Support Vector Machine Lagrange Multipliers and Simplex Volume Decompositions *

Tong Wen, Alan Edelman

Department of Mathematics
Laboratory of Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
tonywen@math.mit.edu
edelman@math.mit.edu

August 2000

Abstract

The Support Vector Machine (SVM) idea has attracted recent attention in solving classification and regression problems. As an example based method, SVMs distinguish two point classes by finding a separating boundary layer, which is determined by points that become known as Support Vectors (SVs). While the computation of the separating boundary layer is formulated as a linearly constrained Quadratic Programming (QP) problem, in practice the corresponding dual problem is computed.

This paper investigates how the solution to the dual problem depends on the geometry. When examples are separable, we will show that the Lagrange multipliers (the unknowns of the dual problem) associated with SVs can be interpreted geometrically as a normalized ratio of simplex volumes, and at the same time a simplex volume decomposition relation must be satisfied. Examples for the two and three dimensional cases are given during the discussion. Besides showing geometric properties of SVMs, we also suggest a way to investigate the distribution of the Lagrange multipliers based on a random matrix model. We finish this paper with a further analysis of how the Lagrange multipliers depend on three critical angles using the Singular Value and CS decompositions.

*This paper is supported by U.S. Army TACOM under the contract DAAHO4-96-C-0086 TCN 99024 awarded from the Battelle-Research Triangle Park and NSF under the grand DMS-9971591.

1 Introduction

How can we compute the distance that separates two sets of points? This problem which arises in such applications as Support Vector Machine (SVM) classifiers [3] can be formulated as a Quadratic Programming (QP) problem. This paper investigates how the Lagrange multipliers that arise in the QP formulation depend on the geometry of those points. Our guiding principle is that we can explain the solutions through the geometry of a linear system rather than the complexity of a QP problem.

In this paper, we first consider the problem of separating $n+1$ non-degenerate points into two sets in R^n, and show that these $n+1$ points can always be separated by the E-separating hyperplanes, which are defined to contain the points from each set respectively. Then we derive the geometric meaning of the Lagrange multipliers associated with the E-separating hyperplanes showing that they can be interpreted as a normalized ratio of simplex volumes. For the E-separating hyperplanes to be optimal, a simplex volume decomposition relation must be satisfied. When examples are separable, a binary SVM classifier is equivalent to the optimal hyperplanes separating two sets of points. After deriving the properties of Support Vectors (SVs), we show that the optimal separating hyperplanes are equivalent to the E-separating hyperplanes computed in the subspace determined by SVs. It follows that all the results we have obtained for the E-separating hyperplanes can be applied to SVM classifiers. Therefore, the Lagrange multiplier associated with each SV can be interpreted geometrically as a normalized ratio of simplex volumes, and at the same time a simplex volume decomposition relation must be satisfied. Moreover, based on a random matrix model we suggest a way to investigate the distribution of the Lagrange multipliers. This paper is finished by a further analysis of how the Lagrange multipliers depend on three critical angles.

As indicated by Figure 1, the organization of this paper might be conveniently characterized by the number of the points we want to separate . In Sections 2 and 3, we first consider the problem of separating $n+1$ non-degenerate points in R^n, then we formulate the separation of arbitrary N points as a QP problem. In Section 4, the geometric and statistical properties of the $n+1$ Lagrange multipliers associated with the E-separating hyperplanes in R^n are derived. Then we show that in Section 6 the QP problem of size N can be reduced to a smaller linear system of size n such that the results obtained in Section 4 can be applied. At the end, we continue our analysis by showing how the Lagrange multipliers associated with the $n+1$ SVs depend on three critical angles.

2 Separating Two Point Sets

We define the distance between two sets of points to be the maximum gap (if it exists) between two parallel hyperplanes that separate them. Finding the

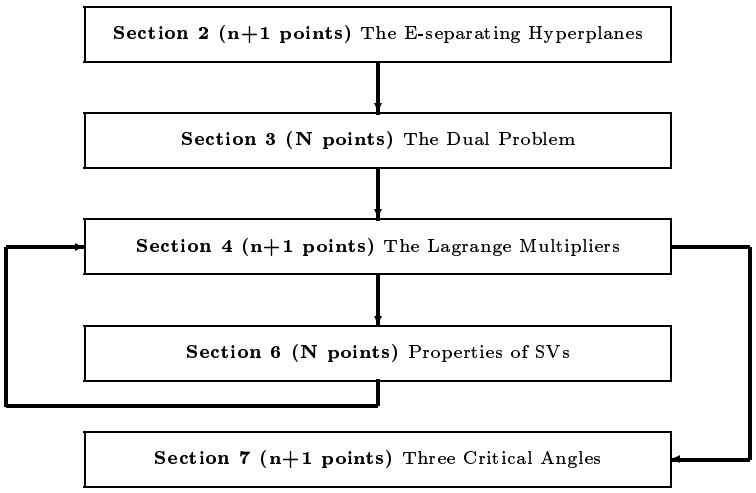


Figure 1: The organization of this paper can be characterized by the number of the points we want to separate. Section 5 provides background material and may be read independently.

pair of optimal separating hyperplanes is formulated as a QP problem. This problem arises in such applications as SVM classifiers (see Section 5).

We begin this section with the problem of how to separate $n+1$ points in R^n with two parallel hyperplanes. We will show that under the condition of non-degeneracy, $n+1$ points can always be separated in R^n by solving a linear system, but this system may or may not give the optimal separating hyperplanes. At the end of this section we formulate the problem of finding the optimal separating hyperplanes as a linearly constrained QP problem.

Formally, assume that there are $n+1$ points $\{x_1, \ldots, x_{n+1}\}$ in R^n, each of which belongs to one of two classes. We use ± 1 to represent each class respectively, and define $I_+ = \{1, \ldots, m\}$ to be the index set corresponding to the positive points and $I_- = \{m+1, \ldots, n+1\}$ for the negative points. To indicate whether x_i is positive or negative, a sign $y_i = \pm 1$ is assigned to each point x_i. It is also assumed that *if any point x_j is taken as the origin, then the resulting n vectors $x_i - x_j$ $(i \neq j)$ are linearly independent.* The problem is how to find the pair of optimal hyperplanes separating the positive and negative points.

Any pair of parallel hyperplanes can be expressed as

$$w^T x + b = \pm 1, \tag{2.1}$$

where $w \in \mathrm{R}^n$ gives the normal direction of the hyperplanes and b is a scalar. Since only the hyperplanes that can separate the two sets of points are considered, we must satisfy the conditions:

$$w^T x_i + b \geq 1 \text{ for } i \in I_+ \tag{2.2}$$

and

$$w^T x_i + b \leq -1 \text{ for } i \in I_-, \tag{2.3}$$

that is,

$$y_i(w^T x_i + b) \geq 1 \text{ for } i = 1, \ldots, n+1. \tag{2.4}$$

An easy consequence from the assumption of linear independence is that the $n+1$ points are separable, i.e., there must exist one pair of hyperplanes satisfying the separability condition (2.4). We state this result as a proposition.

PROPOSITION 2.1 *Suppose we are given $n+1$ points x_i in R^n. Under the condition of non-degeneracy, that is, if any point x_j is taken as the origin then the resulting n vectors $x_i - x_j$ $(i \neq j)$ are linearly independent, these $n+1$ points can always be separated into two sets by a pair of parallel hyperplanes.*

Proof. To demonstrate this, we will find w and b such that equality holds:

$$y_i(w^T x_i + b) = 1 \text{ for } i = 1, \ldots, n+1. \tag{2.5}$$

Without loss of generality, we assume that x_{n+1} is the origin. It follows immediately that $b = -1$. Thus for the remaining n points, the equalities in (2.5) become

$$w^T x_i = 2 \text{ for } i \in I_+, \tag{2.6}$$

and

$$w^T x_i = 0 \text{ for } i \in I_-. \tag{2.7}$$

Let $X = [x_1, \ldots, x_n]$ be a square matrix. The equations (2.6) and (2.7) uniquely determine the unknown vector w by the following linear system:

$$X^T w = \begin{bmatrix} \mathbf{2} \\ \mathbf{0} \end{bmatrix}, \tag{2.8}$$

where the sub-vector $\mathbf{2}$ corresponds to the positive points. From the assumption of linear independence, the matrix X is nonsingular so that the above linear system must have a unique solution w. Therefore, there always exists such a pair of hyperplanes that satisfies the separability condition (2.4). ■

In the following context, we denote the pair of hyperplanes satisfying Equation (2.5) by *E-separating* hyperplanes, the letter "E" represents equality. The two hyperplanes containing the positive and negative points are called positive hyperplane and negative hyperplane respectively.

DEFINITION 1 *Suppose we are given $n+1$ points x_i in R^n such that if any point x_j is taken as the origin then the resulting n vectors $x_i - x_j$ $(i \neq j)$ are linearly independent. The E-separating hyperplanes are the pair of hyperplanes satisfying Equation* (2.5).

Figure 2 shows the two dimensional case. It is easy to see how three vertices of a triangle uniquely determine the pair of lines satisfying Equation (2.5), i.e., the E-separating lines. One might be tempted to believe that the pair of E-separating hyperplanes gives the optimal separator, but a simple example in Figure 2(b) shows that this is not true. Here the pair of solid lines defines a larger gap than the pair of dotted E-separating lines.

