MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1612 August, 1997
C.B.C.L Paper No. 152

Properties of Support Vector Machines

Massimiliano Pontil and Alessandro Verri

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1612.ps.Z

Abstract

Support Vector Machines (SVMs) perform pattern recognition between two point classes by finding a decision surface determined by certain points of the training set, termed *Support Vectors* (SV). This surface, which in some feature space of possibly infinite dimension can be regarded as a hyperplane, is obtained from the solution of a problem of quadratic programming that depends on a regularization parameter. In this paper we study some mathematical properties of support vectors and show that the decision surface can be written as the sum of two orthogonal terms, the first depending only on the *margin vectors* (which are SVs lying on the margin), the second proportional to the regularization parameter. For almost all values of the parameter, this enables us to predict how the decision surface varies for small parameter changes. In the special but important case of feature space of finite dimension m, we also show that there are at most $m+1$ margin vectors and observe that $m+1$ SVs are usually sufficient to fully determine the decision surface. For relatively small m this latter result leads to a consistent reduction of the SV number.

This report describes research done within the Department of Physics, University of Genoa and in collaboration with the Center for Biological and Computational Learning at MIT.

1 Introduction

Support Vector Machines (SVMs) have been recently introduced as a new technique for solving pattern recognition problems [Cortes and Vapnik 1995, Blanz *et al.* 1996, Scholkopf *et al.* 1996, Osuna, Freund and Girosi 1997]. According to the theory od SVMs [Vapnik 1982, Vapnik 1995], while traditional techniques for pattern recognition are based on the minimization of the *empirical risk* – that is, on the attempt to optimize the performance on the training set –, SVMs minimize the *structural risk* – that is, the probability of misclassifying yet-to-be-seen patterns for a fixed but unknown probability distribution of the data. This new induction principle, which is equivalent to minimize an upper bound on the generalization error, relies on the theory of uniform convergence in probability [Vapnik 1982]. What makes SVMs attractive is *(a)* the ability to condense the information contained in the training set, and *(b)* the use of families of decision surfaces of relatively low VC-dimension [Vapnik and Chervonenkis 1971].
In the linear, separable case the key idea of a SVM can be explained in plain words. Given a training set S which contains points of either of two classes, a SVM separates the classes through a hyperplane determined by certain points of S, termed *support vectors*. In the separable case, this hyperplane maximizes the *margin*, or twice the minimum distance of either class from the hyperplane, and all the support vectors lie at the same minimum distance from the hyperplane (and are thus termed *margin vectors*). In real cases, the two classes may not be separable and both the hyperplane and the support vectors are obtained from the solution of a problem of constrained optimization. The solution is a trade-off between the largest margin and the lowest number of errors, trade-off controlled by a regularization parameter.
The aim of this paper is to gain a better understanding of the nature of support vectors, and how the regularization parameter determines the decision surface, in both the linear and nonlinear case. We thus investigate some mathematical properties of support vectors and characterize the dependence of the decision surface on the changes of the regularization parameter. The analysis is first carried out in the simpler linear case and then extended to include nonlinear decision surfaces.
The paper is organized as follows. We first review the theory of SVMs in section 2 and then present our analysis in section 3. Finally, we summarize the conclusions of our work in section 4.

2 Theoretical overview

In this section we recall the basics of the theory of SVM [Vapnik 1995, Cortes and Vapnik 1995] in both the linear and nonlinear case. We start with the simple case of linearly separable sets.

2.1 Optimal separating hyperplane

In what follows we assume we are given a set S of points $\mathbf{x}_i \in \mathbb{R}^n$ with $i = 1, 2, \ldots, N$. Each point $\mathbf{x}_i$ belongs to either of two classes and thus is given a label $y_i \in \{-1, 1\}$. The goal is to establish the equation of a hyperplane that divides S leaving all the points of the same class on the same side while maximizing the minimum distance between either of the two classes and the hyperplane. To this purpose we need some preliminary definitions.

Definition 1. The set S is *linearly separable* if there exist $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$\begin{array}{ll} \mathbf{w} \cdot \mathbf{x}_i + b \geq 1 & \text{if } y_i = 1, \\ \mathbf{w} \cdot \mathbf{x}_i + b \leq -1 & \text{if } y_i = -1. \end{array} \tag{1}$$

In more compact notation, the two inequalities (1) can be rewritten

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \tag{2}$$

for $i = 1, 2, \ldots, N$. The pair $(\mathbf{w}, b)$ defines a hyperplane of equation

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

named *separating hyperplane* (see figure 1(a)). If we denote with w the norm of $\mathbf{w}$, the signed distance d_i of a point $\mathbf{x}_i$ from the separating hyperplane $(\mathbf{w}, b)$ is given by

$$d_i = \frac{\mathbf{w} \cdot \mathbf{x}_i + b}{w}. \tag{3}$$

Combining inequality (2) and equation (3), for all $x_i \in S$ we have

$$y_i d_i \geq \frac{1}{w}. \tag{4}$$

Therefore, $1/w$ is the lower bound on the distance between the points $\mathbf{x}_i$ and the separating hyperplane $(\mathbf{w}, b)$.

