate-transpose) of the matrix A. A proof of this can be found in most advanced undergraduate textbooks on linear algebra.

2.2. Lemmata and Proofs

First, some lemmata:

Lemma 2.1. *Let n be a positive integer and let M and N be real (respectively, complex) $n \times n$ matrices. Assume further that M and N are self-adjoint and positive semi-definite. Then their component-wise product is also self-adjoint and positive semi-definite.*

Proof. Let P denote the indicated component-wise product. For any integers i and j such that $1 \leq i, j \leq n$, $P_{i,j} = M_{i,j}N_{i,j}$. So $P_{j,i} = M_{j,i}N_{j,i} = \overline{M_{i,j}}\,\overline{N_{i,j}} = \overline{M_{i,j}N_{i,j}} = \overline{P_{i,j}}$. Thus, P is self-adjoint.

Let $M = A^*A$ and $N = B^*B$ where A and B are real-valued (respectively, complex-valued) $n \times n$ matrices, which matrices A and B we can always find. Also let x be an arbitrary element of $\mathbf{R}^n$ (respectively $\mathbf{C}^n$). We can assume x is represented as column-vector (an $n \times 1$ matrix). Consider the product x^*Px.

$$
\begin{aligned}
x^*Px = (x^*P)x &= \sum_{j=1}^{n}(\sum_{i=1}^{n}(x^*)_{1,i}P_{i,j})x_{j,1} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n}\overline{x_{i,1}}P_{i,j}x_{j,1} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n}\overline{x_{i,1}}M_{i,j}N_{i,j}x_{j,1} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n}\overline{x_{i,1}}(\sum_{k=1}^{n}(A^*)_{i,k}A_{k,j})(\sum_{l=1}^{n}(B^*)_{i,l}B_{l,j})x_{j,1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\overline{x_{i,1}}(A^*)_{i,k}A_{k,j}(B^*)_{i,l}B_{l,j}x_{j,1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\overline{x_{i,1}}\overline{A_{k,i}}A_{k,j}\overline{B_{l,i}}B_{l,j}x_{j,1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\overline{x_{i,1}}\overline{A_{k,i}B_{l,i}}x_{j,1}A_{k,j}B_{l,j} \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\overline{x_{i,1}}\overline{A_{k,i}B_{l,i}}x_{j,1}A_{k,j}B_{l,j} \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n}(\sum_{i=1}^{n}\overline{x_{i,1}}\overline{A_{k,i}B_{l,i}})(\sum_{j=1}^{n}x_{j,1}A_{k,j}B_{l,j}) \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n}\overline{(\sum_{i=1}^{n}x_{i,1}A_{k,i}B_{l,i})}(\sum_{i=1}^{n}x_{i,1}A_{k,i}B_{l,i}) \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n}\left|\sum_{i=1}^{n}x_{i,1}A_{k,i}B_{l,i}\right|^2 \\
&\geq 0
\end{aligned}
\tag{3}
$$

Since x was arbitrary, P is positive semi-definite, as required. □

Lemma 2.2. *Let n be a positive integer and M a self-adjoint positive semi-definite real $n \times n$ matrix. Let $\{a_k\}_{k=0}^{\infty}$ be a sequence of non-negative numbers. Then the matrix whose i, jth entry for integers $1 \leq i, j \leq n$ is $\sum_{k=0}^{\infty} a_k M_{i,j}^k$ is self-adjoint and positive semi-definite, provided that each such series converges.*

Proof. The $n \times n$ matrix U whose entries are all 1s is certainly self-adjoint viewed as either a real or complex matrix. For any $x \in \mathbf{R}^n$ or $x \in \mathbf{C}^n$, where x is represented as an $n \times 1$ matrix,

$$\begin{aligned}
x^*Ux = (x^*U)x &= \sum_{j=1}^{n}(\sum_{i=1}^{n}(x^*)_{1,i}U_{i,j})x_{j,1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}(x^*)_{1,i}U_{i,j}x_{j,1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\overline{x_{i,1}}\,1\,x_{j,1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\overline{x_{i,1}}x_{j,1} \qquad (4)\\
&= (\sum_{i=1}^{n}\overline{x_{i,1}})(\sum_{j=1}^{n}x_{j,1}) \\
&= \left|\sum_{i=1}^{n}x_{i,1}\right|^2 \\
\geq 0
\end{aligned}$$

So U is clearly positive semi-definite as a real or complex matrix. For any integer $l \geq 0$, define the i, jth element of P_l as $M_{i,j}^l$ for integers $1 \leq i, j \leq n$. P_0 is self-adjoint and positive semi-definite as just demonstrated. If P_i is self-adjoint and positive semi-definite for some integer $i \geq 0$, P_{i+1} being the component-wise product of P_i and M is self-adjoint and positive semi-definite by lemma 2.1. So by mathematical induction, P_l is self-adjoint and positive semi-definite for all integers $l \geq 0$. If c is a real non-negative number and A is a self-adjoint positive semi-definite $n \times n$ matrix, then for any $n \times 1$ matrix x, $x^*Ax \geq 0$, so $c(x^*Ax) = x^*(cA)x \geq 0$ as well, so cA is positive semi-definite and clearly self-adjoint. If A and B are $n \times n$ positive semi-definite self-adjoint matrices, then for any $n \times 1$ matrix x, $x^*(A+B)x = (x^*(A+B))x = (x^*A + x^*B)x = x^*Ax + x^*Bx \geq 0$, so $A+B$ is positive semi-definite (and clearly self-adjoint).

Thus, $\sum_{i=0}^{n} a_i P_i$ is self-adjoint and positive semi-definite.

If $\{A_k\}_{k=1}^{\infty}$ is a convergent sequence of self-adjoint positive semi-definite matrices, $A = \lim_{k\to\infty} A_k$ is clearly self-adjoint. To see that it is positive semi-definite, note that for a fixed $n \times 1$ matrix x, $f_x(X) = x^*Xx$ defines a continuous function from $n \times n$ matrices to the real numbers under any reasonable metric on $n \times n$ matrices. So $f_x(A) = \lim_{k\to\infty} f_x(A_k)$. Since $f_x(A_k) \geq 0$ for all k, it follows that $\lim_{k\to\infty} f_x(A_k)$ must also be non-negative. Since this is true for arbitrary x, it follows that A is positive semi-definite.

Thus, $\lim_{n\to\infty} \sum_{i=0}^{n} a_i P_i$, when it exists, must be a self-adjoint, positive semi-definite $n \times n$ matrix. □

2.3. Main Proof

The proof here continues in a series of theorems leading to the final result. While the earlier lemmata are sufficiently general as to warrant separate treatment, the following theorems are merely scaffolding for the final result.

Theorem 2.3. *Let $x_1, \ldots, x_n$ be a sequence of n linearly independent vectors in a complex inner product space V with inner product $\langle \cdot, \cdot \rangle$. There exists a positive real number K such that for all real $k > K$, the matrix M whose i,jth entry is given by*

$$M_{i,j} = 1 - \left\| \frac{x_i - x_j}{k} \right\|^2 = 1 - \left\langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \right\rangle$$

for integers $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite.

Proof. M is clearly a real symmetric matrix, hence self-adjoint. Since any vector β can be written $\|\beta\| \alpha$ where $\alpha = \frac{\beta}{\|\beta\|}$ if $\beta \neq 0$ and where α is any vector on the unit sphere when $\beta = 0$, to show that for any vector β, $\beta^* M \beta = \|\beta\|^2 (\alpha^* M \alpha) \geq 0$, all that remains to be proven is that for any vector α on the unit sphere, $\alpha^* M \alpha \geq 0$.