How can we find the optimal separating hyperplanes? We know that the distance between the two parallel hyperplanes in (2.1) is $d = \frac{2}{\|w\|}$. Therefore, given any N points belonging to two classes, we can formulate the finding of

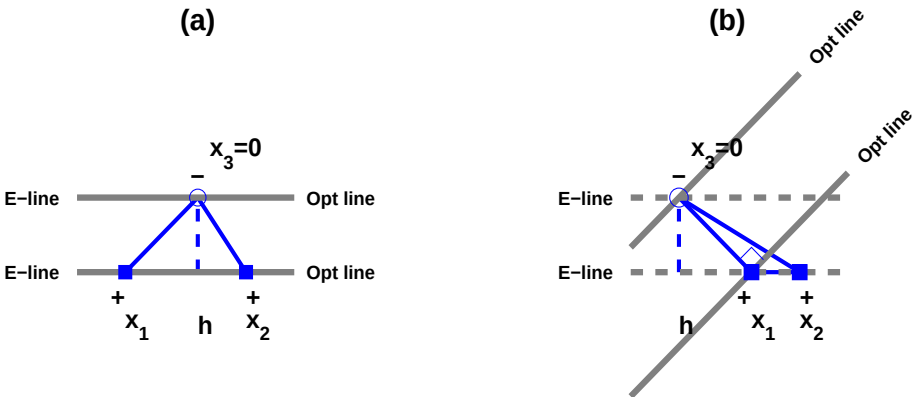


Figure 2: The horizontal lines labeled by "E-line" are the E-separating lines. In both cases, the solid lines labeled by "opt line" are the optimal separator. Squares and circles are used to indicate the positive and negative points respectively.

the optimal separating hyperplanes as the following linearly constrained QP problem:

$$\underset{w,b}{\text{minimize}} \quad f(w,b) \equiv \frac{1}{2}\|w\| \tag{2.9}$$

subject to:

$$y_i(w^T x_i + b) \geq 1, \quad \text{for } i = 1, ..., N. \tag{2.10}$$

If the two sets of points are separable, then the above QP problem must have an optimal solution, which is unique as proved in Section 6. In the following context, it is always assumed that the two sets of points are separable.

3 The Dual Problem and Optimal Conditions

In this section, we derive the dual problem to the linearly constrained QP problem (2.9), which is the one that is actually computed in practice. We also give conditions for judging if a solution is optimal.

Let α_i be the Lagrange Multipliers corresponding to the inequality constraints in (2.10). Then the Lagrangian is

$$L(w,b,\alpha) \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i (w^T x_i + b) + \sum_{i=1}^{N} \alpha_i, \tag{3.1}$$

where $\alpha = [\alpha_1, \ldots, \alpha_N]$ and $\alpha_i \geq 0$. By requiring the gradients of $L(w, b, \alpha)$ with respect to w and b vanish, the following relations are obtained:

$$w = \sum_{i=1}^{N} y_i \alpha_i x_i, \tag{3.2}$$

and

$$\sum_{i=1}^{N} y_i \alpha_i = 0. \tag{3.3}$$

Substituting the above equations into $L(w, b, \alpha)$, we get

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i^T x_j). \tag{3.4}$$

Maximizing $L(\alpha)$ or equivalently minimizing $-L(\alpha)$ subject to $\alpha_i \geq 0$ and Condition (3.3) gives the dual problem:

$$\underset{\alpha}{\text{minimize}} \quad F(\alpha) \equiv \frac{1}{2} \alpha^T H \alpha - \alpha^T \mathbf{1} \tag{3.5}$$

subject to:

$$y^T \alpha = 0, \tag{3.6}$$

$$\alpha \geq \mathbf{0}, \tag{3.7}$$

where $y = [y_1, \ldots, y_N]^T$, $\mathbf{1} = [1, \ldots, 1]^T$ and H is a symmetric semi-positive definite matrix with $H_{ij} = y_i (x_i^T x_j) y_j$. The objective function $F(\alpha)$ is equal to $-L(\alpha)$. From the fact that $f(w, b)$ and $F(\alpha)$ are equal at optimality, it follows that

$$\|w\|^2 = \sum_{j=1}^{N} \alpha_i. \tag{3.8}$$

Thus the distance between the two sets of points is

$$d = \frac{2}{\sqrt{\sum_{i=1}^{N} \alpha_i}}. \tag{3.9}$$

Since the above quadratic programming (QP) problem is convex, the Karush-Kuhn-Tucker (KKT) conditions (*page* 35, [2]) become both necessary and sufficient for α to be optimal. These KKT conditions can be summarized as the following:

$$y^T\alpha = 0, \tag{3.10}$$

$$\alpha \geq 0, \tag{3.11}$$

$$y_i(\sum_{j=1}^{N} \alpha_j y_j x_j^T x_i + b) \geq 1 \ \forall \ i, \tag{3.12}$$

$$y_i(\sum_{j=1}^{N} \alpha_j y_j x_j^T x_i + b) > 1 \text{ only if } \alpha_i = 0. \tag{3.13}$$

If $\alpha_i > 0$, then from KKT conditions (3.12) and (3.13) we know that

$$\sum_{j=1}^{N} \alpha_j y_j x_j^T x_i + b = y_i.$$

Therefore,

$$b = y_i - \sum_{j=1}^{N} \alpha_j y_j x_j^T x_i. \tag{3.14}$$

4 The Lagrange Multipliers of the E-separating Hyperplanes

As already seen in Figure 2, the pair of optimal separating lines depends on the shape of the triangle constructed by x_i. For an acute triangle such as in Figure 2(a), the E-separating lines are optimal. While for an obtuse triangle such as in Figure 2(b), the E-separating lines are not optimal. To investigate how the optimal separator depends on the geometry, we begin with the pair of E-separating hyperplanes. We first derive the geometric meaning of the Lagrange multipliers (the unknowns of the dual problem giving the E-separating hyperplanes), then show that the optimality conditions can be expressed geometrically as a simplex volume decomposition relation, which explains clearly how the optimal solutions in Figure 2 depend on the shape of the triangles.

4.1 The Geometric Meaning of the Lagrange Multipliers

If the inequality constraints in (2.10) are enforced to be active, i.e., (2.10) becomes (2.5), then the QP problem (2.9) will give us the pair of E-separating hyperplanes. Since only equality constraints are involved, the Lagrange multipliers α_i become unconstrained. In this section, we derive the formula for the Lagrange multipliers associated with the E-separating hyperplanes and show that they can be interpreted geometrically as a normalized ratio of simplex volumes.

THEOREM 4.1 *The Lagrange multipliers associated with the E-separating hyperplanes can be expressed as a normalized ratio of simplex volumes.*

Proof. Define h such that it has the same direction as w and its length gives the distance between the E-separating hyperplanes. Again without loss of generality, we assume that x_{n+1} is the origin. It follows that point h must be in the positive hyperplane, i.e.,

$$w^T h = 2. \tag{4.1}$$

From the definition of h, it is true that

$$w = \frac{2}{\|h\|^2} h \text{ and } h = \frac{2}{\|w\|^2} w. \tag{4.2}$$

Define $\beta_i = \alpha_i$ if $i \in I_+$ and $\beta_i = -\alpha_i$ if $i \in I_-$ for $i = 1, \ldots, n$. Then Equation (3.2) becomes

$$w = \sum_{i=1}^{n} \beta_i x_i = X\beta, \tag{4.3}$$

where $\beta = [\beta_1, \ldots, \beta_n]^T$. Combining the two equations above, we obtain the following linear system:

$$X\beta = \frac{2}{\|h\|^2} h. \tag{4.4}$$

The solution to this linear system in terms of determinants tells us that

$$\alpha_l = y_l \beta_l = \frac{2y_l}{\|h\|^2} \frac{\det X_l}{\det X}, \tag{4.5}$$

where $X_l = [x_1, \ldots, x_{l-1}, h, x_{l+1}, \ldots, x_n]$ for $l = 1, \ldots, n$. If $\alpha_1, \ldots, \alpha_n$ are known, α_{n+1} can be determined by Equation (3.6), i.e.,

$$\alpha_{n+1} = \sum_{i \in I_+} \alpha_i - \sum_{i \in (I_- - \{n+1\})} \alpha_i. \tag{4.6}$$

Recall that

$$\det X = \pm n! \text{vol } X.$$

Therefore, α_l can be expressed geometrically as

$$\alpha_l = \frac{\pm 2}{\|h\|^2} \frac{\text{vol } X_l}{\text{vol } X} \tag{4.7}$$

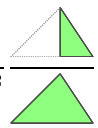
for $l = 1, \ldots, n$, and

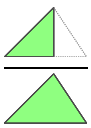
$$\alpha_{n+1} = \frac{2}{\|h\|^2} \frac{\sum_{i \in (I_+ \cup I_- - \{n+1\})} \pm \text{vol } X_i}{\text{vol } X}. \tag{4.8}$$

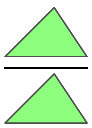
■

Equation (4.7) and (4.8) indicate clearly how the Lagrange multipliers depend on the geometry of points x_i and h. In Figure 2(a), the E-separating lines are optimal. The three corresponding Lagrange multipliers in terms of the ratio of simplex volumes are shown in the following:

$$\alpha_1 = \frac{2}{\|h\|^2}\frac{\quad}{\quad}, \quad \alpha_2 = \frac{2}{\|h\|^2}\frac{\quad}{\quad}, \quad \alpha_3 = \frac{2}{\|h\|^2}\frac{\quad}{\quad} = \frac{2}{\|h\|^2}.$$







4.2 The Simplex Volume Decomposition at Optimality

In this section, we show that the necessary and sufficient condition for the E-separating hyperplanes to be optimal can be expressed geometrically as a simplex volume decomposition relation.