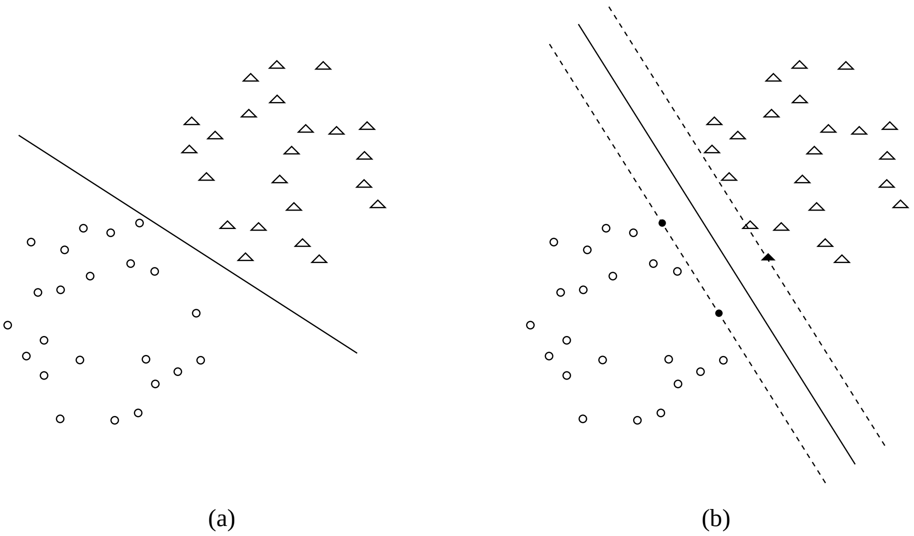

Figure 1: Separating hyperplane and optimal separating hyperplane. Both solid lines in (a) and (b) separate the two identical sets of open circles and triangles, but the solid line in (b) leaves the closest points (the filled circles and triangle) at the maximum distance. The dashed lines in (b) identify the margin.

One might ask why not simply rewrite inequality (2) as

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 0.$$

The purpose of the "1" in the right hand side of inequality (2) is to establish a one-to-one correspondence between separating hyperplanes and their parametric representation. This is done through the notion of canonical representation of a separating hyperplane[1].

Definition 2. Given a separating hyperplane $(\mathbf{w}, b)$ for the linearly separable set S, the *canonical representation* of the separating hyperplane is obtained by rescaling the pair $(\mathbf{w}, b)$ into the pair $(\mathbf{w}', b')$ in such a way that the distance of the closest point equals $1/w'$.

Through this definition we have that

$$\min_{\mathbf{x}_i \in S} \{y_i(\mathbf{w}' \cdot \mathbf{x}_i + b')\} = 1.$$

Consequently, for a separating hyperplane in the canonical representation, the bound in inequality (4) is tight. In what follows we will assume that a separating hyperplane is always given the canonical representation and thus write $(\mathbf{w}, b)$ instead of $(\mathbf{w}', b')$. We are now in a position to define the notion of optimal separating hyperplane.

Definition 3. Given a linearly separable set S, the *optimal separating hyperplane* (OSH) is the separating hyperplane which maximizes the distance of the closest point of S.

Since the distance of the closest point equals $1/w$, the OSH can be regarded as the solution of the problem of maximizing $1/w$ subject to the constraint (2), or

Problem **P1**
Minimize $\frac{1}{2}\mathbf{w} \cdot \mathbf{w}$
subject to $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N$

Two comments are in order. First, if the pair $(\mathbf{w}, b)$ solves **P1**, then for at least one $\mathbf{x}_i \in S$ we have $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$. In particular, this implies that the solution of **P1** is always a separating hyperplane in the canonical representation. Second, the parameter b enters in the constraints but not in the function to be minimized.
The quantity $2/w$, which measures the distance between the two classes in the direction of $\mathbf{w}$, is named *margin*. Hence, the OSH can also be seen as a separating hyperplane which maximizes the margin (see figure 1(b)). We now study the properties of the solution of the problem **P1**.

2.2 Support vectors

Problem **P1** can be solved by means of the classical method of Lagrange multipliers [Bazaraa and Shetty 19
If we denote with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ the N nonnegative Lagrange multipliers associated with the constraints (2), the solution to problem **P1** is equivalent to determining the *saddle point* of the function

$$L = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} - \sum_{i=1}^{N} \alpha_i \{y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1\}. \tag{5}$$

with L = L($\mathbf{w}$, b, $\boldsymbol{\alpha}$). At the saddle point, L has a minimum for $\mathbf{w} = \bar{\mathbf{w}}$ and $b = \bar{b}$ and a maximum for $\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}}$, and thus we can write

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} y_i \alpha_i = 0, \tag{6}$$

[1]This intermediate step toward the derivation of optimal separating hyperplanes is slightly different from the derivation originally developed in [Cortes and Vapnik 1995].

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0 \tag{7}$$

with

$$\frac{\partial L}{\partial \mathbf{w}} = (\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \dots, \frac{\partial L}{\partial w_N}).$$

Substituting equations (6) and (7) into the right hand side of (5), we see that problem **P1** reduces to the maximization of the function

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j,$$

subject to the constraint (6) with $\boldsymbol{\alpha} \geq 0$[2]. This new problem is called *dual problem* and can be formulated as

Problem **P2**

Maximize	$-\frac{1}{2}\boldsymbol{\alpha} \cdot D\boldsymbol{\alpha} + \sum \alpha_i$
subject to	$\sum y_i \alpha_i = 0$
	$\boldsymbol{\alpha} \geq 0,$

where both sums are for $i = 1, 2, \dots, N$, and D is an $N \times N$ matrix such that

$$D_{ij} = y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j. \tag{8}$$

As for the pair $(\bar{\mathbf{w}}, \bar{b})$, from equation (7) it follows that

$$\bar{\mathbf{w}} = \sum_{i=1}^{N} \bar{\alpha}_i y_i \mathbf{x}_i, \tag{9}$$

while $\bar{b}$ can be determined from the Kuhn-Tucker conditions

$$\bar{\alpha}_i \left(y_i(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}) - 1 \right) = 0, \qquad i = 1, 2, \dots, N. \tag{10}$$

Note that the only $\bar{\alpha}_i$ that can be nonzero in equation (10) are those for which the constraints (2) are satisfied with the equality sign. The corresponding points $\mathbf{x}_i$, termed *support vectors*, are the points of S closest to the OSH (see figure $1(b)$).
Given a support vector $\mathbf{x}_j$, the parameter $\bar{b}$ can be obtained from the corresponding Kuhn-Tucker condition as

$$\bar{b} = y_j - \bar{\mathbf{w}} \cdot \mathbf{x}_j.$$

The problem of classifying a new data point $\mathbf{x}$ is now simply solved by computing

$$\operatorname{sign}\left(\bar{\mathbf{w}} \cdot \mathbf{x} + \bar{b}\right). \tag{11}$$

In conclusion, the support vectors condense all the information contained in the training set S which is needed to classify new data points.