For any $\alpha \in \mathbf{C}^n$, with components $\alpha_1, \ldots, \alpha_n$

$$\begin{aligned}
\alpha^* M\alpha &= \sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i} M_{i,j}\alpha_j \\
&= \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i}\alpha_j\right) - \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i}\alpha_j \left\langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k}\right\rangle\right) \\
&= \left(\sum_{i=1}^{n} \overline{\alpha_i}\right)\left(\sum_{j=1}^{n} \alpha_j\right) - (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i}\alpha_j(\langle x_i, x_i\rangle + \langle x_j, x_j\rangle - \langle x_i, x_j\rangle - \langle x_j, x_i\rangle)\right) \\
&= \overline{\left(\sum_{i=1}^{n} \alpha_i\right)}\left(\sum_{j=1}^{n} \alpha_j\right) - (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i}\alpha_j \langle x_i, x_i\rangle\right) - (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i}\alpha_j \langle x_j, x_j\rangle\right) \\
&\quad + (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i}\alpha_j \langle x_i, x_j\rangle\right) + (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i}\alpha_j \langle x_j, x_i\rangle\right) \\
&= \overline{\left(\sum_{i=1}^{n} \alpha_i\right)}\left(\sum_{j=1}^{n} \alpha_j\right) - (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \langle x_i, \alpha_i x_i\rangle\right) - (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\alpha_i} \langle \alpha_j x_j, x_j\rangle\right) \\
&\quad + (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \langle \overline{\alpha_i} x_i, \overline{\alpha_j} x_j\rangle\right) + (1/k^2)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \langle \alpha_j x_j, \alpha_i x_i\rangle\right) \\
&= \overline{\left(\sum_{i=1}^{n} \alpha_i\right)}\left(\sum_{j=1}^{n} \alpha_j\right) - (1/k^2)\left(\sum_{j=1}^{n} \alpha_j\right)\left(\sum_{i=1}^{n} \langle x_i, \alpha_i x_i\rangle\right) \\
&\quad - (1/k^2)\left(\sum_{i=1}^{n} \overline{\alpha_i}\right)\left(\sum_{j=1}^{n} \langle \alpha_j x_j, x_j\rangle\right) \\
&\quad + (1/k^2)\left(\left\langle \sum_{i=1}^{n} \overline{\alpha_i} x_i, \sum_{j=1}^{n} \overline{\alpha_j} x_j\right\rangle + \left\langle \sum_{j=1}^{n} \alpha_j x_j, \sum_{i=1}^{n} \alpha_i x_i\right\rangle\right) \\
&= \overline{\left(\sum_{i=1}^{n} \alpha_i\right)}\left(\sum_{j=1}^{n} \alpha_j\right) - (1/k^2)\left(\sum_{j=1}^{n} \alpha_j\right)\left(\sum_{i=1}^{n} \langle x_i, \alpha_i x_i\rangle\right) \\
&\quad - (1/k^2)\left(\sum_{i=1}^{n} \overline{\alpha_i}\right)\left(\sum_{j=1}^{n} \langle \alpha_j x_j, x_j\rangle\right) \\
&\quad + (1/k^2)\left(\left\langle \sum_{i=1}^{n} \overline{\alpha_i} x_i, \sum_{j=1}^{n} \overline{\alpha_j} x_j\right\rangle + \left\langle \sum_{j=1}^{n} \alpha_j x_j, \sum_{i=1}^{n} \alpha_i x_i\right\rangle\right)
\end{aligned}$$

$$
\begin{aligned}
&= \overline{\left(\sum_{i=1}^{n} \alpha_i\right)} \left(\sum_{j=1}^{n} \alpha_j\right) - (1/k^2) \left(\sum_{j=1}^{n} \alpha_j\right) \overline{\left(\sum_{i=1}^{n} \langle \alpha_i x_i, x_i \rangle\right)} \\
&\quad - (1/k^2) \overline{\left(\sum_{i=1}^{n} \alpha_i\right)} \left(\sum_{j=1}^{n} \langle \alpha_j x_j, x_j \rangle\right) \\
&\quad + (1/k^2) \left(\left\| \sum_{i=1}^{n} \overline{\alpha_i} x_i \right\|^2 + \left\| \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right) \\
&= \overline{\left(\left(\sum_{i=1}^{n} \alpha_i\right) - (1/k^2) \left(\sum_{i=1}^{n} \langle \alpha_i x_i, x_i \rangle\right) \right)} \left(\left(\sum_{j=1}^{n} \alpha_j\right) - (1/k^2) \left(\sum_{j=1}^{n} \langle \alpha_j x_j, x_j \rangle\right) \right) \\
&\quad + (1/k^2) \left(\left\| \sum_{i=1}^{n} \overline{\alpha_i} x_i \right\|^2 + \left\| \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right) - (1/k^4) \left(\overline{\left(\sum_{i=1}^{n} \langle \alpha_i x_i, x_i \rangle\right)} \left(\sum_{j=1}^{n} \langle \alpha_j x_j, x_j \rangle\right) \right) \\
&= \left| \left(\sum_{i=1}^{n} \alpha_i\right) - (1/k^2) \left(\sum_{i=1}^{n} \langle \alpha_i x_i, x_i \rangle\right) \right|^2 + (1/k^2) \left(\left\| \sum_{i=1}^{n} \overline{\alpha_i} x_i \right\|^2 + \left\| \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right) \\
&\quad - (1/k^4) \left| \sum_{i=1}^{n} \langle \alpha_i x_i, x_i \rangle \right|^2
\end{aligned}
$$

So

$$
\alpha^* M \alpha \geq (1/k^2) \left(\left\| \sum_{i=1}^{n} \overline{\alpha_i} x_i \right\|^2 + \left\| \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right) - (1/k^4) \left| \sum_{i=1}^{n} \langle \alpha_i x_i, x_i \rangle \right|^2
$$

Let us now restrict α to the unit sphere. Since the unit sphere (in $\mathbf{C}^n$) is compact and $\langle \cdot \rangle$ is continuous on V, it follows that $\|\sum_{i=1}^{n} \overline{\alpha_i} x_i\|^2 + \|\sum_{i=1}^{n} \alpha_i x_i\|^2$ attains a minimum value of no less than 0. This value cannot be 0, since the zero vector is not on the unit sphere and $x_1, \ldots, x_n$ are linearly independent, so this minimum is strictly greater than 0. Call it $2B$. $|\sum_{i=1}^{n} \langle \alpha_i x_i, x_i \rangle|^2$ also attains a non-negative maximum for α in the unit sphere. Call this maximum U. Then for any α on the unit sphere:

$$
\alpha^* M \alpha \geq (1/k^4)(2Bk^2 - U)
$$

Let $K = \sqrt{\frac{U}{2B}}$ and note that for any $k > K$, $\alpha^* M \alpha$ is non-negative, proving the theorem. □

THEOREM 2.4. *Let $x_1, \ldots, x_n$ be a sequence of n linearly independent vectors in a complex inner product space V and let $\mu \in [0,1]$. There exists a non-negative real number K such that for all real $k > K$, the matrix M whose i, jth entry is given by*

$$M_{i,j} = 1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^{\mu} \tag{5}$$

for integers $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite.