Using KKT condition (3.10), Equation (3.8) can be written as

$$\begin{aligned} \|w\|^2 = \sum_{i=1}^{n+1} \alpha_i &= 2 \sum_{i \in I_+} \alpha_i \\ &= 2 \sum_{i \in I_-} \alpha_i \\ &= \frac{4}{\|h\|^2}. \end{aligned} \tag{4.9}$$

Choosing a different point as the origin will give different simplices $\tilde{X}$ and $\tilde{X}_l$ in Formula (4.7), but the ratio of their determinants is unchanged. Therefore, the Lagrange multipliers α_i do not depend on the choice of the origin. In the following context, we always assume that a positive point is chosen as the origin when α_i with $i \in I_-$ is considered, and a negative point is chosen as the origin when α_i with $i \in I_+$ is considered, so that Formula (4.5) can be applied without considering the case (4.6).

Substituting α_i with Formula (4.5) into Equation (4.9) gives us the following relation:

$$\begin{aligned} \det X &= \sum_{i \in I_+} \det X_i \\ \det \tilde{X} &= -\sum_{i \in I_-} \det \tilde{X}_i, \end{aligned} \tag{4.10}$$

where $\tilde{X}$ and $\tilde{X}_i$ are the simplices obtained by choosing a positive point as the origin. From KKT conditions (3.10) – (3.13) and the definition of the E-separating hyperplanes, we know that if all the Lagrange multipliers α_i are nonnegative then the E-separating hyperplanes must be optimal and vice versa.

If α_i is positive, then by Formula (4.5) $\det X$ and $\det X_i$ must have identical signs if $i \in I_+$, or $\det \tilde{X}$ and $\det \tilde{X}_i$ must have different signs if $i \in I_-$. It follows that when the E-separating hyperplanes are optimal, Relation (4.10) becomes

$$\begin{aligned} \text{vol } X &= \sum_{i \in I_+} \text{vol } X_i \\ \text{vol } \tilde{X} &= \sum_{i \in I_-} \text{vol } \tilde{X}_i. \end{aligned} \tag{4.11}$$

On the other hand, if the above relation is true, we want to show that all the α_i are nonnegative. In (4.11), writing the first equation in terms of determinants, we have

$$|\det X| = \sum_{i \in I_+} |\det X_i|. \tag{4.12}$$

Suppose $\det X = \text{vol } X$. Subtracting the first equation in (4.10) from Equation (4.12) gives us

$$0 = \sum_{i \in I_+} |\det X_i| - \det X_i. \tag{4.13}$$

Since each term is nonnegative, $|\det X_i| - \det X_i$ must be zero, i.e.,

$$\det X_i = \text{vol } X_i.$$

Therefore, α_i must be nonnegative according to (4.5) for $i \in I_+$. The same result can be derived for the case when $\det X = -\text{ vol } X$. By the same reasoning, it is also ture that α_i must be nonnegative for $i \in I_-$ if the second equation in (4.11) is satisfied. Therefore, we can conclude that the E-separating hyperplanes must be optimal.

We state the above results in the following theorem:

THEOREM 4.2 *The pair of E-separating hyperplanes is optimal if and only if the simplex volume decomposition relation* (4.11) *is satisfied.*

If there is only one negative point as in the two dimensional cases shown by Figure 2, the simplex volume decomposition relation (4.11) can be replaced by a simplex decomposition relation:

$$\text{simplex } X = \sum_{i \in I_+} \text{simplex } X_i. \tag{4.14}$$

From the geometry, it is easy to see that this relation is true if and only if h is inside the symplex constructed by the positive points in the positive hyperplane.

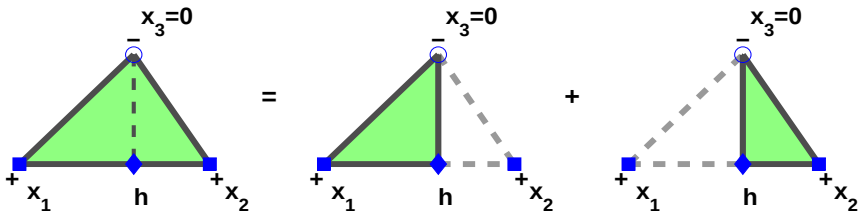


Figure 3: The simplex decomposition for an acute triangle

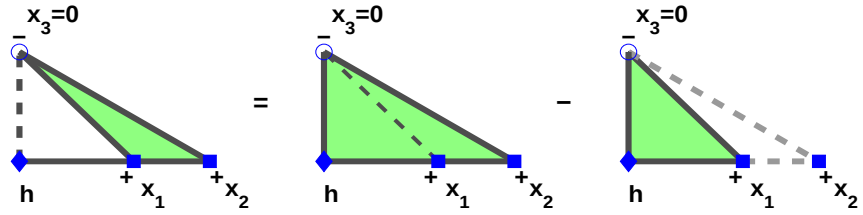


Figure 4: The simplex decomposition for an obtuse triangle

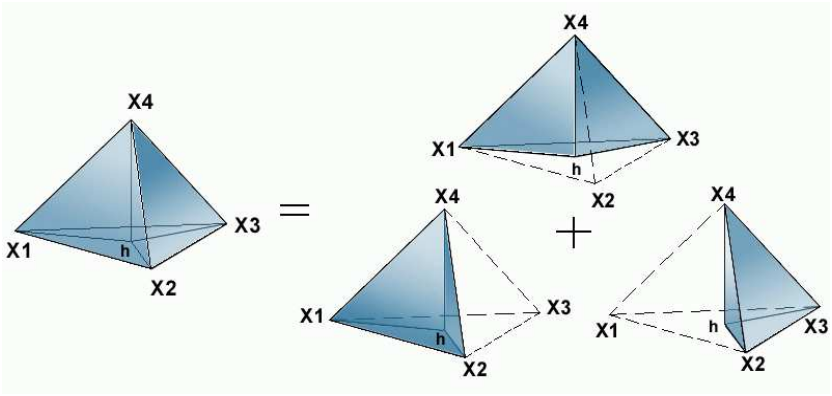


Figure 5: A three dimensional simplex decomposition where h is inside the base triangle

The relation (4.10) for the cases described in Figure 2(a) and Figure 2(b) is illustrated by Figure 3 and Figure 4 respectively. For acute triangles, since h is between x_1 and x_2, the pair of E-separating lines is optimal. While for obtuse triangles, since the above relation is not satisfied, the pair of E-separating lines is not optimal. Figure 5 shows a three dimensional simplex decomposition where the E-planes separating the points $\{x_1, x_2, x_3\}$ and $\{x_4\}$ are optimal.

If the simplex volume decomposition relation is not satisfied, then the E-separating hyperplanes are not optimal. To achieve the optimal separator such as the pair of solid lines in Figure 2(b), some conditions in (2.5) must be relaxed to inequalities.

REMARK 4.1 *It is not true that the equality constraint with the most negative Lagrange multiplier must be relaxed. A counter example can be constructed.*

4.3 A Random Matrix Model: The Lagrange Multiplier Distribution

For the E-separating hyperplanes, α_i can be either positive or negative as indicated by Formula (4.7). It is interesting to study the distribution of sign(α_i).

As a model, we will assume that X is a random matrix whose elements are N(0, 1) (standard normal distribution) and independent. In statistics, $W = X^T X$ is called a Wishart matrix $W_n(n, nI_{n\times n})$ where $I_{n\times n}$ represents a $n \times n$ idendity matrix (*page* 82, [7]). Multiplying Equation (4.4) by X^T, we can see that $\alpha(1:n)$ is determined by the inverse Wishart matrix W^{-1}:

$$\alpha(1:n) = \text{diag}(y)W^{-1}\begin{bmatrix} \mathbf{2} \\ \mathbf{0} \end{bmatrix}. \tag{4.15}$$

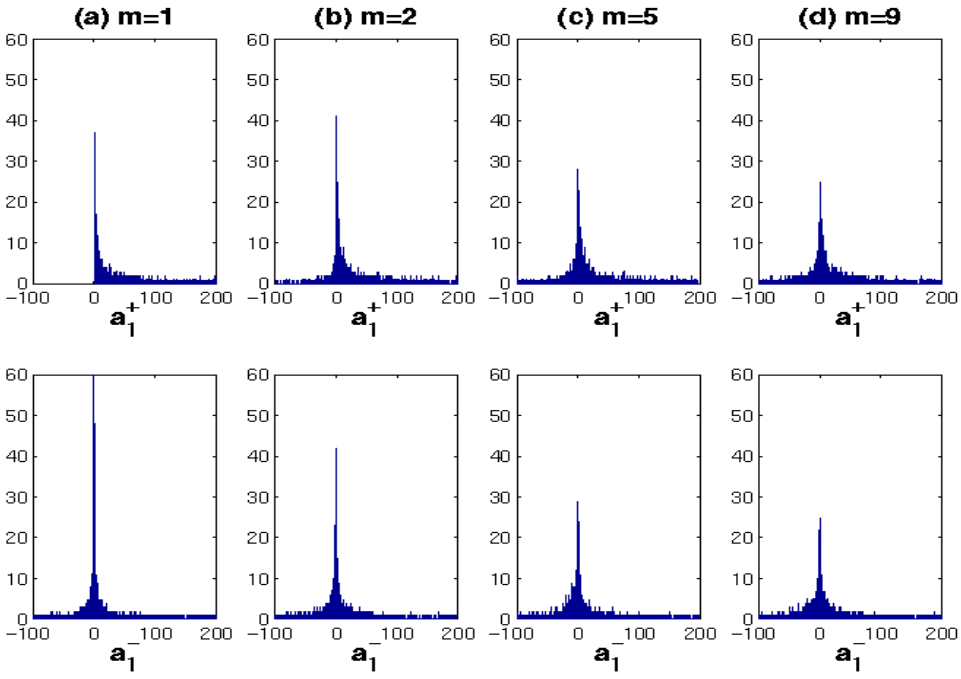


Figure 6: Histograms of α_1^+ and α_1^-. Each column of sub-figures corresponds to a different m. The size of the experiment is 8000 and $n = 12$.