[2]In what follows $\boldsymbol{\alpha} \geq 0$ means $\alpha_i \geq 0$ for every component α_i of any vector $\boldsymbol{\alpha}$.

2.3 Linearly nonseparable case

If the set S is not linearly separable or one simply ignores whether or not the set S is linearly separable, the problem of searching for an OSH is meaningless (there may be no separating hyperplane to start with). Fortunately, the previous analysis can be generalized by introducing N nonnegative variables $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N)$ such that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, N. \tag{12}$$

If the point $\mathbf{x}_i$ satisfies inequality (2), then ξ_i is null and (12) reduces to (2). Instead, if the point $\mathbf{x}_i$ does not satisfy inequality (2), the term $-\xi_i$ is added to the right hand side of (2) to obtain inequality (12). The generalized OSH is then regarded as the solution to

Problem **P3**

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C \sum \xi_i \\ \text{subject to} & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, 2, \ldots, N \\ & \boldsymbol{\xi} \geq 0. \end{array}$$

The term $C \sum \xi_i$, where the sum is for $i = 1, 2, \ldots, N$, can be thought of as some measure of the amount of misclassification. Note that this term leads to a more robust solution, in the statistical sense, than the intuitively more appealing term $C \sum \xi_i^2$. In other words, the term $C \sum \xi_i$ makes the OSH less sensitive to the presence of outliers in the training set. The parameter C can be regarded as a regularization parameter. The OSH tends to maximize the minimum distance $1/w$ for small C, and minimize the number of misclassified points for large C. For intermediate values of C the solution of problem **P3** trades errors for a larger margin. The behavior of the OSH as a function of C will be studied in detail in the next section.
In analogy with what was done for the separable case, problem **P3** can be transformed into the *dual*

Problem **P4**

$$\begin{array}{lll} \text{Maximize} & -\frac{1}{2}\boldsymbol{\alpha} \cdot D\boldsymbol{\alpha} + \sum \alpha_i & \\ \text{subject to} & \sum y_i \alpha_i = 0 & \\ & 0 \leq \alpha_i \leq C, & i = 1, 2, \ldots, N \end{array}$$

with D the same $N \times N$ matrix of the separable case. Note that the dimension of **P4** is given by the size of the training set, while the dimension of the input space gives the rank of D. From the constraints of problem **P4** it follows that if C is sufficiently large and the set S linearly separable, problem **P4** reduces to **P2**.
As for the pair $(\bar{\mathbf{w}}, \bar{b})$, it is easy to find that

$$\bar{\mathbf{w}} = \sum_{i=1}^{N} \bar{\alpha}_i y_i \mathbf{x}_i,$$

while $\bar{b}$ can again be determined from $\bar{\boldsymbol{\alpha}}$, solution of the dual problem **P4**, and from the new Kuhn-Tucker conditions

$$\bar{\alpha}_i \left(y_i(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}) - 1 + \bar{\xi}_i \right) = 0 \tag{13}$$

$$(C - \bar{\alpha}_i)\bar{\xi}_i = 0 \tag{14}$$

where the $\bar{\xi}_i$ are the values of the ξ_i at the saddle point. Similarly to the separable case, the points $\mathbf{x}_i$ for which $\bar{\alpha}_i > 0$ are termed *support vectors*. The main difference is that here we have to distinguish between the support vectors for which $\bar{\alpha}_i < C$ and those for which $\bar{\alpha}_i = C$. In the first case, from condition (14) it follows that $\bar{\xi}_i = 0$, and hence, from condition (13), that the support vectors lie at a distance $1/\bar{w}$ from the OSH. These support vectors are termed *margin vectors*. The support vectors for which $\bar{\alpha}_i = C$, instead, are misclassified points (if $\xi_i > 1$), points correctly classified but closer than $1/\bar{w}$ from the OSH (if $0 < \xi \leq 1$), or, in some degenerate cases, even points lying on the margin (if $\xi_i = 0$). In any event, we refer to all the support vectors for which $\alpha_i = C$ as *errors*. An example of generalized OSH with the relative margin vectors and errors is shown in figure 2. All the points that are not support vectors are correctly classified and lie outside the margin strip.

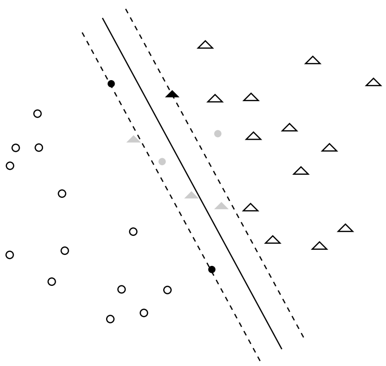

Figure 2: Generalized optimal separating hyperplane. The two sets of circles and triangles are not linearly separable. The solid line is the optimal separating hyperplane, the filled circles and triangles the support vectors (the margin vectors are shown in black, the errors in gray).

We now conclude this section by discussing the extension of the theory to the *nonlinear* case.

2.4 Nonlinear kernels

In most cases,, linear separation in input space is a too restrictive hypothesis to be of practical use. Fortunately, the theory can be extended to nonlinear separating surfaces by mapping the input points into feature points and looking for the OSH in the corresponding feature space [Cortes and Vapnik 1995].
If $\mathbf{x} \in \mathbb{R}^n$ is an input point, we let $\boldsymbol{\varphi}(\mathbf{x})$ be the corresponding feature point with $\boldsymbol{\varphi}$ a mapping from $\mathbb{R}^n$ to a certain space Z (typically a Hilbert space of finite or infinite dimension). In both cases we denote with φ_i the components of $\boldsymbol{\varphi}$. Clearly, to an OSH in Z corresponds a nonlinear separating surface in input space.
At first sight it might seem that this nonlinear surface cannot be determined unless the mapping $\boldsymbol{\varphi}$ is completely known. However, from the formulation of problem **P4** and the classification stage of equation (11), it follows that $\boldsymbol{\varphi}$ enters only in the dot product between feature points, since

$$D_{ij} = y_i y_j \boldsymbol{\varphi}(\mathbf{x}_i) \cdot \boldsymbol{\varphi}(\mathbf{x}_j),$$

and

$$\bar{\mathbf{w}} \cdot \boldsymbol{\varphi}(\mathbf{x}) + \bar{b} = \sum \bar{\alpha}_i y_i \boldsymbol{\varphi}(\mathbf{x}_i) \cdot \boldsymbol{\varphi}(\mathbf{x}) + \bar{b}.$$