PROOF. By the binomial theorem for non-natural-number exponents, we have for $\mu \in (0,1)$:

$$(1+y)^{\mu} = 1 + \sum_{k=1}^{\infty} \left(\prod_{j=0}^{k-1} (\mu - j) \right) (y^k / k!) \tag{6}$$

for any $y \in [-1,1]$, since $\mu > 0$. By inspection, when $\mu = 0$ or $\mu = 1$, the series becomes finite and equal to $(1+y)^{\mu}$ trivially. Let $x = y + 1$. Then for any $x \in [0,2]$, we have:

$$x^{\mu} = 1 + \sum_{k=1}^{\infty} \left(\prod_{j=0}^{k-1} (\mu - j) \right) ((x-1)^k / k!) \tag{7}$$

so

$$x^{\mu} = 1 + \mu \left(\sum_{k=1}^{\infty} \left(\prod_{j=1}^{k-1} (\mu - j) \right) ((x-1)^k / k!) \right) \tag{8}$$

We may rewrite this as:

$$
\begin{aligned}
&x^{\mu} \\
&= 1 + \mu\left(\sum_{k=1}^{\infty}\left(\prod_{j=1}^{k-1}(j-\mu)\right)(-1)^{k-1}((x-1)^k/k!)\right) \\
&= 1 + \mu\left(\sum_{k=1}^{\infty}\left(\prod_{j=1}^{k-1}(j-\mu)\right)(-1)^{k-1}(-1)^k((1-x)^k/k!)\right) \\
&= 1 - \mu\left(\sum_{k=1}^{\infty}\left(\prod_{j=1}^{k-1}(j-\mu)\right)((1-x)^k/k!)\right) \\
&= 1 - \left(\sum_{k=1}^{\infty}\left(\mu\prod_{j=1}^{k-1}(j-\mu)\right)((1-x)^k/k!)\right)
\end{aligned} \tag{9}
$$

and, therefore,

$$
1 - x^{\mu} = \sum_{l=1}^{\infty}\left(\mu\prod_{j=1}^{l-1}(j-\mu)\right)((1-x)^l/l!) \tag{10}
$$

Note that since $\mu \leq 1$, $j - \mu \geq 0$ for any positive integer j. So for any positive integer l, $\mu(\prod_{j=1}^{l-1}(j-\mu))/l! \geq 0$. Now for each pair of integers $1 \leq i, j \leq n$ in turn, let $x = \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle$. The left-hand side of the expansion above is just the i, jth entry of the matrix M defined in the statement of the theorem. The right-hand side is a power-series expansion with non-negative coefficients in $1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle$. So for positive integers k, let $W(k)$ be the matrix whose i, jth entry is $1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle$ for integers $1 \leq i, j \leq n$. Then by lemma 2.2, as long as the series converges for each pair i, j of integers, and as long as $W(k)$ is self-adjoint and positive semi-definite, M as defined in the statement of the theorem will be self-adjoint and positive semi-definite as well.

By lemma 2.3, there exists $K \geq 0$ such that for any $k > K$, the matrix $W(k)$ is self-adjoint and positive semi-definite. Since $x_1, \ldots, x_n$ is a finite set, $\langle x_i - x_j, x_i - x_j \rangle$ attains a maximum value, necessarily non-negative. Call it L. Let $K' = \max(K, \sqrt{L/2})$. Because $K \geq 0$, $K' \geq 0$ as well. Then for any $k > K'$, $k > K$, so $W(k)$ is self-adjoint and positive semi-definite.

$$k > \sqrt{L/2} \geq 0 \tag{11}$$
$$0 < k^{-2} < 2/L$$

Since k is also greater than $\sqrt{L/2}$, $\langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle \leq 2$. Thus, the indicated power series converges for each entry in the matrix whenever $k > K'$.

For any integer $l \geq 1$, define the $n \times n$ matrix P_l whose i, jth entry is given by $W_{i,j}^l$ for integers $1 \leq i, j \leq n$. $W = P_1$ is self-adjoint and positive semi-definite. Assume that P_i is self-adjoint and positive semi-definite for some integer $i \geq 1$. Then P_{i+1} is the component-wise product of P_i and W. By lemma 2.1, P_{i+1} is self-adjoint and positive semi-definite as well. So P_l is self-adjoint and positive semi-definite for each $l \geq 1$.

Consider $(\mu \prod_{j=1}^{l-1}(j - \mu))(1/l!)$. Since $\mu \leq 1$, this product is non-negative. So the matrix P_l' whose i, jth component is given by

$$(\mu \prod_{j=1}^{l-1}(j - \mu))(1/l!)(1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle)^l \tag{12}$$

for $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite as well, as is the matrix S_l given by $\sum_{k=1}^{l} P_k'$. Recall that $\langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle \leq 2$ and certainly it is no less than 0. So the matrix whose i, jth entry is given by

$$\sum_{k=1}^{\infty} \left(\mu \prod_{j=1}^{k-1}(j - \mu) \right) ((1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle)^k / k!) \tag{13}$$

for integers $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite as well, since the limit as l goes to infinity of the i, jth element of S_l converges by equation 10 and this matrix is this component-wise limit. Also by equation 10, this sum is just $1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^\mu$. So for $k > K'$, the matrix whose i, jth element is given by

$$1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^\mu \tag{14}$$

for integers $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite, with $K' \geq 0$, as desired. □

Theorem 2.5. *Let $x_1, \ldots, x_n$ be a sequence of n linearly independent vectors in a complex inner product space V and let $\mu \in [0,1]$. There exists a non-negative real number K such that for all real $k > K$, the matrix M whose i, jth entry is given by*

$$M_{i,j} = e^{1-\langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^{\mu}} \tag{15}$$

for integers $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite.

Proof. The $n \times n$ identity matrix is self-adjoint and positive semi-definite trivially. Let M be the matrix whose i, jth entry for integers $1 \leq i, j \leq n$ is given by $1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^{\mu}$ The power-series expansion of e^x converges everywhere on the real line and each coefficient is non-negative. So when K is that given by theorem 2.4 and $k > K$, since the matrix whose i, jth element is $1 - \langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^{\mu}$ for $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite, the indicated matrix must be self-adjoint and positive semi-definite as well. □

Theorem 2.6. *Let $x_1, \ldots, x_n$ be a sequence of n linearly independent vectors in a complex inner product space V. Let $\mu \in [0,1]$. Then the matrix M whose i, jth entry is given by*

$$M_{i,j} = e^{-\langle x_i - x_j, x_i - x_j \rangle^{\mu}} \tag{16}$$

for $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite.

Proof. By the corollary, there exists K such that for $k > K$, the matrix N whose i, jth component is given by $e^{1-\langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^{\mu}}$ for $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite. Consider the matrix $\frac{1}{e}N$. This matrix must also be self-adjoint and positive semi-definite since $\frac{1}{e}$ is a positive number and is exactly equal to the matrix M whose i, jth entry is $e^{-\langle \frac{x_i - x_j}{k}, \frac{x_i - x_j}{k} \rangle^{\mu}}$ for integers $1 \leq i, j \leq n$.

Now let n be an integer greater than K^μ. Since $K \geq 0$, $n \geq 1$. Since $n > K^\mu$, $n^{1/\mu} > K$. So $-(1/n^2)\langle x_i - x_j, x_i - x_j\rangle^\mu = -((1/n^{2/\mu})\langle x_i - x_j, x_i - x_j\rangle)^\mu = -\langle \frac{x_i - x_j}{n^{1/\mu}}, \frac{x_i - x_j}{n^{1/\mu}}\rangle^\mu$. So the matrix Q whose i, jth entry is given by

$$e^{-(1/n^2)\langle x_i - x_j, x_i - x_j\rangle^\mu} \tag{17}$$

for integers $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite. Since this matrix is self-adjoint and positive semi-definite, it follows that the matrix whose i, jth entry is given by

$$(e^{-(1/n^2)\langle x_i - x_j, x_i - x_j\rangle^\mu})^{n^2} \tag{18}$$

for integers $1 \leq i, j \leq n$ is self-adjoint and positive semi-definite as well, being the n^2 repeated element-by-element product of Q with itself, by lemma 2.1. But this is just the matrix M whose i, jth component is given by

$$e^{-\langle x_i - x_j, x_i - x_j\rangle^\mu} \tag{19}$$

for integers $1 \leq i, j \leq n$. Hence M is self-adjoint and positive semi-definite, as desired. □