Since permuting the columns of X does not affect the distribution of W^{-1}, we can conclude that the Lagrange multipliers associated with the positive points must have identical distributions and so do the Lagrange multipliers associated with the negative points (except α_{n+1}). Therefore, in the following context it is general enough to only consider two Lagrange multipliers α_1 and $\alpha_{\min\{m+1,n\}}$, which are denoted by α_1^+ and α_1^- respectively. Again m gives the number of positive points.

In Figure 6, each column of sub-figures plots experimental histograms of α_1^+ and α_1^- corresponding to a different m. These histograms tell us that the p.d.f. of α_1^+ tends to spread out (have larger variance) and become more symmetric about the y-axis as m increases, and at the same time the p.d.f. of α_1^- always has a similar shape but also tends to spread out as m increases. From the first sub-plot in Figure 6(a), we can see that α_1^+ is always positive at $m = 1$. The above observation is consistent with the approximations to $\Pr(\alpha_i \geq 0)$ for $i = 1, \ldots, n$ as shown in Figure 7. The plot in Figure 7(a) indicates that $\Pr(\alpha_1^+ \geq 0)$ monotonically decreases as m increases since increasing m moves the median of α_1^+ toward the origin, and $\Pr(\alpha_1^- \geq 0) = 0.5$, i.e., the median of α_1^- is zero for any m. While the plot in Figure 7(b) indicates that $\Pr(\alpha_1^+ \geq 0)$

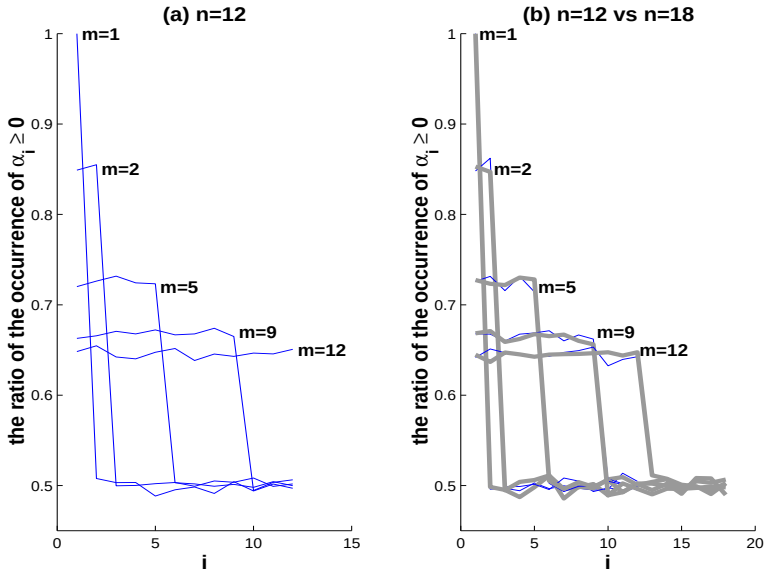


Figure 7: Each curve plots the ratio of the occurrence of $\alpha_i \geq 0$ for $i = 1, \ldots, n$ during 8000 computations at a different m. In sub-figure (b), two cases with $n = 12$ and $n = 18$ are plotted against each other indicating that $\Pr(\alpha_i \geq 0)$ is independent of n.

and $\Pr(\alpha_1^- \geq 0)$ do not depend on n.

Although the density function of W^{-1} is known (*page* 113, [7]), we have not at this time chosen to verify the above observations analytically. To shed light on the distribution of α_i, we examine the elements of W^{-1}. Let $\frac{1}{n}W = T^T T$ be the Cholesky factorization of $\frac{1}{n}W$, where T is an upper-triangular matrix with positive diagonal elements. We know that the elements t_{ij} $(1 \leq i \leq j \leq n)$ of T are all independent, t_{ii}^2 is $\mathcal{X}_{n-i+1}^2$ $(i = 1, \ldots, n)$, and t_{ij} is $\mathrm{N}(0,1)$ $(1 \leq i < j \leq n)$ (Theorem 3.2.14, *page* 99, [7]). Here $\mathcal{X}_{n-i+1}^2$ represents the chi-square distribution with $n-i+1$ degrees of freedom. Again by symmetry, the diagonal elements of W^{-1} must have identical distributions and so do the off-diagonal elements. Therefore, examining one diagonal and one off-diagonal element is enough to derive the distributions of all elements of W^{-1}. From the following equation:

$$W^{-1} = \frac{1}{n} T^{-1} T^{-T},$$

we have

$$W^{-1}(n,n) = \frac{1}{n \times t_{nn}^2}, \tag{4.16}$$

$$W^{-1}(n,n-1) = \frac{-t_{n-1n}}{n \times t_{nn}^2 \times t_{n-1n-1}}. \tag{4.17}$$

It follows that the following theorem is true:

THEOREM 4.3 *Let X be a $n \times n$ random matrix whose elements are* $\mathrm{N}(0,1)$ *and all independent. Define $W = X^T X$. The diagonal elements of W^{-1} have the same distribution as*

$$\frac{1}{n \times r_1}, \tag{4.18}$$

and the off-diagonal elements have the same distribution as

$$\frac{r_3}{n \times r_1 \times \sqrt{r}_2} = \frac{r_4}{\sqrt{2} \times n \times r_1}, \tag{4.19}$$

where r_i $(i = 1,2,3)$ are all independent, r_1 is $\mathcal{X}_1^2$, r_2 is $\mathcal{X}_2^2$, r_3 is $\mathrm{N}(0,1)$ *and $r_4 = \frac{\sqrt{2} \times r_3}{\sqrt{r_2}}$ is the t-distribution with 2 degrees of freedom which is denoted by t_2. Note that the elements of W^{-1} are not independent.*

Assume that $m < n$. Since

$$\alpha_1^+ = 2 \sum_{j=1}^{m} W^{-1}(1,j) \tag{4.20}$$

and

$$\alpha_1^- = -2 \sum_{j=1}^{m} W^{-1}(m+1,j), \tag{4.21}$$

the difference between them is that α_1^+ has one diagonal element in it which has the same distribution as $\frac{1}{n \times \mathcal{X}_1^2}$ while α_1^- does not. This difference can be visualized by the histogram plots of α_1^+ and α_1^- in Figure 6. The first sub-plot of Figure 6 shows what the distribution function of $\frac{1}{n \times \mathcal{X}_1^2}$ looks like. At $m = 1$, it is obvious that $\Pr(\alpha_1^+ \geq 0) = 1$, and

$$\Pr(\alpha_1^- \geq 0) = \Pr(r_4 \geq 0) = 0.5,$$

because the p.d.f. of a t-distribution is symmetic. At $m = 2$,

$$\Pr(\alpha_1^+ \geq 0) = \Pr(1 - \frac{r_4}{\sqrt{2}} \geq 0) = \Pr(r_4 \leq \sqrt{2}).$$

We know that the p.d.f of t_2 is

$$f(r) = \frac{\sqrt{2}}{4(1+\frac{r^2}{2})^{\frac{3}{2}}}, \tag{4.22}$$

where $-\infty < r < \infty$ (*page* 600, [6]). Integrating it from $-\infty$ to $\sqrt{2}$ gives us

$$\Pr(\alpha_1^+ \geq 0) = \frac{\sqrt{2}}{4} + \frac{1}{2} \approx 0.8536 < 1. \tag{4.23}$$

From Theorem 4.3 and the definitions of α_1^+ and α_1^-, it is easy to see that $\Pr(\alpha_1^+ \geq 0)$ and $\Pr(\alpha_1^- \geq 0)$ do not depend on n. The plots in figures 6 and 7 are consistent with these analytic results.

By the analytic and experimental results obtained above, we give the following conjecture:

CONJECTURE 1 *Assume that there are n random points $x_1, \ldots, x_n$ in R^n, where the coordinates x_{ij} of x_j are $\mathrm{N}(0,1)$ and all independent $(i, j = 1, \ldots, n)$. For the E-separating hyperplanes separating the positive points $\{x_1, \ldots, x_m\}$ and the negative points $\{x_{m+1}, \ldots, x_n, 0\}$, $\Pr(\alpha_1^+ \geq 0)$ decreases monotonically as the number of positive points m increases and it does not depend on n; for any $m < n$, $1 \geq \Pr(\alpha_1^+ \geq 0) > \Pr(\alpha_1^- \geq 0) = 0.5$.*

From the above conjecture, it follows that the Lagrange multipliers associated with the positive points have more chance to be positive than the Lagrange multipliers associated with the negative points under the assumed model.

