Manuscript Number: 2115

Formulations of Support Vector Machines: a Note from an Optimization Point of View

Chih-Jen Lin

Department of Computer Science and

Information Engineering

National Taiwan University

Taipei 106, Taiwan (`cjlin@csie.ntu.edu.tw`)

Abstract

In this paper, we discuss issues about formulations of support vector machines (SVM) from an optimization point of view. First, support vector machines map training data into a higher (may be infinite) dimensional space. Currently primal and dual formulations of SVM are derived in the finite dimensional space and readily extended to the infinite dimensional space. We rigorously discuss the primal-dual relation in the infinite dimensional spaces. Second, SVM formulations contain penalty terms which are different from unconstrained penalty functions in optimization. Traditionally unconstrained penalty functions approximate a constrained problem as the penalty parameter increases. We are interested in similar properties for SVM formulations. For two of the most popular SVM formulations, we show that one enjoys properties of exact penalty functions but the other is only like traditional penalty functions which converge when the penalty parameter goes to infinity.

1 Introduction

The support vector machine (SVM) is a new and very promising classification technique. Surveys of SVM are, for example, Vapnik (1995, 1998). Given training vectors $x_i \in R^n, i = 1, \dots, l$ in two classes, and a vector $y \in R^n$ such that $y_i \in \{1, -1\}$, the support vector machine requires the solution of the following quadratic programming problem (Boser et al. (1992)):

$$(P_0) \qquad \min \frac{1}{2}\langle w, w\rangle$$
$$y_i(\langle w, \phi(x_i)\rangle + b) \geq 1, i = 1, \dots, l,$$

where $\langle \cdot, \cdot \rangle$ is the inner product of two finite or infinite vectors.

Training vectors x_i are mapped into a higher dimensional space by the function ϕ. In this new space we seek for a separating hyperplane $\langle w, \phi(x) \rangle + b = 0$ with coefficients w and b such that all training vectors are separated. Note that this higher dimensional space can be finite or *infinite*. In this paper, to stress the difference between finite and infinite inner products, we represent them as $x^T y$ and $\langle x, y \rangle$, respectively.

If in a higher dimensional space data are still not separable, Cortes and Vapnik (1995) proposed to add a penalty term $C \sum_{i=1}^{l} \xi_i$ in the objection function, where C is a large positive number:

$$(P_C) \qquad \begin{aligned} &\min \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^{l} \xi_i \\ &y_i(\langle w, \phi(x_i) \rangle + b) \geq 1 - \xi_i, \\ &\xi_i \geq 0, i = 1, \ldots, l. \end{aligned}$$

The technique of introducing variables ξ_i can be traced back to, for example, Kennete and Mangasarian (1992) for linear programming formulations.

As (P_C) becomes a problem with many (or infinite) variables, currently the standard method (see, for example, Cortes and Vapnik (1995), and Vapnik (1995, 1998)) to solve (P_C) is through the following dual which is a finite problem with l variables:

$$(D_C) \qquad \begin{aligned} &\min \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ &y^T \alpha = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, l, \end{aligned}$$

where $Q_{ij} \equiv y_i y_j \langle \phi(x_i), \phi(x_j) \rangle$ and e is the vector of all ones. If α is a solution of (D_C), $w \equiv \sum_{i=1}^{l} y_i \alpha_i \phi(x_i)$ is a solution of (P_C). Usually $K(x_i, x_j) \equiv \langle \phi(x_i), \phi(x_j) \rangle$ is called the kernel. Though $\phi(x)$ is in a very high dimensional space, if a special ϕ is used, $K(x_i, x_j)$ is finite and can be easily calculated. Algorithms for solving (D_C) can be seen in, for example, papers in the collection by Schölkopf et al. (1998). Similarly, the dual of (P_0) is as follows:

$$(D_0) \qquad \begin{aligned} &\min \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ &y^T \alpha = 0, \alpha_i \geq 0, i = 1, \ldots, l. \end{aligned}$$

However, in the above references, the derivation of (D_C) was through the standard convex analysis in finite dimensional spaces. Note that $\phi(x)$ and w can be vectors in an *infinite dimensional Hilbert space*. Since properties of convex analysis in finite spaces may not always hold in infinite spaces, we must be careful on any generalization. In this paper, all results are derived for infinite dimensional formulations.