Theorem 2.7. *Let ϵ be any positive number. Let $x_1, \ldots, x_n$ be a sequence of n vectors in $\mathbf{C}^n$. Then there exists a sequence of vectors $y_1, \ldots, y_n$ such that $\|y_i - x_i\| < \epsilon$ for $1 \leq i \leq n$ and such that $y_1, \ldots, y_n$ is a linearly independent set.*

Proof. Let $I = \{1, \ldots, n\}$ and for any $J \subseteq I$, $V_J = \{x_k \mid k \in J\}$. Note that $\text{span}(V_I) = \text{span}(V_I)$, and I is finite, so there must be a set $J \subseteq I$ with minimal cardinality such that $\text{span}(V_J) = \text{span}(V_I)$. Since J has minimal cardinality, V_J must be a linearly independent set of vectors — if not, then $\sum_{k \in J} c_k x_k = 0$ where for each $k \in J$, c_k is a real number and at least one such $c_k \neq 0$, and one can express that x_k as a linear combination of the remaining elements of J, hence the span of V_J is the same as the span of $V_{J\setminus\{k\}}$, a contradiction of the minimal cardinality of J. Since V_J is a linearly independent subset of $\mathbf{C}^n$, there exists a set of vectors C of $\mathbf{C}^n$ of size $n - |V_J|$ such that $V_J \cup C$ forms a basis for $\mathbf{C}^n$. Without loss of generality, we may assume that each element of C has norm less than ϵ. Since $|C| = n - |V_J| = |I \setminus J|$, there exists a bijection d from $I \setminus J$ to C. Define the sequence $(y_1, \ldots, y_n)$ by setting

$$y_k = \begin{cases} x_k & \text{if } k \in J \\ d_k + x_k & \text{if } k \in I \setminus J \end{cases} \tag{20}$$

for all $k \in I$. This sequence of vectors is linearly independent, for if $0 = \sum_{k\in I} c_k y_k$ for some sequence of scalars c, then

$$0 = \sum_{k\in J} c_k x_k + \sum_{k\in I\setminus J} c_k(d_k + x_k) = \sum_{k\in I} c_k x_k + \sum_{k\in I\setminus J} c_k d_k \tag{21}$$

Since $\sum_{k\in I} c_k x_k \in \text{span}(S)$, it is in the span of V_J as well, so $\sum_{k\in I} c_k x_k = \sum_{k\in J} c'_k x_k$ for some real constants c'_k. Therefore $0 = \sum_{k\in J} c'_k x_k + \sum_{k\in I\setminus J} c_k d_k$. Since $V_J \cup C$ spans $\mathbf{C}^n$, its vectors are linearly independent, hence $c'_k = 0$ for $k \in J$ and $c_k = 0$ for each $k \in I \setminus J$. So

$$0 = \sum_{k\in J} c_k x_k + \sum_{k\in I\setminus J} c_k(d_k + x_k) = \sum_{k\in J} c_k x_k \tag{22}$$

Since V_J is a linearly independent set of vectors, it follows that $c_k = 0$ for all $k \in J$. So if $0 = \sum_{k\in I} c_k y_k$, $c_k = 0$ for all $k \in I$ and the sequence $(y_1, \ldots, y_n)$ is linearly independent. Finally, if $k \in J$, $\|y_k - x_k\| = \|x_k - x_k\| = 0$ and if $k \in I \setminus J$, $\|y_k - x_k\| = \|d_k\| < \epsilon$, as desired. □

Theorem 2.8. *Let $x_1, \ldots, x_n$ be a sequence of arbitrary vectors in a complex inner product space V. Let $\mu \in [0, 1]$. Then the matrix M whose i, jth entry is given by*

$$M_{i,j} = e^{-\langle x_i - x_j, x_i - x_j\rangle^\mu} \tag{23}$$

is positive semi-definite.

PROOF. Let $S = \operatorname{span}(x_1, \ldots, x_n)$. S is a finite-dimensional inner product space of dimension no greater than n. Call this dimension m. Then there is a linear mapping from S to $\mathbf{C}^m$ that preserves inner products (necessarily a bijection). The standard linear embedding of $\mathbf{C}^m$ into $\mathbf{C}^n$ also preserves inner products. So there exists a linear inner-product preserving mapping from S to $\mathbf{C}^n$. Pick an arbitrary such mapping and for each $1 \leq i \leq n$, let x'_i be the vector in $\mathbf{C}^n$ corresponding to x_i in V. Then the matrix M whose i, jth component is given by

$$e^{-\langle x'_i - x'_j, x'_i - x'_j \rangle^{\mu}} \tag{24}$$

is identical to M in every respect.

Assume M is not positive semi-definite. Then there exists $\alpha \in \mathbf{C}^n$ such that $\alpha^* M \alpha < 0$. Let $N(y_1, \ldots, y_n)$ be for each $y_1, \ldots y_n \in \mathbf{C}^n$ the matrix whose i, jth component is given by $e^{-\langle y_i - y_j, y_i - y_j \rangle^{\mu}}$. Consider $\alpha^* N(y_1, \ldots, y_n)\alpha$. This is a continuous function of the arguments $y_1, \ldots, y_n$. So there exists $\delta > 0$ such that for any $(y_1, \ldots, y_n)$ within δ of $(x'_1, \ldots, x'_n)$, $\alpha^* N(y_1, \ldots, y_n)\alpha < 0$ as well. Lemma 2.7 tells us that there exists a linearly independent set of vectors $y_1, \ldots, y_n$ satisfying this condition. But by theorem 2.6, the linear independence of the sequence $(y_1, \ldots, y_n)$ guarantees that $N(y_1, \ldots, y_n)$ is positive semi-definite, a contradiction of $\alpha^* N(y_1, \ldots, y_n)\alpha < 0$. So M must be positive semi-definite. □

Our main result for this chapter is Theorem 2.9, below:

THEOREM 2.9. *The function $f(x,y) = e^{-\alpha\|x-y\|^\lambda}$ is a kernel on complex inner product spaces V when $0 \leq \lambda \leq 2$, $\alpha > 0$, and $\|\cdot\|$ indicates the norm derived from the inner product.*

PROOF. $e^{-\alpha\|x-y\|^\lambda} = e^{-\left\|\alpha^{1/\lambda}(x-y)\right\|^\lambda} = e^{-\left\|(\alpha^{1/\lambda}x)-(\alpha^{1/\lambda}y)\right\|^\lambda}$ for any $x, y \in V$, so if $g(x,y) = e^{-\|x-y\|^\lambda}$ is a kernel on V, f must be as well. $g(x,x) = 1$ for all $x \in V$. $e^{-\|x-y\|^\lambda} = e^{-\langle x-y,x-y\rangle^{(\lambda/2)}}$ where $\lambda/2 \in [0,1]$. So for any sequence $x_1, \ldots, x_n$ of vectors in V, the matrix whose i,jth entry is given by $g(x_i, x_j)$ is positive semi-definite by theorem 2.8. So g is a kernel on V, as must be f, QED. □

CHAPTER 3

MAIN ALGORITHM

3.1. Introduction

In a practical construction of a particular support vector machine, the data we use to classify each individual are generally encoded as a vector of real numbers in some fashion, as shown in example 1.4. $\mathbf{Z}$ and $\mathbf{Q}$ in that particular example are identified with their embeddings in $\mathbf{R}$ and each individual is assigned a vector from $\mathbf{R}^5$. In the general case, the feature space is taken to be a subset (perhaps not explicitly defined and usually proper) of $\mathbf{R}^\Xi$ for some natural number Ξ, and each individual is associated with a vector from this space. From each of the disjoint classes, we take a finite set of representatives and use these as the members of C_1 and C_2 (using the language of Chapter 1). (Since the goal of a support vector machine is to generate a test for membership in C_1 or C_2, we don't usually have an explicit description of the categories C_1 and C_2. If such a description were available, it could much more directly be converted into a membership testing function, obviating the construction of the support vector machine!) Additionally, for any given execution of the support vector machine, we restrict our space $\mathcal{F}$ to include only C_1, C_2, and the (finite!) set of individuals to be tested for membership by the support vector machine. Thus, $\mathcal{F}$ is a merely finite set. Finally, we select a kernel of some variety based on our needs and theoretical appreciation of which kernel leads to an appropriate separating function.