5 Classification Problems and Support Vector Machines

This section contains background material that may be familiar to some specialists. We have chosen to include this material for the benefit of the many readers who are not already familiar with the underlying problem. Before introducing Support Vector Machines and showing how they relate to the distance problem, we review classification problems by going through an example.

5.1 A Classification Problem and Its Bayesian Solution

Imagine that inside a newly invented Las Vegas machine, there are M generators, each of which can randomly generate k-dimensional vectors x ($x \in \mathcal{I} \subseteq \mathrm{R}^k$) using its own probabilistic model. A color H_m is assigned to each generator for $m = 0, \ldots, M-1$. If x is generated by the mth generator, then we color it with H_m. In the following context we will use color H_m to represent the class of vectors generated by the mth generator. At each time, the machine randomly

turns on one generator and outputs a vector x. The game is to guess the color of x. Since x can be output by any generator, we are not able to tell the right answer all the time. The best thing we can do is to minimize the error.

Given an observation x, a classification problem is to identify the class this x belongs to. Denoted by $\hat{H}(\cdot)$, a deterministic solution to the above problem called decision rule is a function that uniquely maps every k-dimensional vector in $\mathcal{I}$ to one of the M colors, i.e., $\hat{H}$: $\mathcal{I} \rightarrow \{H_0, H_1, ..., H_{M-1}\}$. If the ideal decision rule is $H(\cdot)$, then the function $\hat{H}(\cdot)$ can be considered as an estimation of $H(\cdot)$.

Let us define P_m to be the probability that the machine chooses the mth generator, i.e., $P_m = \Pr[x \in H_m]$, and characterize the probabilistic model underlying each generator by a probability density function $p_m(x)$. If all the statistics P_m and $p_m(x)$ are known, then the optimal decision rule can be derived analytically to minimize the expectation of some cost. For instance, if we define that the cost of the correct answer ($\hat{H}(x) = H(x)$) is zero, and the cost of the wrong answer ($\hat{H}(x) \neq H(x)$) is one, then the expectation of the cost is just the probability of error. Of course for casinos, this probablity of error must be set higher than $\frac{1}{2}$. By Bayesian rule, it can be shown that the optimal solution minimizing the probability of error is in the form of a Likelihood Ratio Test [12]. For the case where all the P_m are equal, the solution is simply

$$\hat{H}(x) = H_{opt} \text{ with } opt = \underset{m}{\text{argmax}} \;\; p_m(x)$$

for $m = 0, \ldots, M-1$.

Each decision rule $\hat{H}(\cdot)$ decomposes the domain $\mathcal{I}$ into M regions:

$$Z_m = \{x |\; \hat{H}(x) = H_m\}.$$

Therefore, geometrically $\hat{H}(\cdot)$ corresponds to the boundary separating these regions Z_m. For example, let us consider the binary case where

$$p_0(x) = \mathrm{N}(m_0, \sigma^2 \mathbf{I}),$$

$$p_1(x) = \mathrm{N}(m_1, \sigma^2 \mathbf{I}),$$

and $P_0 = P_1 = \frac{1}{2}$. The boundary is just the hyperplane orthogonally bisecting the line $\overline{m_1 m_0}$ as shown in Figure 8.

5.2 Support Vector Machine Classifiers

In practice, there are many cases where the statistics P_m and $p_m(x)$ are unknown. The only available information is a finite set of examples $\{(x_i, H_i),\ x_i \in \mathrm{R}^q,\ H_i \in \{H_0, \ldots, H_{M-1}\} \text{ for } i = 1, \ldots, N\}$. Using these examples a SVM classifier estimates the optimal decision rule by finding a boundary layer that

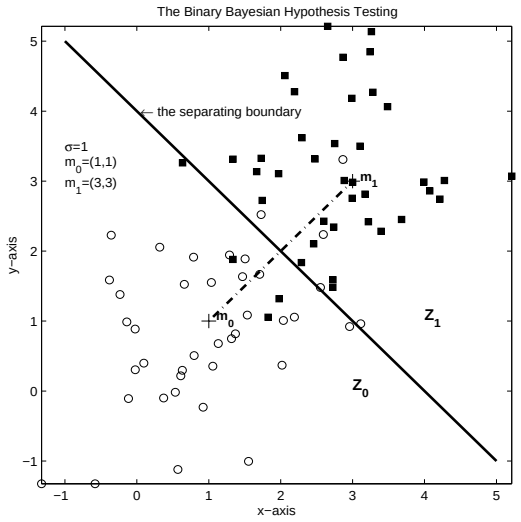


Figure 8: A two dimensional Bayesian separating boundary. The circles represent the instances of color H_0, and the squares represent the instances of color H_1.

separates the M subsets of examples. Without loss of generality, we only consider the binary case. Again, $y_i = \pm 1$ is assigned to each point x_i to represent its class. If the positive and negative points are separable by hyperplanes, then a linear SVM classifier is just the pair of separating hyperplanes that defines the distance between the two sets of points. Therefore, when examples are separable, SVMs can be formulated as the same QP problem as (2.9), and its dual problem is the same as (3.5). As an example, A two dimensional linear SVM classifier is shown in Figure 9. For the inseparable case, [3] could be a good reference.

Nonlinearity is introduced by mapping every point in R^q into a higher dimensional space R^Q where a hyperplane has more degrees of freedom. The map Φ is implicitly defined by a positive definite function $k(\cdot)$ which gives the inner product of two mapped points in R^Q, i.e.,

$$k(x_1, x_2) = \Phi(x_1)^T \Phi(x_2) \text{ where } x_1, x_2 \in R^q.$$

The same problem is computed in R^Q to find the pair of hyperplanes separating the mapped examples $(\Phi(x_i), y_i)$ with the maximum gap. All the key relations are kept same except that the inner product $x_1^T x_2$ is replaced by $k(x_1, x_2)$. Thus, the QP problem (2.9) becomes

$$\underset{w,b}{\text{minimize}} \quad f(w, b) = \frac{1}{2}\|w\|^2 \tag{5.1}$$

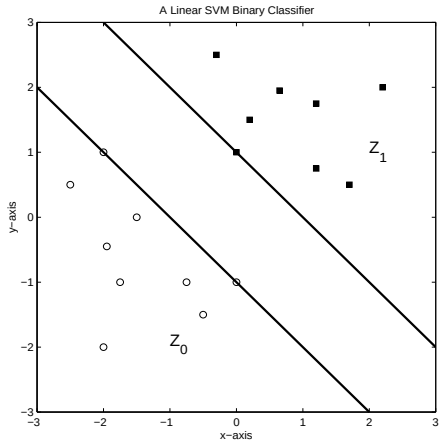


Figure 9: A linear SVM classifier. The circles represent the instances of color H_0, and the squares represent the instances of color H_1.

subject to:

$$y_i(w^T\Phi(x_i)+b) \geq 1, \quad \text{for } i=1,...,N \text{ and } w \in \mathrm{R}^Q. \tag{5.2}$$

The corresponding dual problem is the same as (3.5) except that the Hessian matrix is defined by $H_{ij} = y_i k(x_i, x_j) y_j$. The resulting nonlinear SVM separator is given by

$$\sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b = \pm 1.$$

Since only the inner product $k(.)$ is involved in computing the dual problem, we do not need to know the exact form of Φ.

From the above argument, we see that a nonlinear SVM is equivalent to the optimal separating hyperplanes in a higher dimensional space. Therefore, it is general enough to only consider the linear case. We point out again that KKT conditions $(3.10) - (3.13)$ are the if and only-if conditions for α and b to be the optimal solution. Points x_i are called *support vectors* (SVs) if their Lagrange multipliers α_i are positive. From KKT condition (3.13), SVs must be in the two optimal hyperplanes. In Figure $2(a)$ x_1, x_2 and x_3 are SVs, while in Figure $2(b)$ only x_1 and x_3 are SVs. An important property of SVM classifiers is that discarding the examples corresponding to non SVs will not change the optimal hyperplanes. We will prove it and other properties of SVs in Section 6. Once α and b are computed, the SVM decision rule is the following:

1. if $\sum_{j=1}^{N} \alpha_j y_j x_j^T x + b \leq -1$ then $\hat{H}(x) = H_0$.
2. if $\sum_{j=1}^{N} \alpha_j y_j x_j^T x + b \geq 1$ then $\hat{H}(x) = H_1$.

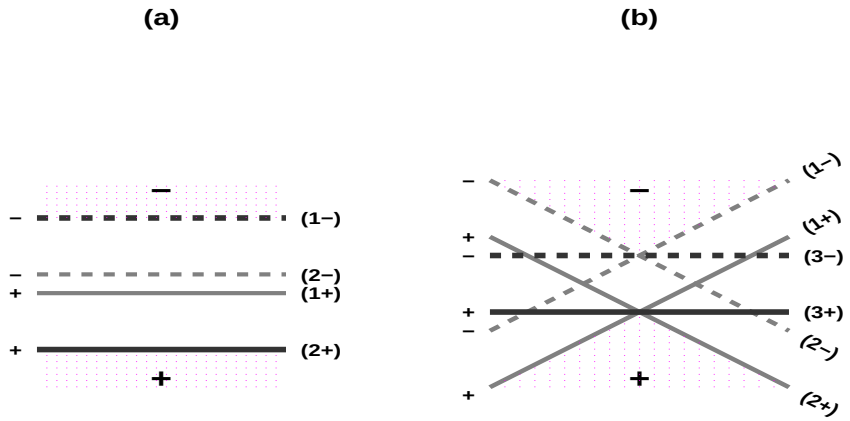


Figure 10: The pair of optimal separating hyperplanes is unique.