The first issue is on the relation between (P_C) and $(D_C), C \geq 0$. For example, we may ask whether it might happen that (P_0) has an optimal solution but (D_0) is not solvable. Note that it is possible that a convex programming problem has a lower bound on all objective values but has no feasible points achieving this bound. In Section 2, we show that (P_0) and (D_0) are either both solvable, or (P_0) is infeasible and (D_0) is unbounded. In addition, (P_C) and (D_C), $C > 0$ are both solvable.

Originally by Cortes and Vapnik (1995), and recently by, for example, Friess et al. (1998), a different SVM formulation is as follows:

$$(\bar{P}_C) \quad \begin{aligned} &\min \frac{1}{2}\langle w, w\rangle + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \\ &y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i, \\ &i = 1, \dots, l. \end{aligned} \qquad (\bar{D}_C) \quad \begin{aligned} &\min \frac{1}{2}\alpha^T(Q + \frac{I}{C})\alpha - e^T\alpha \\ &y^T\alpha = 0, \alpha_i \geq 0, \\ &i = 1, \dots, l. \end{aligned}$$

Here I is the identity matrix. Using similar techniques, we show that both $(\bar{P}_C)$ and $(\bar{D}_C)$, $C > 0$ are always solvable.

The second topic of this paper is on the penalty parameter. In (P_C) (or $(\bar{P}_C)$), the penalty term $C\sum_{i=1}^{l}\xi_i$ (or $\frac{C}{2}\sum_{i=1}^{l}\xi_i^2$) is added to reduce the occurrence of wrongly classified data. Note that such an approach is different from traditional penalty functions which modify constrained problems to unconstrained ones. As the penalty parameter goes to infinity, the solutions of this sequence of unconstrained problems approximate a solution of the original problem. Here we are interested in similar properties for SVM formulations. In Section 3, we show that for (P_C), after C is greater than a certain number, all solutions of (P_C) are a solution of (P_0). In other words, (P_C) is like an *exact penalty function* approach. On the other hand, this result may not hold for $(\bar{P}_C)$ but as $C \to \infty$, solutions of $(\bar{P}_C)$ still converge to a solution of (P_0).

We hope that this note demonstrates a framework on optimization properties of SVM formulations. Hence when there are new formulations proposed, similar tools and techniques can be applied without many modifications. Finally we make conclusions and discussions in Section 4.

2 Primal-Dual Relation of SVM Formulations

For the discussion in the rest of this paper, first we need the following theorem which shows the necessary and sufficient optimality conditions of a convex optimization problem. This is an extension of the KKT condition in finite dimensional spaces. Note that the standard necessary condition usually requires some constraint qualifications (e.g. Slater condition). However, for problems with linear constraints, such an assumption is not necessary.

Theorem 2.1 *Consider an optimization problem*

$$\min f(x) \text{ s.t. } \langle v_i, x\rangle - b_i \geq 0, i = 1, \dots, l,$$

where $x \in X$, *a Hilbert space,* $v_i \in X, i = 1, \dots, l$ *and* $b_i \in R, i = 1, \dots, l$. *If* f *is a real convex and differentiable function, then a feasible element* x *is an optimal solution if and only if there exist real numbers* $\lambda_1, \dots, \lambda_l$ *such that*

$$\nabla f(x) - \lambda_1 v_1 - \cdots - \lambda_l v_l = 0$$
$$\lambda_i \geq 0, \lambda_i(\langle v_i, x\rangle - b_i) = 0, \forall i = 1, \dots, l.$$

Proof. The necessary condition when constraints are linear can be seen in, for example, Girsanov (1972, Theorem 11.3). The sufficient condition does not require any constraint qualification. An example of proof can be seen in Corollary 1.1 of Barbu and Precupanu (1986, Chapter 3). □

Note that the definition of $\nabla f(x)$ here is in the sense of Frechet differentiability. The above theorem can be extended to linear equality constraints as a linear equality can be written as two linear inequalities.