Since we start with a feature space and a kernel, we conceptually use the approach outlined in Section 1.6 on kernel functions. Let the vector spaces V, U, W, Z be as defined in that discussion, and let Φ and $\mathcal{F}$ be as in that section also. We use the approach of that section to create the Hilbert space V in which we will work and the mapping Φ from the feature space to V. The restriction of $\mathcal{F}$ to being a finite set has useful consequences for the construction of our support vector machine. Among these are that the space U from that section, the space of real-valued functions on $\mathcal{F}$ with finite support, is spanned by a finite set: the set $\{f_x \mid x \in \mathcal{F} \wedge f_x \colon \mathcal{F} \to \mathbf{R} \wedge f_x(v) = \begin{cases} 1, & v = x \\ 0, & v \neq x \end{cases}$ for each $v \in \mathcal{F}\}$. So U is finite-dimensional, implying that the quotient $W = U/Z$ from that section is a finite-dimensional real vector space, and also implying that the function $l \colon U \times U \to \mathbf{R}$ defined in that section is a continuous function from U^2 (under any reasonable metric making U homeomorphic to $\mathbf{R}^n$ for some natural number n) to $\mathbf{R}$. Thus, Z is closed and therefore the quotient space $W = U/Z$ is already a complete space. So the Hilbert space V, with the norm derived from the inner product based on the function L defined in that section, is isomorphic to W, hence may be identified with $W = U/Z$. Since V is a real finite-dimensional inner product space, we may fruitfully regard it as isometrically isomorphic to $\mathbf{R}^\sigma$ for some natural number σ.

For the balance of this chapter, then, let V be a real vector space of finite dimension constructed as in the last paragraph, let m and n be positive integers, and let $V_X = \{x_i\}_{i=1}^m$ and $V_Y = \{y_i\}_{i=1}^n$ be two finite sequences of distinct vectors from V. (We identify V_x as the image of C_1 under Φ and V_y as the image of C_2 under Φ.) Also let X and Y be the convex hulls of V_X and V_Y respectively and assume that $X \cap Y = \emptyset$. Let D be the convex hull of the set $\{x_i - y_j \mid i \in Ix \wedge j \in Iy\}$ and recall that D may also be characterized by lemma 1.2 as $\{x - y \mid x \in X \wedge y \in Y\}$. (Incidentally, V's dimension may change based on (among other things) the element under test from the feature space.)

The finiteness of V_X and V_Y renders X and Y compact in the following way: for any non-negative integer q, let the set $\Delta^q = \{(c_0, c_1, \ldots, c_q) \mid (c_0, c_1, \ldots, c_q) \in [0,1]^{q+1} \wedge (\forall i)(i \in \{0, \ldots, q\} \Rightarrow c_i \geq 0) \wedge \sum_{i=0}^q c_i = 1\}$. Δ^q is a closed and bounded subset of $\mathbf{R}^q$ for any q, hence is compact. The functions $\varphi_X \colon \Delta^{m-1} \to X$ given by $\varphi_X((c_0, \ldots, c_{m-1})) = \sum_{i=0}^{m-1} c_i x_{i+1}$ and $\varphi_Y \colon \Delta^{n-1} \to Y$ given by $\varphi_Y((c_0, \ldots, c_{n-1})) = \sum_{i=0}^{n-1} c_i y_{i+1}$ are clearly continuous, as they are the composition of the continuous functions of scalar multiplication and vector addition. The compactness of Δ^{m-1} and Δ^{n-1} and the continuity of φ_X and φ_Y render the images $\varphi_X(\Delta^{m-1})$ and $\varphi_Y(\Delta^{n-1})$ both compact. $\varphi_X(\Delta^{m-1})$ is clearly the convex hull of the finite sequence V_X and $\varphi_Y(\Delta^{n-1})$ is just as clearly the convex hull of V_Y. So X and Y are compact. Since X and Y are compact, their product is compact. Since $s \colon X \times Y \to V$ given by $s(x, y) = x - y$ for any $(x, y) \in X \times Y$ is continuous. So the image $s(X \times Y)$ is compact as well. $s(X \times Y)$ is also equal to D, defined earlier. Thus, X, Y, and D are compact sets. Let φ_X, φ_Y, and Δ^q for non-negative integers q retain their definitions from this section for the rest of the chapter.

Since $\|\cdot\|$ restricted to D is a continuous function from D to $\mathbf{R}$, and D is compact, it attains a minimum value at some point $v \in D$. Since $X \cap Y = \emptyset$, $0 \notin D$, so $\|v\| \neq 0$. Hence, X and Y have separation strictly greater than 0. So by theorem 1.7, X and Y can be well-separated by a hyperplane. By lemma 1.2, D is convex and by lemma 1.1, every minimizing sequence for $\|\cdot\|$ from D converges to this element v of D where $\|\cdot\|$ attains its minimum value. By theorem 1.7, this value for v in effect defines a separating hyperplane. The separating function corresponding to this vector v is expressed by equation 1. In this expression, $f(x) = \langle \Phi(x), v \rangle = \lim_{n \to \infty} \langle \Phi(x), \sum_{i=1}^{q_n} \alpha_{n,i}(w_{n,i} - c_{n,i}) \rangle$, where x is an arbitrary element from the feature space, recall that the reason that the limit was taken is that, in the general case, v is merely the limit of a sequence of elements from D — not necessarily an element of D itself. Since we know here that v is a member of D, and since D is the convex hull of $\{w - c \mid w \in V_x \wedge c \in V_y\}$, we can write v directly as $\sum_{i=1}^{q} \alpha_i(w_i - c_i)$ where q is a particular natural number, $\{w_i\}_{i=1}^{q}$ is a particular sequence of elements from V_X, $\{c_i\}_{i=1}^{q}$ is a particular sequence of elements from V_Y, and $\{\alpha_i\}_{i=1}^{q}$ is a particular sequence of positive numbers that sum to 1. We can also write $v = \sum_{i=1}^{q} \alpha_i(w_i - c_i) = (\sum_{i=1}^{q} \alpha_i w_i) - (\sum_{i=1}^{q} \alpha_i c_i)$. We can express $\sum_{i=1}^{q} \alpha_i w_i$ as $\sum_{i=1}^{m} \gamma_i x_i = m_X$, where $m_X \in X$ and $\{\gamma_i\}_{i=1}^{m}$ is a particular sequence of non-negative real numbers that sum to 1, and $\sum_{i=1}^{q} \alpha_i c_i$ as $\sum_{j=1}^{n} \epsilon_j y_j = m_Y$, where $m_Y \in Y$ and $\{\epsilon_j\}_{j=1}^{n}$ is a particular sequence of non-negative real numbers that sums to 1. (In particular, for each $k \in \{1, \ldots, m\}$, $\gamma_k = \sum_{\{i \mid i \in \{1,\ldots,q\} \wedge w_i = x_k\}} \alpha_i$ and for each $k \in \{1, \ldots, n\}$, $\epsilon_k = \sum_{\{i \mid i \in \{1,\ldots,q\} \wedge c_i = y_k\}} \alpha_i$.) Fix for the remainder of the chapter these definitions of $\{\gamma_i\}_{i=1}^{m}$, $\{\epsilon_j\}_{j=1}^{n}$, m_X, and m_Y.