3. For the case when x fells in between the two hyperplanes, it is inconclusive so that x can be classified either as H_0 or H_1.

6 Support Vectors and Simplex Volume Decompositions

Recall that in this paper examples are assumed to be separable so that the optimal solution to Problem (2.9) must exist, and that points x_i with positive Lagrange multipliers are called SVs. In this section, we first derive the properties of SVs, then improve the standard definition of SVs so as to insist that they be non-redundant. In order to apply the results obtained in Section 4, we show that the optimal separating hyperplanes are equivalent to the E-separating hyperplanes computed in the subspace determined by SVs.

LEMMA 6.1 *The pair of optimal separating hyperplanes is unique.*

The proof is illustrated geometrically by Figure 10. Suppose that we have two different pairs of optimal hyperplanes (1) and (2). If they are parallel to each other as indicated in Figure 10(a), then it is obvious that the negative hyperplane of (1) and the positive hyperplane of (2) defines a larger gap, which is contradictory to the assumption that (1) and (2) are optimal. Similarly, if they intersect as shown in Figure 10(b), then the pair of hyperplanes (3) is wider, which is a contradiction again. Therefore by contradiction, we can conclude that the pair of optimal separating hyperplanes is unique.

PROPERTY 1 *The pair of optimal separating hyperplanes is independent of points that are not SVs.*

Proof. Given a set of SVs $\{x_{l_1}, \ldots, x_{l_{n+1}}\}$ and their Lagrange multipliers $\alpha^* = [\alpha_{l_1}, \ldots, \alpha_{l_{n+1}}]^T$. Discarding the $N - (n+1)$ points that are not SVs removes the corresponding constraints in the QP problem (2.9):

$$\underset{w,b}{\text{minimize}} \quad f(w,b) = \frac{1}{2}\|w\|^2 \tag{6.1}$$

subject to:

$$y_i(w^T x_i + b) \geq 1, \quad \text{for } i = l_1, ..., l_{n+1}.$$

It follows that the dual problem becomes

$$\underset{\alpha}{\text{minimize}} \quad F(\alpha) \equiv \frac{1}{2}\alpha^T H \alpha - \alpha^T \mathbf{1} \tag{6.2}$$

subject to:

$$\begin{aligned} y^T \alpha &= 0, \\ \alpha &\geq \mathbf{0}, \end{aligned}$$

where $\alpha = [\alpha_1, \ldots, \alpha_{n+1}]^T$ and $H_{i,j} = y_{l_i}(x_{l_i}^T x_{l_j})y_{l_j}$.

In order to prove that solving Problem (6.1) still gives us the same optimal separating hyperplanes, we need to show that α^* is an optimal solution to the above dual problem. Define $I_{sv} = \{l_1, \ldots, l_{n+1}\}$. If α is an optimal solution to Problem (6.2), then by inflating it with zeroes, the resulting $\tilde{\alpha}$ is feasible to the original dual problem (3.5), where

$$\tilde{\alpha}_i = \begin{cases} \alpha_i & \text{if } i \in I_{sv} \\ 0 & \text{otherwise.} \end{cases}$$

With another fact that α^* is defined to be feasible to Problem (6.2), we can conclude that α^* must be an optimal solution to Problem (6.2). By Lemma 6.1, Problem (6.1) (with less constraints) determines the same optimal hyperplanes as Problem (2.9) does. ■

From Property 1 and Lemma 6.1, we know that the optimal separating hyperplanes can be computed by the following equality constrained QP problem if we know which points are SVs:

$$\underset{w}{\text{minimize}} \quad f(w) = \frac{1}{2}\|w\|^2 \tag{6.3}$$

subject to:

$$y_i(w^T x_i - 1) = 1, \quad \text{for } i = l_1, ..., l_n,$$

where $x_{l_{n+1}}$ is assumed to be zero ($b = -1$). Since only equality constraints are involved, the Lagrange multipliers become unconstrained in the corresponding dual problem:

$$\underset{\alpha}{\text{minimize}} \quad F(\alpha) = \frac{1}{2}\alpha^T H \alpha - \alpha^T \begin{bmatrix} \mathbf{2} \\ \mathbf{0} \end{bmatrix}, \tag{6.4}$$

which can be solved by the following linear system if H is nonsingular:

$$H\alpha = \begin{bmatrix} \mathbf{2} \\ \mathbf{0} \end{bmatrix}. \tag{6.5}$$

Since $H\alpha = X^T w$ ($x_{l_{n+1}} = 0$), we have

$$X^T w = \begin{bmatrix} \mathbf{2} \\ \mathbf{0} \end{bmatrix}. \tag{6.6}$$

We recognize that Equation (6.6) is just the vector form of the equality constraints in Problem (6.3), which tells us that the optimal w is given by the least square solution to Equation (6.6).

Before giving the second property of SVs, we define a set of SVs to be non-redundant if they are not degenerate as defined in Proposition 2.1.

PROPERTY 2 *SVs may be redundant.*

Considering the case shown in Figure 2(a), we know that all the three points are SVs. Suppose that another positive point x_4 is given as a new example such that it is on the positive separating line and between x_1 and x_2. By the definition of x_4, we have

$$x_4 = c_1 x_1 + c_2 x_2, \tag{6.7}$$

where $c_1,\ c_2 > 0$ and $c_1 + c_2 = 1$. From the geometry, it is easy to see that adding x_4 will not change the optimal solution. Equation (3.2) and (6.7) together tell us that

$$\begin{aligned} w &= \alpha_1 x_1 + \alpha_2 x_2 \\ &= (\alpha_1 - c_1\alpha_4)x_1 + (\alpha_2 - c_2\alpha_4)x_2 + \alpha_4 x_4 \\ &= \tilde{\alpha_1} x_1 + \tilde{\alpha_2} x_2 + \alpha_4 x_4. \end{aligned} \tag{6.8}$$

Therefore, as long as $0 < \alpha_4 < \min\{\frac{\alpha_1}{c_1}, \frac{\alpha_2}{c_2}\}$, by KKT conditions (3.10) – (3.13) we know that $\alpha = [\tilde{\alpha_1}, \tilde{\alpha_2}, \tilde{\alpha_1} + \tilde{\alpha_2} + \alpha_4, \alpha_4]^T$ is an optimal solution to the corresponding dual problem. Since $\alpha_4 > 0$, x_4 is a SV like others. But we know that it linearly depends on x_1 and x_2. Note that $\alpha = [\alpha_1, \alpha_2, \alpha_1 + \alpha_2, 0]^T$ is also an optimal solution.

REMARK 6.1 *The optimal solution to the dual problem* (3.5) *may not be unique, but the primal problem* (2.9) *must have a unique solution.*

PROPERTY 3 *There exists such a set of SVs that are not redundant.*

Proof. Again, let $\{x_{l_1}, \ldots, x_{l_{n+1}}\}$ be a set of SVs. Equation (4.4) tells us that

$$X\beta = \frac{2}{\|h\|^2}h,$$

where h and β are the same as defined in Section 4, and $X = [x_{l_1}, \ldots, x_{l_n}]$ is a $q \times n$ matrix (we assume that $x_{l_{n+1}} = 0$). Let us redefine the negative points by $x_{l_j} = -x_{l_j}$ so that the above equation can be written as

$$\sum_{i=1}^{n} x_{l_i}\alpha_{l_i} = \frac{2}{\|h\|^2}h. \tag{6.9}$$

Since every negative point x_{l_j} has the following decomposition:

$$x_{l_j} = \tilde{x}_{l_j} - h,$$

where $\tilde{x}_{l_j}$ is the projection of x_{l_j} onto the positive optimal separating hyperplane, Equation (6.9) becomes

$$\begin{aligned}\sum_{l_i \in I_+} x_{l_i}\alpha_{l_i} + \sum_{l_i \in (I_- - \{l_{n+1}\})} \tilde{x}_{l_i}\alpha_{l_i} &= (\frac{2}{\|h\|^2} + \sum_{l_i \in (I_- \{l_{n+1}\})} \alpha_{l_i})h \\ &= c_h h,\end{aligned} \tag{6.10}$$

where $c_h > 0$. Since all the points in the left hand side of Equation (6.10) are in the positive optimal hyperplane and α_{l_i} is positive, h (the projection of $x_{l_{n+1}}$ onto the positive optimal hyperplane) must be inside the polytope constructed by these points. Knowing that among these points the ones that construct the smallest polytope containing h correspond to the smallest basis that can linearly express h, we choose the points x_{l_i} with the same indices as the candidates to be linearly independent SVs, and denote them by $\{x_{c_1}, \ldots, x_{c_{\tilde{n}}}\}$. A three dimensional case is shown in Figure 11. Since h is inside the smallest polytope constructed by the projection of x_{c_i} onto the positive optimal hyperplane, we are guaranteed that the following equation has a positive solution:

$$\sum_{i=1}^{\tilde{n}} x_{c_i}\alpha_{c_i} = \frac{2}{\|h\|^2}h.$$

It follows that the points we have chosen (if $c_i \in I_-$ replace x_{c_i} by $-x_{c_i}$) are still SVs and they are linearly independent. ■