Then the following theorem shows the relation between (P_0) and (D_0):

Theorem 2.2 *For problems (P_0) and (D_0), only one of the following two situations can happen:*

$$\begin{array}{l} (P_0) \text{ solvable and } (D_0) \text{ solvable,} \\ (P_0) \text{ infeasible and } (D_0) \text{ unbounded.} \end{array}$$

On the other hand, (P_C) and (D_C), $C > 0$, are both solvable.

Proof. First we prove that if (D_0) is solvable, (P_0) is solvable. Suppose that an optimal solution of (D_0) is α, by Theorem 2.1 (necessary condition), there exists a b such that

$$\alpha_i(Q\alpha - e - by)_i = 0, i = 1, \dots, l, \text{ and } Q\alpha - e - by \geq 0.$$

By defining $w \equiv \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)$, we have a (w, b) which satisfies the optimality condition of (P_0). From Theorem 2.1 (sufficient condition), (w, b) is an optimal solution of (P_0). Hence (P_0) is solvable.

Next we show that if (P_0) is feasible, (D_0) is solvable. Consider the following problem

$$\max_{\alpha \geq 0} \{ \inf_{w,b} [\frac{1}{2} \langle w, w \rangle - \sum_{i=1}^{l} \alpha_i (y_i (\langle w, \phi(x_i) \rangle + b) - 1)] \}. \tag{2.1}$$

When $y^T\alpha = 0$, $\inf_{w,b}[\frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{l} \alpha_i(y_i(\langle w, \phi(x_i)\rangle + b) - 1)]$ has a global minimum at $w = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)$ (see, for example, (Debnath and Mikusinski 1999, Example 9.3.3)). Thus (2.1) is equivalent to (D_0). It can be clearly seen that, if (P_0) is feasible with any feasible solution w^*, $\frac{1}{2}\langle w^*, w^*\rangle$ is an upper bound of (2.1). That is, $-\frac{1}{2}\langle w^*, w^*\rangle$ is a lower bound of (D_0). From Corollary 27.3.1 of Rockafellar (1970), any feasible bounded finite dimensional space quadratic convex function over a polyhedral attains at least an optimal solution. Since (D_0) is always feasible, we know that (D_0) is solvable.

Furthermore, as a feasible (P_0) implies a solvable (D_0) and a solvable (D_0) implies a solvable (P_0), we know that if (P_0) is feasible, its minimum is attained. Therefore, we conclude that there are only two possible situations:

$$\begin{array}{l} (P_0) \text{ solvable and } (D_0) \text{ solvable,} \\ (P_0) \text{ infeasible and } (D_0) \text{ unsolvable.} \end{array}$$

We then use the same property that if (D_0) is bounded, it is solvable. Therefore, (D_0) unsolvable means (D_0) is unbounded so the proof is complete.

Since (P_C) and (D_C), $C > 0$, are both feasible, by the same arguments above, they are both solvable. $\square$

Using a similar method, we can prove that $(\bar{P}_C)$ and $(\bar{D}_C)$ are also both solvable.

3 Penalty Parameters

In (P_C) (or $(\bar{P}_C)$), the penalty term $C \sum_{i=1}^{l} \xi_i$ (or $\frac{C}{2} \sum_{i=1}^{l} \xi_i^2$) is used to reduce the occurrence of wrongly classified data. However, this is different from the penalty approach in mathematical programming. Originally this method approximates a constrained problem by unconstrained minimization. When the penalty parameter is large enough, minima of unconstrained problems are close to a solution. Hence it is important to know that, if (P_0) is separable, how close a solution of (P_C) is to a solution of (P_0) as C increases. This can be a criteria to judge the appropriateness of using the penalty term and will be the main topic of this section. First we prove the following technical lemma:

Lemma 3.1 *For any $C \geq 0$, The vector w of optimal solutions of (P_C) is unique.*