Our goal is, by the previous paragraph, the determination of v by creating a minimizing sequence for $\|\cdot\|$ on D. Any such sequence will converge to $m_X - m_Y$ as shown earlier. One technique from the literature exploits the two ways of expressing D noted in lemma

1.2: instead of thinking of D as the "difference of convex hulls" of V_x and V_y (expressed as $\{x - y \mid x \in X \wedge y \in Y\}$, this technique concentrates on expressing D as the "convex hull of the differences" of V_x and V_y (expressed as the convex hull of $\{x - y \mid x \in V_x \wedge y \in V_y\}$). A method for determining v by seeking where the minimum value of $\|\cdot\|^2$ is attained is used: KKT, or a similar quadratic programming algorithm. (Norm-minimization for convex polytopes appears to be a well-studied problem.) However, each technique that treats D as the convex hull of a single polytope is applied to the convex hull of a set that is generally of size mn. We present a different approach below (an algorithm proposed by Dr. Kallman) that avoids the product-based proliferation. Our approach is an extension of the first algorithm shown in [4].

3.2. Lemmata and Proofs

We set forth a procedure for determining $m_X - m_Y$ to any desired degree of precision. It produces a sequence of pairs of elements $\{(a_i, b_i)\}_{i=1}^{z}$, where z may be a positive integer or ∞, and each element of the sequence belongs to $X \times Y$. The sequence $\{a_i - b_i\}_{i=1}^{z}$ converges to $m_X - m_Y$, if z is ∞, or terminates with $m_X - m_Y$, if z is finite. The proof of correctness for this procedure is a novel variant of that in [4].

Let $(x_*, y_*) \in X \times Y$. Define $c_X \colon X \times Y \to V_X$ be given by letting $c_X(x_*, y_*)$ be the element of least index x_p in V_X such that $\langle x_p - x_*, y_* - x_* \rangle = \max_{i \in \{1,\ldots,m\}} \langle x_i - x_*, y_* - x_* \rangle$. Similarly, let $c_Y \colon X \times Y \to V_y$ be given by letting $c_Y(x_*, y_*)$ be the element of least index y_q in V_Y such that $\langle y_q - y_*, x_* - y_* \rangle = \max_{i \in \{1,\ldots,n\}} \langle y_i - y_*, x_* - y_* \rangle$.

Now define $p_X \colon X \times Y \to \mathbf{R}$ by $p_X(x_*, y_*) = \langle c_X(x_*, y_*) - x_*, y_* - x_* \rangle$ and $p_Y \colon X \times Y \to \mathbf{R}$ by $p_Y(x_*, y_*) = \langle c_Y(x_*, y_*) - y_*, x_* - y_* \rangle$.

LEMMA 3.1. *For any $(x_*, y_*) \in X \times Y$, $x \in X$, and $y \in Y$, $\langle x - x_*, y_* - x_* \rangle \leq p_X(x_*, y_*)$ and $\langle y - y_*, x_* - y_* \rangle \leq p_Y(x_*, y_*)$.*

PROOF. We prove only that $\langle x - x_*, y_* - x_* \rangle \leq p_X(x_*, y_*)$, as the other proof is virtually identical.

Since $x \in X$, there exists a sequence $\{\alpha_i\}_{i=1}^m$ of non-negative real numbers such that $\sum_{i=1}^m \alpha_i = 1$ and $\sum_{i=1}^m \alpha_i x_i = x$. So:

$$\begin{aligned} p_X(x_*, y_*) &= \langle c_X(x_*, y_*) - x_*, y_* - x_* \rangle \\ &= \sum_{i=1}^m \alpha_i \langle c_X(x_*, y_*) - x_*, y_* - x_* \rangle \\ &\geq \sum_{i=1}^m \alpha_i \langle x_i - x_*, y_* - x_* \rangle \end{aligned}$$

and

$$\begin{aligned} p_X(x_*, y_*) &\geq \sum_{i=1}^m \alpha_i \langle x_i - x_*, y_* - x_* \rangle \\ &= \sum_{i=1}^m \alpha_i \left(\langle x_i, y_* - x_* \rangle - \langle x_*, y_* - x_* \rangle \right) \\ &= \left(\sum_{i=1}^m \alpha_i \langle x_i, y_* - x_* \rangle \right) - \left(\sum_{i=1}^m \alpha_i \langle x_*, y_* - x_* \rangle \right) \\ &= \left\langle \sum_{i=1}^m \alpha_i x_i, y_* - x_* \right\rangle - \langle x_*, y_* - x_* \rangle \\ &= \langle x, y_* - x_* \rangle - \langle x_*, y_* - x_* \rangle \\ &= \langle x - x_*, y_* - x_* \rangle \end{aligned}$$

□

Lemma 3.2. *For each $(x_*, y_*) \in X \times Y$, $p_X(x_*, y_*) \geq 0$ and $p_Y(x_*, y_*) \geq 0$.*

Proof. By lemma 3.1, $\langle x_* - x_*, y_* - x_* \rangle \leq p_X(x_*, y_*)$ and $\langle y_* - y_*, x_* - y_* \rangle \leq p_Y(x_*, y_*)$. Since each given inner product is clearly 0, the lemma follows. □

Lemma 3.3. *For any $(x_*, y_*) \in X \times Y$,*

$$\sqrt{2(p_X(x_*, y_*) + p_Y(x_*, y_*))} \geq \|(x_* - y_*) - (m_X - m_Y)\| \geq \left| \|x_* - y_*\| - \|m_X - m_Y\| \right|$$

Proof. By lemma 3.1, we have

$$\langle x_* - m_X, x_* - y_* \rangle = \langle m_X - x_*, y_* - x_* \rangle \leq p_X(x_*, y_*)$$

and

$$\langle m_Y - y_*, x_* - y_* \rangle \le p_Y(x_*, y_*)$$

So

$$\langle (x_* - y_*) - (m_X - m_Y), x_* - y_* \rangle$$

$$= \langle x_* - m_X, x_* - y_* \rangle + \langle m_Y - y_*, x_* - y_* \rangle$$

$$\le p_X(x_*, y_*) + p_Y(x_*, y_*)$$

and

$$\langle (x_* - y_*) - (m_X - m_Y), x_* - y_* \rangle$$

$$= \|x_* - y_*\|^2 - \langle m_X - m_Y, x_* - y_* \rangle$$

$$\le p_X(x_*, y_*) + p_Y(x_*, y_*)$$

Since $\|m_X - m_Y\|^2 \le \|x_* - y_*\|^2$, $\|m_X - m_Y\|^2 - \langle m_X - m_Y, x_* - y_* \rangle \le p_X(x_*, y_*) + p_Y(x_*, y_*)$, as well, so

$$2(p_X(x_*, y_*) + p_Y(x_*, y_*)) \ge$$

$$\|x_* - y_*\|^2 - \langle m_X - m_Y, x_* - y_* \rangle +$$

$$\|m_X - m_Y\|^2 - \langle m_X - m_Y, x_* - y_* \rangle =$$

$$\|x_* - y_*\|^2 - 2 \langle m_X - m_Y, x_* - y_* \rangle + \|m_X - m_Y\|^2$$

$$\|(x_* - y_*) - (m_X - m_Y)\|^2 \ge 0$$

hence $\sqrt{2(p_X(x_*, y_*) + p_Y(x_*, y_*))} \ge \|(x_* - y_*) - (m_X - m_Y)\|$. The triangle inequality gives us $\|(x_* - y_*) - (m_X - m_Y)\| \ge |\,\|(x_* - y_*)\| - \|(m_X - m_Y)\|\,|$ and the lemma has been proved. □

We define a sequence of elements of $X \times Y$, possibly finite, in the following way: Let x_{*_0} be a random element of X and y_{*_0} be a random element of Y. Given $x_{*_k} \in X$ and $y_{*_k} \in Y$, let $x_* = x_{*_k}$ and $y_* = y_{*_k}$. Let $x_p = c_X(x_*, y_*)$ and $y_q = c_Y(x_*, y_*)$. If $p_X(x_*, y_*) = \langle x_p - x_*, y_* - x_* \rangle = 0$ and $p_Y(x_*, y_*) = \langle y_q - y_*, x_* - y_* \rangle = 0$, the sequence terminates. Otherwise, since $\|\lambda x_p + (1-\lambda)x_* - (\mu y_q + (1-\mu)y_*)\|$, viewed as a function of (λ, μ) is a continuous function from $[0,1]^2$ to $\mathbf{R}$, it attains a minimum value, and let $\lambda, \mu \in [0,1]$ be otherwise arbitrary numbers such that $\|\lambda x_p + (1-\lambda)x_* - (\mu y_q + (1-\mu)y_*)\|$ is minimized. Then let $x_{*_{k+1}} = \lambda x_p + (1-\lambda)x_*$ and $y_{*_{k+1}} = \mu y_q + (1-\mu)y_*$. Thus the sequence is defined inductively.