DEFINITION 2 *By* Property 3*, we define SVs to be non-redundant in addition to the requirement that their corresponding Lagrange multipliers be positive.*

LEMMA 6.2 *The optimal separating hyperplanes are equivalent to the E-separating hyperplanes computed in the subspace determined by SVs.*

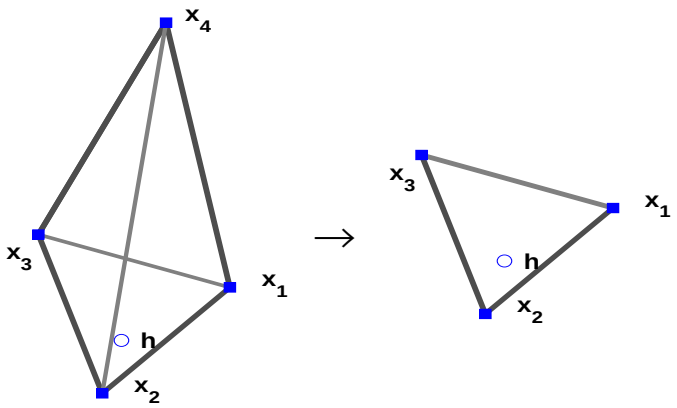


Figure 11: On the positive optimal plane, x_1, x_2 and x_3 define the smallest triangle containing h. It follows that x_4 is redundant and $\{x_1, x_2, x_3, x_5\}$ are SVs.

Proof. Given a set of SVs $\{x_{l_1}, \ldots, x_{l_{n+1}}\}$, we can determine the optimal separating hyperplanes by computing the least square solution to (6.6). Let $X = Q_{q\times n}\hat{X}_{n\times n}$ be the compact QR decomposition of X and $\hat{w} = Q^T w$, where $\hat{X}$ is nonsingular by the definition of SVs. If P is a matrix that completes Q, i.e., $W = [Q, P]$ is square orthogonal, then w has the form:

$$w = Q\hat{w} + P\hat{r},$$

where $\hat{w} \in \mathrm{R}^n$ and $\hat{r} \in \mathrm{R}^{q-n}$. From Equation (6.6), we have

$$\hat{X}^T \hat{w} = \begin{bmatrix} \mathbf{2} \\ \mathbf{0} \end{bmatrix}. \tag{6.11}$$

Since $\hat{X}$ is nonsingular, the above system has a unique solution:

$$\hat{w} = \hat{X}^{-T} \begin{bmatrix} \mathbf{2} \\ \mathbf{0} \end{bmatrix}.$$

It is easy to see that to minimize

$$\|w\|^2 = \|\hat{w}\|^2 + \|\hat{r}\|^2 \text{ for } \hat{r} \in \mathrm{R}^{q-n},$$

$\hat{r}$ must be zero, i.e., the least square solution is $w = Q\hat{w}$. Therefore, once we are given a set of SVs, the original QP problem can be reduced to a linear system (6.11). From results in Section 4, we recognize that this is exactly the linear system that defines the E-separating hyperplanes in R^n that separate the $n+1$ projected SVs $\{\hat{x_{l_1}}, \ldots, \hat{x_{l_n}}, 0\}$. Thus we can conclude that the pair of optimal hyperplanes (in R^q) is equivalent to the pair of E-separating hyperplanes (in R^n) in the sense that $w = Q\hat{w}$ and $\hat{w} = Q^T w$. ■

From Lemma 6.2 and the results about the E-separating hyperplanes in Section 4, it immediately follows that the following theorem is true:

THEOREM 6.1 *Given $n+1$ SVs $\{x_1, \ldots, x_{n+1}\}$ in R^q. Define $X = [x_1 - x_{n+1}, \ldots, x_n - x_{n+1}]$, $I_+ = \{1, \ldots, m\}$ and $I_- = \{m+1, \ldots, n+1\}$. If $X = Q_{q\times n}\hat{X}_{n\times n}$ is the compact QR decomposition of X, then the Lagrange multiplier associated with each SV is determined by*

$$\alpha_l = \frac{2}{\|h\|^2}\frac{\text{vol } \hat{X}_l}{\text{vol } \hat{X}} \tag{6.12}$$

for $l = 1, \ldots, n$, and

$$\alpha_{n+1} = \sum_{i\in I_+}\alpha_i - \sum_{i\in I_- - \{n+1\}}\alpha_i. \tag{6.13}$$

Where $\hat{X}_l = [\hat{x}_1, \ldots, \hat{x}_{l-1}, h, \hat{x}_{l+1}, \ldots, \hat{x}_n]$, and Qh defines the normal direction of the optimal separating hyperplanes and its length gives the distance between them. The pair of optimal separating hyperplanes is given by

$$(Q\hat{X}^{-T}\begin{bmatrix}\mathbf{2}\\ \mathbf{0}\end{bmatrix})^T x + b = \pm 1, \tag{6.14}$$

where b is determined by (3.14). Moreover, the following simplex volume decomposition relation must be satisfied at optimality:

$$\begin{aligned}\text{vol } \hat{X} &= \sum_{i\in I_+}\text{vol } \hat{X}_i,\\ \text{vol } \hat{X}^+ &= \sum_{i\in I_-}\text{vol } \hat{X}_i^+,\end{aligned} \tag{6.15}$$

where the superscript "+" indicates that a positive point is chosen as the origin in the subspace determined by SVs.

REMARK 6.2 *The above argument shows that the Lagrange multiplier associated with each SV can be interpreted geometrically as a normalized ratio of simplex volumes, and at optimality a simplex volume decomposition relation must be satisfied.*

7 A Trigonometric Interpretation of α_l

If we inflate (or deflate) the geometry, i.e., each point is multiplied by a factor c such that $x_i = cx_i$, then from Formula (4.7) and Formula (4.8), we know that each α_i will be changed by a factor $\frac{1}{c^2}$. Since the volumes of X_i and X inflate (or deflate) in the same way, their ratio will be a constant. Define γ_i to be

$$\gamma_i = \frac{\det X_i}{\det X} \quad (i = 1, \ldots, n),$$

where X and X_i are the same as defined in Section 4. We first show that γ_i depends on two angles. Then by introducing another one, we show that the Lagrange multipliers α_i associated with the E-separating hyperplanes can be expressed in terms of three angles.

We need the following theorem during our derivation. Since its proof is not directly related to the main theme of this paper, we put it in the appendix.

THEOREM 7.1 *Let $X_1 = [x_1, \ldots, x_m]$ and $X_2 = [x_{m+1}, \ldots, x_n]$ be a partition of a general square matrix $X = [x_1, \ldots, x_n]$ with $m > n - m$. We denote the compact QR decompositions of X_1 and X_2 by $X_1 = Q_1^{n\times m} R_1^{m\times m}$ and $X_2 = Q_2^{n\times(n-m)} R_2^{(n-m)\times(n-m)}$ respectively. After defining the determinants of X_1 and X_2 by*

$$\det X_1 = \pm \det R_1$$

and

$$\det X_2 = \pm \det R_2,$$

we have the following decomposition of $\det X$*:*

$$\det X = \det X_1 \det X_2 \prod_{i=1}^{n-m} \sin\theta_i, \tag{7.1}$$

where θ_i are the principal angles between the subspaces spanned by X_1 and X_2.

Let

$$X_i^{(1)} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, h]$$

and

$$X^{(1)} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, x_i]$$

be matrices obtained by permuting the columns of X_i and X respectively. Since $X_i^{(1)}$ and $X^{(1)}$ are obtained by the same permutation and permuting the columns of a matrix may only change the sign of its determinant, γ_i has the same form in terms of $X_i^{(1)}$ and $X^{(1)}$:

$$\gamma_i = \frac{\det X_i^{(1)}}{\det X^{(1)}}.$$

Using the decomposition theorem 7.1 and denoting that

$$X_i^{(2)} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n],$$

we have

$$\begin{aligned} \gamma_i = \frac{\det X_i^{(2)} \det h \sin\theta_i}{\det X_i^{(2)} \det x_i \sin\psi_i} &= \frac{\det h \sin\theta_i}{\det x_i \sin\psi_i} \\ &= \pm \frac{\|h\|}{\|x_i\|} \frac{\sin\theta_i}{\sin\psi_i}, \end{aligned} \tag{7.2}$$

where θ_i is the principle angle between the subspaces spanned by $X_i^{(2)}$ and h, and ψ_i is the principle angle between the subspaces spanned by $X_i^{(2)}$ and x_i. If p_i is a vector perpendicular to the $n-1$ dimensional subspace spanned by $X_i^{(2)}$ and we redefine θ_i to be the angle between p_i and h, and ψ_i to be the angle between p_i and x_i, then the formula becomes

$$\gamma_i = \pm \frac{\|h\|}{\|x_i\|} \frac{\cos\theta_i}{\cos\psi_i}. \tag{7.3}$$

It is not hard to verify the above formula from Equation (4.4). Define p_i such that

$$X^T p_i = e_i. \tag{7.4}$$

Multiplying both sides of Equation (4.4) with p_i^T gives us:

$$\begin{aligned} p_i^T X\beta &= \frac{2}{\|h\|^2} p_i^T h \\ \beta_i &= \frac{2}{\|h\|^2} p_i^T h \\ &= \frac{2}{\|h\|^2} \|p_i\| \|h\| \cos\theta_i. \end{aligned} \tag{7.5}$$

It follows that

$$\gamma_i = \|p_i\| \|h\| \cos\theta_i. \tag{7.6}$$

By the definition of p_i, we know that

$$\|p_i\| = \frac{1}{\cos\psi_i \|x_i\|}.$$

Substituting $\|p_i\|$ into (7.6) with the above result, we have

$$\gamma_i = \frac{\|h\| \cos\theta_i}{\|x_i\| \cos\psi_i}. \tag{7.7}$$

REMARK 7.1 *The definition of* p_i (7.4) *removes the* $\pm$ *sign in Formula* (7.2).