Proof. Assume (w, b, ξ) and $(\bar{w}, \bar{b}, \bar{\xi})$ with $w \neq \bar{w}$ are both optimal solutions of (P_C). For any $0 \leq \alpha \leq 1$, $(\hat{w}, \hat{b}, \hat{\xi}) = \alpha(w, b, \xi) + (1-\alpha)(\bar{w}, \bar{b}, \bar{\xi})$ is feasible to (P_C). Therefore,

$$\frac{1}{2}\langle \hat{w}, \hat{w} \rangle + C \sum_{i=1}^{l} \hat{\xi}_i \geq \frac{1}{2}\langle w, w \rangle + C \sum_{i=1}^{l} \xi_i = \frac{1}{2}\langle \bar{w}, \bar{w} \rangle + C \sum_{i=1}^{l} \bar{\xi}_i. \tag{3.1}$$

Note that (3.1) can be treated as two inequalities. By multiplying the first part by α and the second part by $(1 - \alpha)$,

$$\frac{1}{2}\alpha^2 \langle w, w \rangle + \alpha(1-\alpha)\langle w, \bar{w} \rangle + \frac{1}{2}(1-\alpha)^2 \langle \bar{w}, \bar{w} \rangle \geq \frac{1}{2}\alpha \langle w, w \rangle + \frac{1}{2}(1-\alpha)\langle \bar{w}, \bar{w} \rangle.$$

Thus,

$$\alpha(1-\alpha)\langle w - \bar{w}, w - \bar{w} \rangle \leq 0, \ \forall 0 \leq \alpha \leq 1.$$

Since $w - \bar{w}$ is in a Hilbert space, $w - \bar{w} = 0$. This causes the contradiction so w must be unique. $\square$

For other variables such as b and ξ of (P_C), and α of (D_C), there are always examples where multiple solutions occur. Next we present a relation between solutions of $(P_C), C > 0$ and those of (P_0).

Theorem 3.2 *If (P_0) is solvable, there is a C^* such that for all $C \geq C^*$, any solution (w, b) of (P_0) is a solution of (P_C).*

Proof. If (w, b) is a solution of (P), by Theorem 2.1, there are α and b such that

$$w = \sum_{i=1}^{l} y_i \alpha_i \phi(x_i), y^T \alpha = 0, \alpha \geq 0$$
$$\alpha_i (y_i \langle \phi(x_i), w \rangle + b - 1) = 0, i = 1, \ldots, l.$$

Let $C^* > \max(\alpha_i)$. For all $C > C^*$, we can set $\xi = 0$ and obtain

$$w = \sum_{i=1}^{l} y_i \alpha_i \phi(x_i),\ y^T \alpha = 0$$
$$\alpha_i (y_i \langle \phi(x_i), w \rangle + b - 1 + \xi_i) = 0$$
$$0 \leq \alpha_i \leq C, \xi_i (C - \alpha_i) = 0, i = 1, \ldots, l.$$

Therefore, from Theorem 2.1, (w, b, ξ) is also a solution of (P_C). $\square$

The above theorem proves a property similar to exact penalty functions. That is, after the penalty parameter is large enough, any solution of the original problem is also a solution of penalty functions. However, according to Di Pillo (1994), they are only weak exact penalty functions, though most early literature on penalty functions studied this result. Since the penalty approach is an attempt to solve (P_0), it is more important to study the converse property which assures that solutions of (P_C) are solutions of (P_0). It is shown in the following theorem and corollary:

Theorem 3.3 *If (P_0) is solvable, there is a C^* such that for all $C \geq C^*$, any solution (w, b, ξ) of (P_C) satisfies that $\xi = 0$ and (w, b) is a solution of (P_0).*

Proof. Consider any C^* such that Theorem 3.2 holds and assume (w, b) is a solution of (P_0). From Theorem 3.2, $(w, b, 0)$ is also an optimal solution of (P_C). Now if we solve (P_C) and obtain $(\bar{w}, \bar{b}, \bar{\xi})$. From Lemma 3.1, $\bar{w} = w$. Therefore,

$$\frac{1}{2}\langle \bar{w}, \bar{w} \rangle + C \sum_{i=1}^{l} \bar{\xi}_i \leq \frac{1}{2}\langle w, w \rangle.$$