Lemma 3.4. *The sequence $\|x_{*_k} - y_{*_k}\|$ is monotone non-increasing.*

Proof. For any k, if $x_{*_{k+1}}$ and $y_{*_{k+1}}$ exist, then recall that $x_{*_{k+1}} = \lambda c_X(x_{*_k}, y_{*_k}) + (1-\lambda)x_{*_k}$ and $y_{*_{k+1}} = \mu c_Y(x_{*_k}, y_{*_k}) + (1-\mu)y_{*_k}$, where $\lambda \in [0,1]$ and $\mu \in [0,1]$ and $\|\lambda c_X(x_{*_k}, y_{*_k}) + (1-\lambda)x_{*_k} - (\mu c_Y(x_{*_k}, y_{*_k}) + (1-\mu)y_{*_k})\|$ is minimal. So

$$\begin{aligned}
&\left\|x_{*_{k+1}} - y_{*_{k+1}}\right\| \\
&\quad = \min_{(\lambda,\mu)\in[0,1]^2} \|\lambda c_X(x_{*_k}, y_{*_k}) + (1-\lambda)x_{*_k} - (\mu c_Y(x_{*_k}, y_{*_k}) + (1-\mu)y_{*_k})\| \\
&\quad \le \|\lambda c_X(x_{*_k}, y_{*_k}) + (1-\lambda)x_{*_k} - (\mu c_Y(x_{*_k}, y_{*_k}) + (1-\mu)y_{*_k})\|\Big|_{(\lambda,\mu)=(0,0)} \\
&\quad = \|x_{*_k} - y_{*_k}\|
\end{aligned}$$

So for any k such that x_{*_k}, y_{*_k}, $x_{*_{k+1}}$, and $y_{*_{k+1}}$ exist, $\left\|x_{*_{k+1}} - y_{*_{k+1}}\right\| \le \|x_{*_k} - y_{*_k}\|$. Thus, by a trivial induction proof, it follows that the indicated sequence is monotone non-increasing.

□

Lemma 3.5. *Let $x_{*_k}, y_{*_k}, x_{*_{k+1}}, y_{*_{k+1}}$ all exist. Then*

$$\begin{aligned}
&\|x_{*_k} - y_{*_k}\|^2 - \left\|x_{*_{k+1}} - y_{*_{k+1}}\right\|^2 \ge \\
&\qquad\qquad \min(p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k}), \\
&\qquad\qquad\qquad (p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k}))^2/(\mathrm{diam}(X) + \mathrm{diam}(Y) + 1)^2)
\end{aligned}$$

Proof. Let $x_* = x_{*_k}$ and $y_* = y_{*_k}$. Let $x_p = c_X(x_*, y_*)$ and $y_q = c_Y(x_*, y_*)$. Then

$$\begin{aligned}
&\left\|x_{*_k} - y_{*_k}\right\|^2 - \left\|x_{*_{k+1}} - y_{*_{k+1}}\right\|^2 \\
&= \|x_* - y_*\|^2 - \left(\min_{(\lambda,\mu)\in[0,1]\times[0,1]} \|x_* - y_* + \lambda(x_p - x_*) - \mu(y_q - y_*)\|\right)^2 \\
&= \|x_* - y_*\|^2 - \left(\min_{(\lambda,\mu)\in[0,1]\times[0,1]} \|x_* - y_* + \lambda(x_p - x_*) - \mu(y_q - y_*)\|^2\right) \\
&= \max_{(\lambda,\mu)\in[0,1]\times[0,1]} \|x_* - y_*\|^2 - \|x_* - y_* + \lambda(x_p - x_*) - \mu(y_q - y_*)\|^2 \\
&\geq \max_{\lambda\in[0,1]} \|x_* - y_*\|^2 - \|x_* - y_* + \lambda((x_p - x_*) - (y_q - y_*))\|^2 \\
&= \max_{\lambda\in[0,1]} \|x_* - y_*\|^2 - (\|x_* - y_*\|^2 + \lambda^2 \|(x_p - x_*) - (y_q - y_*)\|^2 + \\
&\qquad 2\lambda \langle x_* - y_*, (x_p - x_*) - (y_q - y_*)\rangle) \\
&= \max_{\lambda\in[0,1]} -(\lambda^2 \|(x_p - x_*) - (y_q - y_*)\|^2 + \\
&\qquad 2\lambda \langle x_* - y_*, x_p - x_*\rangle - 2\lambda \langle x_* - y_*, y_q - y_*\rangle) \\
&= \max_{\lambda\in[0,1]} -(\lambda^2 \|(x_p - x_*) - (y_q - y_*)\|^2 - 2\lambda p_X(x_*, y_*) - 2\lambda p_Y(x_*, y_*)) \\
&= \max_{\lambda\in[0,1]} 2\lambda(p_X(x_*, y_*) + p_Y(x_*, y_*)) - \lambda^2 \|(x_p - x_*) - (y_q - y_*)\|^2
\end{aligned}$$

Consider the expression

$$E(\lambda) = 2\lambda(p_X(x_*, y_*) + p_Y(x_*, y_*)) - \lambda^2 \|(x_p - x_*) - (y_q - y_*)\|^2$$

. $p_X(x_*, y_*) + p_Y(x_*, y_*) \geq 0$ by lemma 3.2.

If $\|(x_p - x_*) - (y_q - y_*)\|^2 = 0$, $E(\lambda)$ clearly has its maximum value when $\lambda = 1$ and this maximum value is clearly $2(p_X(x_*, y_*) + p_Y(x_*, y_*))$, which is no less than $p_X(x_*, y_*) + p_Y(x_*, y_*)$.

Otherwise, $\|(x_p - x_*) - (y_q - y_*)\|^2 > 0$ and elementary considerations dictate that $E(\lambda)$ attains its minimum value at

$$\lambda = \min(\max((p_X(x_*, y_*) + p_Y(x_*, y_*))/\|(x_p - x_*) - (y_q - y_*)\|^2, 0), 1)$$

Since $p_X(x_*, y_*) + p_Y(x_*, y_*) \geq 0$, this is just

$$\min((p_X(x_*, y_*) + p_Y(x_*, y_*))/\left\|(x_p - x_*) - (y_q - y_*)\right\|^2, 1)$$

If $p_X(x_*, y_*) + p_Y(x_*, y_*) \geq \|(x_p - x_*) - (y_q - y_*)\|^2$, $E(\lambda)$ attains its maximum at $\lambda = 1$. Then, $p_X(x_*, y_*) + p_Y(x_*, y_*) \geq \|(x_p - x_*) - (y_q - y_*)\|^2$, so

$$p_X(x_*, y_*) + p_Y(x_*, y_*) - \|(x_p - x_*) - (y_q - y_*)\|^2 \geq 0$$

and

$$\begin{aligned} E(1) &= 2(p_X(x_*, y_*) + p_Y(x_*, y_*)) - \|(x_p - x_*) - (y_q - y_*)\|^2 \\ &= p_X(x_*, y_*) + p_Y(x_*, y_*) + p_X(x_*, y_*) + p_Y(x_*, y_*) - \|(x_p - x_*) - (y_q - y_*)\|^2 \\ &\geq p_X(x_*, y_*) + p_Y(x_*, y_*) \end{aligned}$$

so E has maximum value no less than $p_X(x_*, y_*) + p_Y(x_*, y_*)$.