From the above results, we can express α_i in terms of $\cos\theta_i$ and $\cos\psi_i$ as the following for $i = 1, \ldots, n$:

$$\begin{aligned} \alpha_i &= \frac{2y_i}{\|h\|^2} \frac{\|h\|}{\|x_i\|} \frac{\cos\theta_i}{\cos\psi_i} \\ &= \frac{y_i}{\frac{1}{2}\|h\|\|x_i\|} \frac{\cos\theta_i}{\cos\psi_i}. \end{aligned} \tag{7.8}$$

Let ϕ_i be the angle between p_i and x_i such that

$$\text{Area}(\triangle x_i x_{n+1} h) = \frac{1}{2}\|h\|\|x_i\| \sin\phi_i. \tag{7.9}$$

Then we have the following formula for α_i:

$$\alpha_i = \frac{y_i}{\text{Area}(\triangle x_i x_{n+1} h)} \frac{\sin\phi_i \cos\theta_i}{\cos\psi_i}. \tag{7.10}$$

It follows that the sign of α_i is determined by $y_i \frac{\cos\theta_i}{\cos\psi_i}$. Therefore, for the corresponding E-separating hyperplanes to be optimal $y_i \frac{\cos\theta_i}{\cos\psi_i}$ must be nonnegative.

The above argument has proved the following theorem:

THEOREM 7.2 *Define p_l such that $\hat{X}^T p_l = e_l$. Then the formula* (6.12) *in Theorem* 6.1 *can be written in terms of three angles ϕ_l, θ_l and ψ_l:*

$$\alpha_l = \frac{y_l}{\text{Area}(\triangle \hat{x}_l \hat{x}_{n+1} h)} \frac{\sin\phi_l \cos\theta_l}{\cos\psi_l}. \tag{7.11}$$

where θ_l is the angle between p_l and h, ψ_l is the angle between p_l and $\hat{x}_l$, and ϕ_l is the angle between $\hat{x}_l$ and h.

Theorem 7.2 shows that α_l can be decomposed into two terms. One term containing three angles is independent of the scale of the coordinates, while the other one is reciprocal to the square of the scale.

As an example, we consider the acute triangle case shown in Figure 2(a). Since points x_1 and x_2 are positive, we have

$$\cos\phi_i = \frac{\|h\|}{\|x_i\|} \quad (i = 1, 2).$$

Thus α_i can be expressed in terms of the following angles:

$$\begin{aligned}
\alpha_1 &= \frac{\sin\angle x_1 x_3 h}{\text{Area}(\triangle x_1 x_3 h)} \times \frac{\cos\angle x_3 x_2 x_1}{\sin\angle x_1 x_3 x_2} \\
&= \frac{2}{h^2} \times \frac{\cos\angle x_1 x_3 h \cos\angle x_3 x_2 x_1}{\sin\angle x_1 x_3 x_2} = \frac{2}{h^2} \times \frac{\cos\angle x_1 x_3 h \sin\angle x_2 x_3 h}{\sin\angle x_1 x_3 x_2} \\
\alpha_2 &= \frac{\sin\angle x_2 x_3 h}{\text{Area}(\triangle x_2 x_3 h)} \times \frac{\cos\angle x_3 x_1 x_2}{\sin\angle x_1 x_3 x_2} \\
&= \frac{2}{h^2} \times \frac{\cos\angle x_2 x_3 h \cos\angle x_3 x_1 x_2}{\sin\angle x_1 x_3 x_2} = \frac{2}{h^2} \times \frac{\sin\angle x_3 x_1 x_3 h \cos\angle x_2 x_3 h}{\sin\angle x_1 x_3 x_2} \\
\alpha_3 &= \alpha_1 + \alpha_2 = \frac{2}{h^2}.
\end{aligned} \tag{7.12}$$

8 Conclusions and Future Work

Knowing which ones are Support Vectors is equivalent to compressing the corresponding Quadratic Programming problem into a smaller linear system. In this paper, we have shown how the optimal solution to the SVM dual problem depends on the geometry of SVs in terms of both simplex volumes and angles, and that SVs must satisfy a simplex volume decomposition relation. Following Section 4.3, we are interested in deriving more statistical properties of the Lagrange Multipliers as future work.

Acknowledgment

We want to thank David Gorsich of U.S. Army TARDEC (Tank-Automotive Research, Development and Engineering Center) for bringing this problem to our attention, giving valuable advice, and providing financial support to this research. During the writing of this paper, discussions with Lizhao Zhang and Ryan M. Rifkin have been very helpful for us. We also want to thank Prof. Tomaso Poggio for his lectures on SVM at Massachusetts Institute of Technology.

A The Proof of Theorem 7.1

Let $X_1 = [x_1, \ldots, x_m]$ and $X_2 = [x_{m+1}, \ldots, x_n]$ be a partition of a general square matrix $X = [x_1, \ldots, x_n]$ with $m > n - m$. We denote the compact QR decompositions of X_1 and X_2 by $X_1 = Q_1^{n\times m} R_1^{m\times m}$ and $X_2 = Q_2^{n\times(n-m)} R_2^{(n-m)\times(n-m)}$ respectively. If $Q_3^{n\times(n-m)}$ and $Q_4^{n\times m}$ are rectangular matrices that complete Q_1 and Q_2, i.e., $P = [Q_1, Q_3]$ and $T = [Q_2, Q_4]$ are square orthogonal, then we have

$$P^T X = \begin{bmatrix} R_1 & Q_1^T Q_2 R_2 \\ \mathbf{0} & Q_3^T Q_2 R_2 \end{bmatrix}.$$

It follows that

$$\det X = \pm \det P^T X = \pm \det R_1 \det R_2 \det Q_3^T Q_2.$$

Let us consider the matrix $W = \begin{bmatrix} Q_1^T Q_2 \\ Q_3^T Q_2 \end{bmatrix}$. It is easy to show that the columns of W are orthonormal. By the CS decomposition theorem (*page* 77, [5]), we know that there exist orthogonal matrices $U_1 \in \mathrm{R}^{m\times m}$, $U_2 \in \mathrm{R}^{(n-m)\times(n-m)}$ and $V_1 \in \mathrm{R}^{(n-m)\times(n-m)}$ such that

$$\begin{bmatrix} U_1 & \mathbf{0} \\ \mathbf{0} & U_2 \end{bmatrix}^T \begin{bmatrix} Q_1^T Q_2 \\ Q_3^T Q_2 \end{bmatrix} V_1 = \begin{bmatrix} C \\ S \end{bmatrix}, \tag{A.1}$$

where

$$C = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_{n-m}),$$

$$S = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_{n-m}),$$

and θ_i are the principal angles between subspaces spanned by X_1 and X_2 (*page* 603, [5]). Therefore,

$$\det Q_3^T Q_2 = \pm \det U_2 Q_3^T Q_2 V_1 = \pm \det S = \pm \prod_{i=1}^{n-m} \sin\theta_i, \tag{A.2}$$

and

$$\det X = \pm \det R_1 \det R_2 \prod_{i=1}^{n-m} \sin\theta_i. \tag{A.3}$$

Define the determinants of X_1 and X_2 respectively by

$$\det X_1 = \pm \det R_1$$

and

$$\det X_2 = \pm \det R_2,$$

so that we have the following decomposition of $\det X$:

$$\det X = \det X_1 \det X_2 \prod_{i=1}^{n-m} \sin\theta_i. \tag{A.4}$$

References

[1] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, 1995.

[2] John C. G. Boot. *Quadratic Programming.* Rand McNally & Company, Chicago, 1964.

[3] Christopher J.C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition.* Knowledge Discovery and Data Mining, 2(2), 1998.

[4] Theodoros Evgeniou, Massimiliano Pontil, Tomaso Poggio. *A Unified Framework for Regularization Networks and Support Vector Machines.* CBCL Paper No. 171, MIT, 1999.

[5] Gene H. Golub, Charles F. Van Loan. *Matrix Computation.* The Johns Hopkins University Press, Baltimore and London, 1996.

[6] Robert V. Hogg, Elliot A. Tanis. *Probability and Statistical Inference.* Prentice-Hall, Inc, New Jersey, 1997.

[7] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory.* John Wiley & Sons, Inc, 1982.

[8] Bernhard Scholkopf, Sebastian Mika, Chris J.C. Bergurges. *Input Space Vs. Feature Space in Kernel-based Methods. IEEE Transactions on Neural Networks*, 1999.

[9] Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data* (in Russian). Nauka, Moscow, 1979.
(English translation: Springer Verlag, New York, 1982.)

[10] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

[11] Vladimir N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc, 1998.

[12] Alan S. Willsky, Gregory W. Wornell, Jeffrey H. Shapiro. *Stochastic Processes, Detection and Estimation.* 6.432 Course Notes, MIT, Fall 1998.