Thus, $\bar{\xi} = 0$ and $(\bar{w}, \bar{b})$ is an optimal solution of (P_0). $\Box$

Corollary 3.4 *If (P_0) is solvable and*

$$C^* \equiv \min\{ \max_{i=1,\dots,l} \alpha_i \mid \alpha \textit{ is a solution of } (D_0)\},$$

each of $(P_C), C < C^*$ has a different solution w from the solution of (P_0). On the contrary, $(P_C), C \geq C^*$ has the same solution w as that of (P_0).

Proof. When $C \geq C^*$, the result is given by Theorem 3.3. For the case of $C < C^*$, assume there is a (P_C) with a solution (w_C, b_C, ξ_C) such that w_C is also part of a solution of (P_0). If (w, b) is a solution of (P_0), (w, b) is feasible to (P_C) so

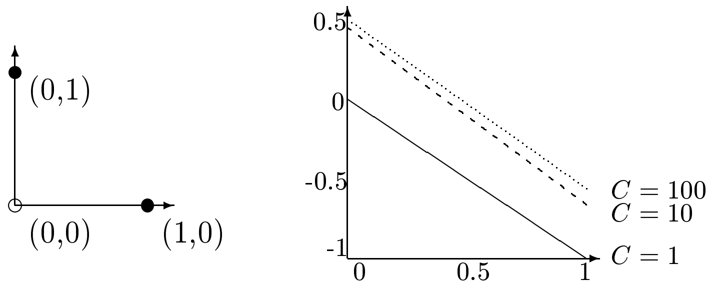
$$\frac{1}{2}\langle w_C, w_C \rangle + C \sum_{i=1}^{l} (\xi_C)_i \leq \frac{1}{2}\langle w, w \rangle.$$

This implies $\xi_C = 0$. Hence any solution of (D_C) is also a solution of (D_0). Since $C < C^*$, this contradicts to the definition of C^*. $\Box$

As we only show that vectors w are different, when $C < C^*$, it is still possible to obtain the same separating hyperplane. Note that $\langle w, \phi(x) \rangle + b = 0$ is the hyperplane so any $\alpha(w, b), \alpha \neq 0$, are appropriate coefficients. However, when $C < C^*$, the separating plane generated by (P_C) can also be very different from that of (P_0). For example, consider a problem with three points, 0, 1, and 100 where 0 is in the first class, and 1 and 100 are in the other class, and $K(x_i, x_j) = x_i^T x_j$. When $C \geq 2$, $\alpha = [2, 2, 0]$ and the hyperplane is $2x - 1 = 0$. When $1 \leq C \leq 2$, $\alpha = [C, C, 0]$ and $w = C$. As b can be any number which satisfies $-1 \leq b \leq 1 - C$, $b = -C/2$ can be a choice. Therefore, $2x - 1 = 0$ can still be a solution. However, when C becomes smaller, the hyperplane starts moving toward the training point 100. In addition, the distance between two planes

$\langle w, x \rangle + b = \pm 1$ becomes very large. This shows that if small Cs are used, some careful analysis should be conducted.

Next we switch the target to $(\bar{P}_C)$ and $(\bar{D}_C)$. Interestingly they are not like exact penalty functions. In the following example, we show that no matter how large C is, solutions of $(\bar{P}_C)$ are not solutions of (P_0). In Figure 3.1(a), there



(a) Training data (b) Separating hyperplanes

Figure 3.1: A simple example

are three training vectors in two different classes. If $Q_{ij} = y_i y_j x_i^T x_j$, the optimal separating hyperplane for (P_0) is $[2\ 2]x - 1 = 0$. The optimality condition of $(\bar{P}_C)$ is

$$\begin{bmatrix} \frac{1}{C} & 0 & 0 \\ 0 & 1+\frac{1}{C} & 0 \\ 0 & 0 & 1+\frac{1}{C} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + b \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}.$$