Consequently, if we find an expression for the dot product in feature space which uses the points in input space only, that is

$$\boldsymbol{\varphi}(\mathbf{x}_i) \cdot \boldsymbol{\varphi}(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j), \tag{15}$$

full knowledge of $\boldsymbol{\varphi}$ is not necessary. The symmetric function K in equation (15) is called *kernel*. The nonlinear separating surface can be found as the solution of problem **P4** with $D_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, while the classification stage reduces to computing

$$\text{sign}\left(\sum \bar{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x}) + \bar{b}\right).$$

Therefore, the extension of the theory to the nonlinear case is reduced to finding kernels which identify certain families of decision surfaces and can be written as in equation (15). A useful criterion for deciding whether a kernel can be written as in equation (15) is given by Mercer's theorem [Courant and Hilbert 1981, Cortes and Vapnik 1995]: a kernel $\mathrm{K}(\mathbf{x}, \mathbf{y})$, with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, is a dot product in some feature space, or $K(\mathbf{x}, \mathbf{y}) = \boldsymbol{\varphi}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{y})$, if and only if

$$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x}) \quad \text{and} \quad \int\int K(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0, \ \forall f \in L^2.$$

Given such a kernel K, a possible set of functions $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots)$ satisfying equation (15) can be determined from the eigenfunctions $\hat{\varphi}_i$ solution of the eigenvalue problem

$$\int K(\mathbf{x}, \mathbf{y}) \hat{\varphi}_i(\mathbf{x}) d\mathbf{x} = \lambda_i \hat{\varphi}_i(\mathbf{y}), \tag{16}$$

with $\varphi_i = \sqrt{\lambda_i}\hat{\varphi}_i$. If the set of eigenfunctions $\hat{\boldsymbol{\varphi}}$ is finite, the kernel K is said to be *finite* and can be rewritten as

$$K(\mathbf{x}, \mathbf{y}) = \sum \lambda_i \hat{\varphi}_i(\mathbf{x}) \hat{\varphi}_i(\mathbf{y}), \tag{17}$$

where the sum ranges over the set of eigenfunctions. In the general case, the set $\boldsymbol{\varphi}$ is infinite, the kernel is said to be *infinite*, and the sum in equation (17) becomes a series or an integral. We now give two simple examples of kernels. The first is the *polynomial* kernel

$$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^d, \qquad \mathbf{x}, \mathbf{y} \in [-a, a]^d.$$

It can easily be verified that the polynomial kernel satisfies Mercer's theorem and is finite. The separating surface in input space is a polynomial surface of degree d. In this case a mapping $\boldsymbol{\varphi}$ can be determined directly from the definition of K. In the particular case $n = 2$ and $d = 2$, for example, if $\mathbf{x} = (x_1, x_2)$ we can write

$$\boldsymbol{\varphi}(\mathbf{x}) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2\right).$$

The second example is the *Gaussian* kernel

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right),$$

for some $\sigma \in \mathbb{R}$. The Gaussian kernel clearly satisfies Mercer's theorem, but is infinite because equation (16) has a *continuum* of eigenvalues. It is easy to verify that in this case the eigenvalues are given by the normalized Fourier Transform of the Gaussian, $\sqrt{2\pi}\sigma \exp(-\|\mathbf{s}\|^2\sigma^2/2)$, with $\exp(i\mathbf{x} \cdot \mathbf{s})$ as corresponding eigenfunctions. The separating surface in input space is a weighted sum of Gaussians centered on the support vectors.

We are now fully equipped to discuss some mathematical properties of the solution of problem **P4**.

3 Mathematical properties

The goal is to study the dependence of the OSH on the parameter C. We first deal with the linear case and then extend the analysis to nonlinear kernels.

3.1 Lagrange multiplier of a margin vector

We start by establishing a simple but important result on the Lagrange multipliers of the margin vectors. We want to show that *the Lagrange multiplier associated with a margin vector is a step-wise linear function of the regularization parameter C.* To prove it, we need a few preliminary definitions. Since there is no risk of confusion, we now write $\boldsymbol{\alpha}$, b, and $\mathbf{w}$ instead of $\bar{\boldsymbol{\alpha}}$, $\bar{b}$, and $\bar{\mathbf{w}}$. We introduce the sets of support vector indexes

$$I = \{i : 0 < \alpha_i < C\} \quad \text{and} \quad J = \{i : \alpha_i = C\},$$

and let $M+1$ and E be the number of indexes in I and J respectively. The set I identifies the $M+1$ margin vectors, while J the E errors. While E can also be equal to 0, we suppose that there are at least two margin vectors (that is, $M > 0$). This last hypothesis may not be satisfied for highly degenerate configurations of points and small values of C, but does not appear to be restrictive in cases of interest. Finally, and with no further loss of generality, we assume that all the points are support vectors[3] and, hence, that $M + 1 + E = N$.
We start by sorting the support vectors so that

$$I = I^* \bigcup \{N\} \text{ and } J = \{M+1, M+2, \ldots, N-1\},$$

with $I^* = \{1, 2, \ldots, M\}$, and labeling the points so that $y_N = -1$. The Kuhn-Tucker conditions (13) for $i \in I$ tell us that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1. \tag{18}$$

Equation (18), by means of (8) and (9), can be rewritten as

$$\sum_{j=1}^{N} \alpha_j D_{ji} + y_i b = 1. \tag{19}$$

From the equality constraint $\sum y_i \alpha_i = 0$, instead, and since $y_N = -1$ we have

$$\alpha_N = \sum_{i=1}^{N-1} \alpha_i y_i. \tag{20}$$

At the same time, from equation (19) with $i = N$ we get

$$b = \sum_{j=1}^{N} \alpha_j D_{jN} - 1. \tag{21}$$

Plugging equations (20) and (21) into (19) we obtain

$$\sum_{j=1}^{N-1} \alpha_j H_{ji} = 1 + y_i, \qquad i \in I^*. \tag{22}$$

[3]This follows from the fact that if the points with $\alpha_i = 0$ are discarded, problem **P4** has still the same solution.