Finally, if $p_X(x_*, y_*) + p_Y(x_*, y_*) < \|(x_p - x_*) - (y_q - y_*)\|^2$, E attains its minimum value at $\lambda = (p_X(x_*, y_*) + p_Y(x_*, y_*))/\|(x_p - x_*) - (y_q - y_*)\|^2$ and the attained value is $(p_X(x_*, y_*) + p_Y(x_*, y_*))^2/\|(x_p - x_*) - (y_q - y_*)\|^2$. We have:

$$\begin{aligned} \operatorname{diam}(X) + \operatorname{diam}(Y) + 1 &\geq \operatorname{diam}(X) + \operatorname{diam}(Y) \\ &\geq \|x_p - x_*\| + \|y_* - y_q\| \\ &\geq \|x_p - x_* + y_* - y_q\| \geq 0 \end{aligned}$$

implying that $1/\|x_p - x_* - (y_q - y_*)\|^2 \geq 1/(\operatorname{diam}(X) + \operatorname{diam}(Y) + 1)^2$, which implies

$$\begin{aligned} &(p_X(x_*, y_*) + p_Y(x_*, y_*))^2/\|(x_p - x_*) - (y_q - y_*)\|^2 \geq \\ &\qquad (p_X(x_*, y_*) + p_Y(x_*, y_*))^2/(\operatorname{diam}(X) + \operatorname{diam}(Y) + 1)^2 \end{aligned}$$

Thus the conclusion follows. □

Theorem 3.6. *The sequence $x_{*_k} - y_{*_k}$ either terminates with value $m_X - m_Y$ if the sequence is finite, or converges to $m_X - m_Y$ if the sequence is not finite.*

Proof. If the sequence terminates, then for the index k of the last element of the sequence, $p_X(x_{*_k}, y_{*_k}) = p_Y(x_{*_k}, y_{*_k}) = 0$. By lemma 3.3 therefore,

$$0 = \sqrt{2(p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k}))} \geq \|(x_* - y_*) - (m_X - m_Y)\| \geq 0$$

so $\|(x_* - y_*) - (m_X - m_Y)\| = 0$ and $x_* - y_* = m_X - m_Y$.

Otherwise, the sequence does not terminate. Consider the sequence $\{\gamma_n\}_{n=0}^{\infty}$ given by $\gamma_n = p_X(x_{*_n}, y_{*_n}) + p_Y(x_{*_n}, y_{*_n})$. Assume for purposes of contradiction that the sequence γ does not converge to 0. Then there exists $\omega > 0$ such that for any index n there exists an index $m \geq n$ such that $|\gamma_m - 0| \geq \omega$. Since $\gamma_m = p_X(x_{*_m}, y_{*_m}) + p_Y(x_{*_m}, y_{*_m}) \geq 0$ by lemma 3.2, this implies that $\gamma_m \geq \omega$. Let

$$\mu = \min(\omega, \omega^2/(\mathrm{diam}(X) + \mathrm{diam}(Y) + 1)^2)$$

and note that $\mu > 0$. Also note that for any k such that $\gamma_k \geq \omega$,

$$\begin{aligned} \mu \leq \min(& p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k}), \\ & (p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k}))^2/(\mathrm{diam}(X) + \mathrm{diam}(Y) + 1)^2) \leq \\ & \|x_{*_k} - y_{*_k}\|^2 - \left\|x_{*_{k+1}} - y_{*_{k+1}}\right\|^2 \end{aligned}$$

where the last inequality follows by lemma 3.5. Define a function r from non-negative integers to non-negative integers in the following way: Let $r(0)$ be the least integer greater than or equal to 0 such that $\gamma_{r(0)} \geq \omega$. This index always exists by hypothesis. Given that $r(n)$ is defined, let $r(n+1)$ be the least integer greater than or equal to $r(n) + 1$ such that $\gamma_{r(n+1)} \geq \omega$. This index also exists by hypothesis. Note that for each non-negative integer n, $r(n)$ is such that $\gamma_{r(n)} \geq \omega$, hence $\left\|x_{*_{r(n)}} - y_{*_{r(n)}}\right\|^2 - \left\|x_{*_{r(n+1)}} - y_{*_{r(n+1)}}\right\|^2 \geq \mu$. Let z be the least non-negative integer strictly greater than $\|x_{*_0} - y_{*_0}\|^2 / \mu$ and consider the sum

$$G = \sum_{i=0}^{r(z-1)} \|x_{*_i} - y_{*_i}\|^2 - \left\|x_{*_{i+1}} - y_{*_{i+1}}\right\|^2$$

By lemma 3.4, this is a sum with exclusively non-negative terms. The index i attains each value in the set $R = \{r(0), \ldots, r(z-1)\}$. So by the previous note, for each such i, at least μ is added to the sum G. Hence the indicated sum contains at least z terms with value no less than μ, so the indicated sum G is at least $z\mu$. Since $\mu > 0$ and $z > \|x_{*_0} - y_{*_0}\|^2/\mu$, $G \geq z\mu > \mu(\|x_{*_0} - y_{*_0}\|^2/\mu) = \|x_{*_0} - y_{*_0}\|^2$. Now

$$\begin{aligned}
\left\|x_{*_{r(z-1)+1}} - y_{*_{r(z-1)+1}}\right\|^2 &= \|x_{*_0} - y_{*_0}\|^2 + \sum_{i=0}^{r(z-1)} \left\|x_{*_{i+1}} - y_{*_{i+1}}\right\|^2 - \|x_{*_i} - y_{*_i}\|^2 \\
&= \|x_{*_0} - y_{*_0}\|^2 - G \\
&\leq \|x_{*_0} - y_{*_0}\|^2 - z\mu \\
&< \|x_{*_0} - y_{*_0}\|^2 - \|x_{*_0} - y_{*_0}\|^2 \\
&< 0
\end{aligned}$$

an absurdity. Hence the sequence $p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k})$ must converge to 0. Since this sequence converges to 0 and is non-negative, its continuous image

$$\sqrt{2(p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k}))}$$

also converges to 0. By lemma 3.3,

$$0 \leq \|(x_{*_k} - y_{*_k}) - (m_X - m_Y)\| \leq \sqrt{2(p_X(x_{*_k}, y_{*_k}) + p_Y(x_{*_k}, y_{*_k}))}$$

for all indices k, so by the Sandwich Theorem, the sequence $\|(x_{*_k} - y_{*_k}) - (m_X - m_Y)\|$ converges to 0 as well. Thus, the sequence $x_{*_k} - y_{*_k}$ converges to $m_X - m_Y$. □

BIBLIOGRAPHY

[1] Salomon Bochner, *Stable laws of probability and completely monotone functions*, Duke Mathematical Journal 3 (1937), no. 4, 726–728.

[2] Robert R. Kallman, *Geometric methods for support vector machines*, preprint.

[3] Paul Lévy, *Calcul des probabilités*, Gauthier-Villars, Paris, France, 1925.

[4] B.F. Mitchell, V.F Dem'yanov, and V.N. Malozemov, *Finding the point of a polyhedron nearest the origin*, SIAM Journal on Control 12 (February 1974), no. 12, 19 – 26.