At the solution, all three vectors are support vectors so $\lambda = 0$. Solving the above equation with $y^T \alpha = 0$ obtains

$$b = \frac{C-1}{-C-3}, \alpha_1 = (1+b)C, \alpha_2 = \alpha_3 = \frac{(1-b)C}{C+1} = \frac{2C}{C+3}.$$

The separating hyperplane is

$$[\frac{C(1-b)}{C+1}\ \frac{C(1-b)}{C+1}]x + b = 0.$$

That is,

$$[2C\ 2C]x + (1-C) = 0. \tag{3.2}$$

No matter what C is, (3.2) is different from $[2\ 2]x - 1 = 0$. However, it can be seen in Figure 3.1(b) that as $C \to \infty$, (3.2) approaches $[2\ 2]x - 1 = 0$. In the following theorems we show that $(\bar{P}_C)$ has only properties of traditional penalty functions. That is, as $C \to \infty$, solutions of $(\bar{P}_C)$ approach solutions of (P_0).

Lemma 3.5 *Let $(w_{C_j}, b_{C_j}, \xi_{C_j})$ be a solution of $(\bar{P}_{C_j})$, α_{C_j} be the solution of $(\bar{D}_{C_j})$,$j = 1, 2, \ldots$, and assume $\lim_{j\to\infty} C_j = \infty$. If (P_0) is solvable and $\lim_{j\to\infty} \alpha_{C_j}$ exists, then $\lim_{j\to\infty} w_{C_j}$ exists and is the solution of (P_0).*

Proof. From the optimality condition (Theorem 2.1) of $(\bar{D}_C)$, $\alpha_{C_j} = C\xi_{C_j}$. Since $\lim_{j\to\infty} \alpha_{C_j}$ exists, $\lim_{j\to\infty} \xi_{C_j} = 0$. As $w_{C_j} = \sum_{i=1}^{l} y_i(\alpha_{C_j})_i \phi(x_i)$ and $\lim_{j\to\infty} \alpha_{C_j}$ exists, $\lim_{j\to\infty} w_{C_j}$ exists and we can denote it as w. Note that

$$\max_{i:y_i=1}(1 - (\xi_{C_j})_i - \langle w_{C_j}, \phi(x_i)\rangle) \le b_{C_j} \le \min_{i:y_i=-1}(-1 + (\xi_{C_j})_i + \langle w_{C_j}, \phi(x_i)\rangle).$$

As $C_j \to \infty$, $\lim_{j\to\infty} \xi_{C_j} = 0$ so

$$\max_{i:y_i=1}(1 - \langle w, \phi(x_i)\rangle) \le \min_{i:y_i=-1}(-1 + \langle w, \phi(x_i)\rangle).$$

Hence w is feasible to (P_0). Furthermore, as (P_0) is solvable, we can assume $\bar{w}$ is a solution. Since $\bar{w}$ is feasible to all $(\bar{P}_C), C > 0$ and w_{C_j} is the optimal solution of $(\bar{P}_{C_j})$, we have

$$\frac{1}{2}\langle w_{C_j}, w_{C_j}\rangle \le \frac{1}{2}\langle \bar{w}, \bar{w}\rangle.$$

As $j \to \infty$, $\frac{1}{2}\langle w, w\rangle \le \frac{1}{2}\langle \bar{w}, \bar{w}\rangle$. Hence $\lim_{j\to\infty} w_{C_j} = w$ is an optimal solution of (P_0). $\square$

Theorem 3.6 *Let (w_C, b_C, ξ_C) be a solution of $(\bar{P}_C)$. If (P_0) is solvable, then $\lim_{C\to\infty} w_C$ exists and is the solution of (P_0).*