where H is the $(N-1)\times(N-1)$ matrix

$$H_{ij} = y_i y_j (\mathbf{x}_i - \mathbf{x}_N)\cdot(\mathbf{x}_j - \mathbf{x}_N). \tag{23}$$

Notice that H can be written as

$$H = \begin{pmatrix} H_M & H_{ME} \\ H_{ME}^\top & H_E \end{pmatrix},$$

H_M being the $M\times M$ submatrix between margin vectors, H_E the $E\times E$ submatrix between errors, and H_{ME} the $M\times E$ submatrix between margin vectors and errors. Separating the sum on margin vectors and errors in equation (22), we find:

$$\sum_{j\in I}\alpha_j H_{ji} + C\sum_{j\in J} H_{ji} = 1 + y_i, \qquad i\in I^*. \tag{24}$$

In vector notation equation (24) rewrites

$$H_M\boldsymbol{\alpha}_M + CH_{ME}\mathbf{1}_E = \mathbf{1}_M + \mathbf{y}_M,$$

with $\boldsymbol{\alpha}_M = (\alpha_1,\alpha_2,\ldots,\alpha_M)$, $\mathbf{y}_M = (y_1,y_2,\ldots,y_M)$, and $\mathbf{1}_M$ and $\mathbf{1}_E$ the M- and E-vectors with all the components equal to unit.
Assuming that the matrix H_M is invertible (see the Appendix for a proof of this fact) we have

$$\boldsymbol{\alpha}_M = H_M^{-1}(\mathbf{1}_M + \mathbf{y}_M) - CH_M^{-1}H_{ME}\mathbf{1}_E. \tag{25}$$

From equation (25) we infer that the *Lagrange multiplier associated with a margin vector can always be written as the sum of two terms.* As made clear by the subscript M, the first term depends only on the margin vectors, while the second is proportional to C and depends on both the margin vectors and errors.
An important consequence of the existence of H_M^{-1} is that the vectors $\mathbf{x}_i - \mathbf{x}_N$, $= 1,2,\ldots,M$ are linearly independent. As a corollary, *the number of margin vectors cannot exceed* $n+1$, that is $M\le n$. Notice that this does not mean that the number of points lying on the margin cannot exceed $n+1$. In degenerate cases, there may be points lying on the margin with $\alpha = 0$, or even support vectors lying on the margin with $\alpha = C$.

3.2 Dependence on the regularization parameter

We are now in a position to study the dependence of the OSH on the parameter C. We first show that the normal to the OSH can be written as the sum of two orthogonal vectors.

3.2.1 Orthogonal decomposition

In components equation (25) can be rewritten

$$\alpha_i = r_i + g_i C \qquad i\in I^*, \tag{26}$$

with

$$\mathbf{r}_M = H_M^{-1}(\mathbf{1}_M + \mathbf{y}_M) \tag{27}$$

and

$$\mathbf{g}_M = -H_M^{-1} H_{ME} \mathbf{1}_E. \tag{28}$$

Notice that the r_i and g_i are not necessarily positive (although they cannot be both negative). If we define

$$r_N = \sum_{i \in I^*} r_i y_i \tag{29}$$

$$g_N = \sum_{i \in I^*} g_i y_i + \sum_{i \in J} y_i, \tag{30}$$

then equation (26) is also true for the margin vector of index N as

$$r_N + g_N C = \sum_{i \in I^*} r_i y_i + \sum_{i \in I^*} g_i y_i C + \sum_{i \in J} y_i C = \sum_{i \in I^*} y_i \alpha_i + C \sum_{i \in J} y_i = \alpha_N,$$

where the last equality is due to the constraint (6) and the fact that $\alpha_i = C$ for all $i \in J$. Plugging equation (26) into (9) and separating the constant and linear term we obtain

$$\mathbf{w} = \mathbf{w}_1 + C\mathbf{w}_2, \tag{31}$$

with

$$\mathbf{w}_1 = \sum_{i \in I} r_i y_i \mathbf{x}_i, \tag{32}$$

$$\mathbf{w}_2 = \sum_{i \in J} y_i \mathbf{x}_i + \sum_{i \in I} g_i y_i \mathbf{x}_i. \tag{33}$$

It can easily be seen that $\mathbf{w}_1$ and $\mathbf{w}_2$ are orthogonal. Substituting equations (29) and (30) into (32) and (33) respectively, one obtains

$$\mathbf{w}_1 = \sum_{i \in I^*} r_i y_i (\mathbf{x}_i - \mathbf{x}_N),$$

$$\mathbf{w}_2 = \sum_{i \in J} y_i (\mathbf{x}_i - \mathbf{x}_N) + \sum_{i \in I^*} g_i y_i (\mathbf{x}_i - \mathbf{x}_N).$$

Then, through the definition of H_M and H_{ME} we have

$$\mathbf{w}_1 \cdot \mathbf{w}_2 = \mathbf{r}_M H_{ME} \mathbf{1}_E + \mathbf{r}_M H_M \mathbf{g}_M. \tag{34}$$

Plugging equation (28) in (34) it follows immediately that $\mathbf{w}_1 \cdot \mathbf{w}_2 = 0$.

3.2.2 Changing the regularization parameter

We now study the effect of small changes of the regularization parameter C on the OSH. Since C is the only free parameter of SVMs, this study is relevant from both the theoretical and practical viewpoint. In what follows we let C take on values over the positive real axis $\mathbb{R}^+$. First, we notice that the possible choices of support vectors for all possible values of C (distinguishing between margin vectors and errors) are finite. If we neglect degenerate configurations of support vectors, this implies that $\mathbb{R}^+$ can be partitioned in a finite number of disjoint interval, each characterized by a fixed set of support vectors. Notice that the rightmost interval is necessarily unbounded.

After this preliminary observation we can already conclude that, with the exception of the C values corresponding to the interval ends, the set of support vectors does not vary for small changes of C. But through the previous analysis we can also study the dependence of the normal vector $\mathbf{w}$ on the parameter C. From equation (31) it follows that if C changes by δC and the margin vectors and errors remain the same, the normal vector $\mathbf{w}$ changes by $\delta C \mathbf{w}_2$ along the direction of $\mathbf{w}_2$. We can make this statement more precise distinguishing between two cases.
In the first case we let M reach the maximum value n. Since H_M has always maximum rank, we have $n+1$ independent Kuhn-Tucker conditions like equation (18) and the OSH is completely determined by the $n+1$ margin vectors. Consequently, since for almost all C the set of support vectors remains the same for small changes of C, $\mathbf{w}_2$ must vanish and we have

$$\mathbf{w} = \sum_{i \in I} r_i y_i \mathbf{x}_i. \tag{35}$$

Equation (35) tells us that if $M = n$ the OSH is fixed and unambiguously identified by the $n+1$ margin vectors. The fact that the OSH is fixed makes it possible to determine the maximum interval around C, say $(C_1, C_2]$, in which the OSH is given by equation (35). To this purpose it is sufficient to compute the r_i and g_i from equations (27) and (28) and find C_1 and C_2 as the minimum and maximum C for which the α_i associated with the margin vector $\mathbf{x}_i$ satisfy the constraint $0 < \alpha_i \leq C$.
In the second case, we have $M < n$. The OSH is now given by equation (31) with $\mathbf{w}_2 \neq 0$. Thus for a small change δC the new OSH $\mathbf{w}'$ can be written as

$$\mathbf{w}' = \mathbf{w} + \delta C \mathbf{w}_2. \tag{36}$$

Equation (36) tells us that if $M < n$ the OSH changes of an amount $\delta C \mathbf{w}_2$. Here again there exists a maximum interval $(C_1, C_2]$ around C in which the OSH is given by equation (36). Similarly to the previous case, one could determine the minimum and maximum C for which the α_i associated with the margin vectors satisfy the constraint $0 < \alpha_i \leq C$. However, since to a changing OSH might correspond a new set of support vectors, these minimum and maximum values are only a lower and upper bound for C_1 and C_2 respectively.
Finally, we observe that even if $M < n$, the OSH can always be written as a linear combination of $n+1$ support vectors, for example by adding $n+1-M$ errors.

3.2.3 A numerical example

We now illustrate both cases by means of the numerical example with $n = 2$ shown in figure 3. figure 3(a) shows the OSH found for the displayed training set with $C = 4.0$. The support vectors are denoted by the filled circles and triangles (the margin vectors in black, the errors in grey). In accordance with equation (35), since there are 3 margin vectors the OSH is fixed. Straightforward computations predict that the OSH must remain the same for $2.7 < C \leq 4.5$. This prediction has been verified numerically.
Figure 3(b) shows the new OSH obtained for C just outside the interval $(2.7, 4.5]$ ($C = 4.8$). Notice that the errors are the same of figure 3(a), while there are only two margin vectors. As we have just discussed, the OSH should now change for small variations of C as predicted by equation (36). This has been verified numerically and figure 3(c) displays the OSHs obtained from equation (36) and from direct solution of the problem **P4** for $C = 6.7$. The two OSH coincide within numerical precision.

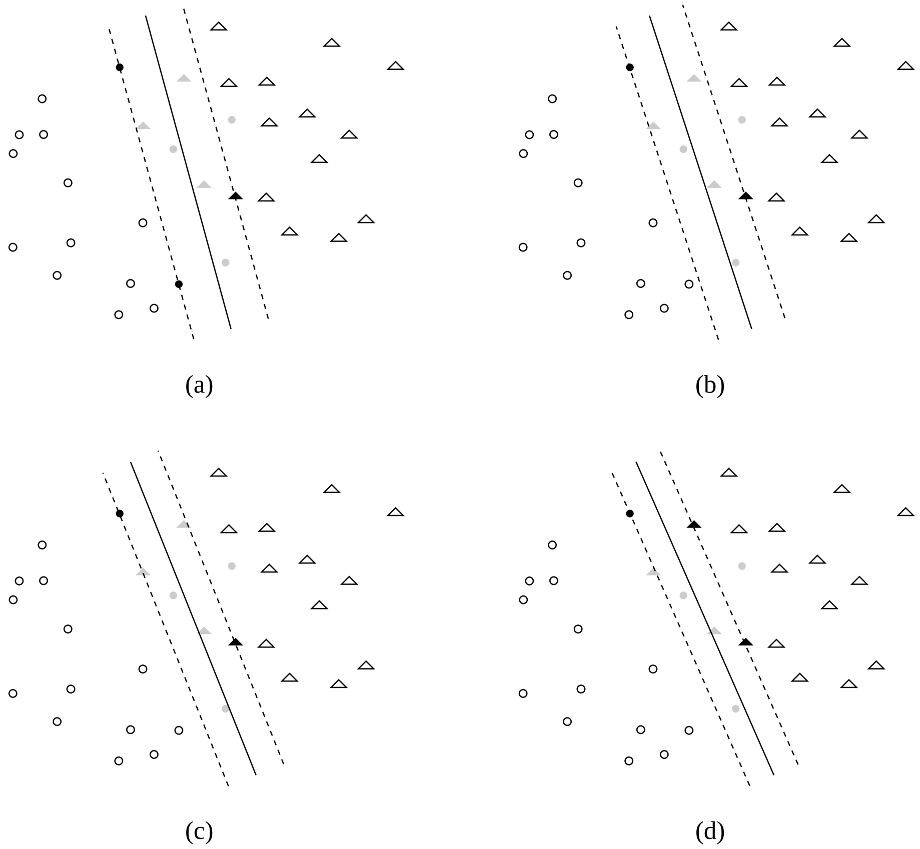

Figure 3: Optimal separating hyperplane for $C = 4.0$ (a), $C = 4.8$ (b), $C = 6.7$ (c), and $C = 7.5$ (d) respectively. Legend as in figure 2.

For a larger variation of C ($C > 7.0$, see figure 3(d)) the number of margin vectors goes back to 3 and the solution is again fixed. Notice that in this last transition one of the errors became a margin vector (the error in the upper part of the margin strip of figure 3(c) is a margin vector in figure 3(d)).
As mentioned in the previous section, it is worthwhile noticing that the solutions with smaller C (see figure 3(a) and (b)) have a larger margin, while the solutions with larger C (see figure 3(c) and (d)) have a smaller number of errors.

3.3 Extension to nonlinear kernels

We now extend the presented analysis to the case of nonlinear kernels.

Lagrange multiplier of a margin vector We start by observing that the same decomposition of the Lagrange multiplier of a margin vector derived in the linear case holds true for nonlinear kernels. Note that the matrix H of equation (23) rewrites

$$H_{ij} = y_i y_j \left(K(\mathbf{x}_i, \mathbf{x}_j) - K(\mathbf{x}_j, \mathbf{x}_N) - K(\mathbf{x}_i, \mathbf{x}_N) + K(\mathbf{x}_N, \mathbf{x}_N) \right), \tag{37}$$

while equations (25) to (30) remain unchanged.

Orthogonal decomposition More care is needed for the extension of the orthogonal decomposition of $\mathbf{w}$ and the study of the behavior of the separating surface as a function of C. This is because, in the nonlinear case, *it may not be possible to recover an explicit expression for* $\mathbf{w}$. However, this does not pose major problems because all the expressions involving $\mathbf{w}$ are effectively dot products between feature points and can be computed by means of the kernel K. Indeed, if we take the dot product between $\mathbf{w}$ and $\varphi(\mathbf{x})$, we obtain

$$\mathbf{w} \cdot \boldsymbol{\varphi}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}),$$

that can be written as

$$\begin{aligned} \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) &= \sum_{i \in I} r_i y_i K(\mathbf{x}_i, \mathbf{x}) \\ &+ C\left(\sum_{j \in J} y_j K(\mathbf{x}_j, \mathbf{x}) + \sum_{i \in I} g_i y_i K(\mathbf{x}_i, \mathbf{x})\right). \end{aligned} \tag{38}$$

The two terms in the r.h.s. of equation (38) are the counterparts of equations (32) and (33) defining $\mathbf{w}_1$ and $\mathbf{w}_2$ respectively. Note that even if the explicit expression for $\mathbf{w}_1$ and $\mathbf{w}_2$ cannot be given, the orthogonality relation (34) remains true. This can be seen from the fact that the r.h.s. of equation (34) depends on the matrix H which, in the nonlinear case, is rewritten as in equation (37). In this respect, the two terms in the r.h.s. of equation (38) can be regarded as orthogonal.

Changing the regularization parameter So far, all the results derived in the linear case carried through the case of nonlinear kernels. For the dependence of the separating surface on the parameter C, instead, it is convenient to distinguish between finite and infinite kernels.
For finite kernels, all the results obtained in the linear case are still valid and can be rederived simply replacing n, dimension of input space, with m, dimension of feature space. For example, if $M = m$, the OSH in feature space does not change for small changes of C and the second term in the r.h.s of equation (38) vanishes for all $\mathbf{x}$. Furthermore, the interval $(C_1, C_2]$, within which the OSH is fixed, can be determined exactly as in the linear case.
For kernels of infinite dimension, instead, a finite number of margin vectors is not sufficient to fully determine the OSH. Consequently and differently from the finite case, the OSH is never fixed and the second term of equation (38) does not vanish. For a small change δC, the dot product $\mathbf{w} \cdot \boldsymbol{\varphi}(\mathbf{x})$ changes of the amount

$$\delta C\left(\sum_{j \in J} y_j K(\mathbf{x}_j, \mathbf{x}) + \sum_{i \in I} g_i y_i K(\mathbf{x}_i, \mathbf{x})\right).$$

In summary, all the results derived in the linear case can be extended without major changes to the nonlinear case, with the exception of the properties depending on the finiteness of the dimension of the linear case, like the upper bound on the number of margin vectors, properties that are still true for finite kernels only.

4 Conclusions

In the case of pattern recognition, SVMs depend only one free parameter, the regularization parameter C. In this paper we have discussed some mathematical properties of support vectors

useful to characterize the behavior of the decision surface with respect to C. We have identified a special subset of support vectors, the margin vectors, whose Lagrange multiplier are strictly smaller than the regularization parameter C. We have shown that the margin vectors are always linearly independent and that the decision surface can be written as the sum of two orthogonal terms, the first depending only on the margin vectors, the second proportional to the regularization parameter. For almost all values of the parameter, this enabled us to predict how the decision surface varies for small parameter changes. In general we found that the solution is usually stable with respect to small changes of C.
The obtained results can be more conveniently summarized distinguishing between finite and infinte kernels. For kernels of finite dimension m, it turned out that $m+1$ is the least upper bound for the number of margin vectors $(M+1)$ and the behavior of the OSH as a function of C depends on whether $M=m$ or $M<m$. If $M=m$, the $M+1$ margin vectors are sufficient to fully determine the equation of the OSH in feature space and for almost all values of C the OSH does not vary for small changes of C. If $M<m$, instead, the OSH varies of an amount proportional to the change δC in a direction identified by both the margin vectors and errors. In both cases it is worthwhile observing that the number of support vectors effectively needed to identify the decision surface is never greater than $m+1$. This latter result may be useful to reduce the number of support vectors effectively needed to perform recognition.
For infinite kernels, the margin vectors are still linearly independent but there is no upper bound on their number. For small changes of C the OSH is not fixed and varies as in the case $M<m$ of finite kernels.

Acknowledgements. Edgar Osuna read the manuscript and made useful remarks. This work has been partially supported by a grant from the Agenzia Spaziale Italiana.

Appendix

In this appendix we sketch the proof of the existence of H_M^{-1}. First, we need to (a) transform the original dual problem **P4** into a Linear Complementary Problem (LCP), and (b) derive the explicit expression for the matrix G which defines the polyhedral set on which the solution of the LCP lies.
Let us define $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{N-1})$ and remind that $\alpha_N = \sum y_i \alpha_i$ where the sum ranges over $i = 1, 2, \ldots, N-1$. We let N_1 and N_2 be the number of points with positive and negative labels respectively. We start by rewriting problem **P4** without the equality constraint as

Problem **P5**

$$\text{Minimize} \qquad \frac{1}{2}\boldsymbol{\alpha} \cdot H\boldsymbol{\alpha} - 2\sum_{i \in I^+} \alpha_i$$

$$\begin{array}{lll} \text{subject to} & -\sum_{i=1}^{N-1} y_i\alpha_i \le 0, \ \sum_{i=1}^{N-1} y_i\alpha_i \le C & \\ & \alpha_i \le C, & i = 1, 2, \ldots, N-1 \\ & \alpha_i \ge 0, & i = 1, 2, \ldots, N-1 \end{array}$$

with I^+ the set of indexes corresponding to the α_i for which $y_i = 1$. Then, we let u_+, u_-, $\mathbf{u} = (u_1, u_2, \ldots, u_{N-1})$, and $\mathbf{v} = (v_1, v_2, \ldots, v_{N-1})$ be the $2N$ Lagrange multipliers associated with

the constraints of problem **P5** respectively.
The LCP associated with problem **P5** is obtained by

1. setting equal to 0 the gradient of the Lagrangian associated with problem **P5**, or

$$\sum_{j=1}^{N-1} \alpha_j H_{ji} - 1 + y_i(u_+ - u_-) - y_i + u_i - v_i = 0,$$

   and

2. introducing the $N+1$ slack variables[4] s_+, s_-, and $\mathbf{s} = (s_1, s_2, \ldots, s_{N-1})$, satisfying

$$s_+ + \sum_{i=1}^{N-1} \alpha_i y_i = 0,$$

$$s_- - \sum_{i=1}^{N-1} \alpha_i y_i = C,$$

   and

$$s_i + \alpha_i = C,$$

   along with the associated complementary conditions

$$s_- u_- = s_+ u_+ = 0,$$

$$s_i u_i = 0,$$

   and

$$\alpha_i v_i = 0,$$

   for each $i = 1, 2, \ldots, N-1$.

The solution of problem **P5** can be obtained as the solution of the LCP

Problem **P6**		
Solve	$\mathbf{t} - M\mathbf{z} = \mathbf{q}$	
subject to	$\mathbf{t}, \mathbf{z} \geq 0$	
	$t_i z_i = 0,$	$i = 1, 2, \ldots, 2N,$

with $\mathbf{t} = (s_-, s_+, \mathbf{s}, \mathbf{v})$, $\mathbf{z} = (u_-, u_+, \mathbf{u}, \boldsymbol{\alpha})$,

$$M = \begin{pmatrix} 0 & -A \\ A^\top & H \end{pmatrix},$$

$$A = \begin{pmatrix} -y_1 & \cdots & -y_{N-1} \\ y_1 & \cdots & y_{N-1} \\ & & \\ & I_{N-1} & \\ & & \end{pmatrix},$$

[4]In the constrained optimization jargon, a slack variable is a nonnegative variable that turns an inequality into an equality constraint.

$\mathbf{q} = (\mathbf{b}, \mathbf{k})$,

$$\mathbf{b} = (0, \overbrace{C, \ldots, C}^{N+1}), \quad \text{and} \quad \mathbf{k} = (\overbrace{-2, \ldots, -2}^{N_1}, \overbrace{0, \ldots, 0}^{N_2 - 1}).$$

Similarly to the case of linear programming, a solution to Problem **P6** is a vertex of a polyhedral set. In addition, the solution must also satisfy the complementarity conditions. In the case of problem **P6**, a solution vector $\mathbf{p} = (\mathbf{t}, \mathbf{z})$ is a vertex of the polyhedral set $S = \{\mathbf{p} : G\mathbf{p} = \mathbf{q}, \mathbf{p} \geq 0\}$, with $G = [I_{2N}, -M]$, $\mathbf{p} = (\mathbf{p}_B, \mathbf{p}_N)$, $\mathbf{p}_B = B^{-1}\mathbf{q}$, $\mathbf{p}_N = 0$, and B is the $2N \times 2N$ matrix defined by the columns of G corresponding to the $2N$ active variables.
Through simple but lengthy calculations, it can be seen that the matrix H_M is a submatrix of B and H_M^{-1} a submatrix of B^{-1}. The existence of H_M^{-1} is thus ensured by the existence of B^{-1}.

References

[Bazaraa and Shetty 1979] Bazaraa, M. and Shetty, C.M. *Nonlinear programming* (John Wiley, New York, 1979).

[Blanz *et al.* 1996] Blanz, V., Scholkopf, B., Bulthoff, H., Burges, C., Vapnik, V.N. and Vetter, T. *Comparison of view–based object recognition algorithms using realistic 3D models.* Proc of ICANN'96. LNCS Vol. 1112, 1996, pp. 251–256.

[Cortes and Vapnik 1995] Cortes, C. and Vapnik, V.N 1995. "Support Vector Network," *Machine learning* **20**: 1-25.

[Courant and Hilbert 1981] Courant, R. and Hilbert, D. *Methods of Mathematical Physics* (John Wiley, New York, 1989).

[Osuna, Freund and Girosi 1997] Osuna, E., Freund, R. and Girosi, F. "Training Support Vector Machines: an Applications to Face Detection." CVPR97.

[Scholkopf *et al.* 1996] Scholkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. and Vapnik. V.N. *Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers.* AI Memo No. 1599; CBCL Paper No. 142, Massachusetts Institute of Technology, Cambridge, 1996.

[Vapnik 1982] Vapnik, V.N. *Estimation of Dependencies Based on Empirical Data* (Springer-Verlag, New York, 1982).

[Vapnik 1995] Vapnik, V.N. *The Nature of Statistical Learning Theory* (Springer-Verlag, New York, 1995).

[Vapnik and Chervonenkis 1971] Vapnik, V.N. and Chervonenkis, A.Ja. 1971. "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab Appl.*, Vol. 16, pp. 264–280.