Proof. First we prove that $\{\alpha_C \mid C \ge \epsilon\}$ is bounded, where α_C is a solution of (D_C) and ϵ is any positive number. If $\bar{\alpha}$ is a solution of (D_0), $\bar{\alpha}$ is feasible to $(\bar{D}_C)$ and α_C is feasible to (D_0) so

$$\begin{aligned} & \frac{1}{2}\bar{\alpha}^T Q\bar{\alpha} - e^T\bar{\alpha} \le \frac{1}{2}\alpha_C^T Q\alpha_C - e^T\alpha_C \\ \le\ & \frac{1}{2}\alpha_C^T(Q + \frac{I}{C})\alpha_C - e^T\alpha_C \le \frac{1}{2}\bar{\alpha}^T(Q + \frac{I}{C})\bar{\alpha} - e^T\bar{\alpha}. \end{aligned}$$

The above formula provides lower and upper bounds on $\frac{1}{2}\alpha_C^T(Q+\frac{I}{C})\alpha_C - e^T\alpha_C$ when $C \geq \epsilon > 0$. From the optimality condition (Theorem 2.1) of $(\bar{D}_C)$,

$$-\frac{1}{2}\alpha_C^T(Q+\frac{I}{C})\alpha_C + e^T\alpha_C = \frac{1}{2}\alpha_C^T(Q+\frac{I}{C})\alpha_C.$$

Hence $\frac{1}{2}\alpha_C^T(Q+\frac{I}{C})\alpha_C$ is bounded and so as $e^T\alpha_C$. Since $\alpha_C \geq 0$, this implies that $\{\alpha_C\}$ is bounded after C is large enough.

Since $\{\alpha_C \mid C \geq \epsilon\}$ is bounded, there exists a convergent sequence $\{\alpha_{C_j}\}$. From Lemma 3.5, $\lim_{j\to\infty} w_{C_j}$ exists, where w_{C_j} is part of a solution of $(\bar{P}_{C_j})$. If $\lim_{C\to\infty} w_C \neq w$, there is a sequence $\{w_{C_k}\}$ and $\delta > 0$ such that $\|w_{C_k} - w\| \geq \delta$, for all $k = 1, 2, \ldots$. However, $\{\alpha_{C_k}\}$ is bounded so we can find a subsequence of $\{\alpha_{C_k}\}$ which satisfies conditions of Lemma 3.5. Hence this subsequence converges to a solution of (P_0), that is, w. This contradicts to the assumption that $\|w_{C_k} - w\| \geq \delta, \forall k$. Therefore, $\lim_{C\to\infty} w_C = w$ $\square$

4 Conclusions

In this paper, through the use of optimization techniques, we have a better understanding about formulations of support vector machines. This work can be a good example on analyzing future SVM formulations by optimization techniques.

SVM formulations are very simple convex quadratic problems in Hilbert spaces. Thus we can use the simple Wolfe dual (2.1) to derive the primal-dual relation instead of using the standard conjugate duality. For different SVM formulations, the results of Theorem 2.2 could be quite different. For example, for the new ν-SVM proposed by Schölkopf et al. (2000), the primal problem can be unbounded while the dual problem can be infeasible (see, Crisp and Burges (1999), and (Chang and Lin 2000)).

Acknowledgments

This work was supported in part by the National Science Council of Taiwan via the grant NSC-88-2213-E-002-097. The author thanks Jorge Moré for pointing him to the reference (Di Pillo 1994) and some helpful discussions. He also thanks Chih-Wei Hsu and Steve Wright for their comments.

References

Barbu, V. and T. Precupanu (1986). *Convexity and Optimization in Banach Spaces*. Dordrecht: D. Reidel.

Bennett, K. P. and O. L. Mangasarian (1992). Robust linear programming discrimination of two linear inseparable sets. *Optimization Methods and Software 1*, 23–34.

Boser, B., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Proceeding of the Fifth Annual Workshop on Computational Learning Theory*.

Chang, C.-C. and C.-J. Lin (2000). Some analysis on ν-support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

Cortes, C. and V. Vapnik (1995). Support-vector network. *Machine Learning 20*, 273–297.

Crisp, D. J. and C. J. C. Burges (1999). A geometric interpretation of ν-svm classifiers. In *NIPS99*.

Debnath, L. and P. Mikusinski (1999). *Introduction to Hilbert Spaces with Applications* (Second ed.). San Diego: Academic Press.

Di Pillo, G. (1994). Exact penalty methods. In I. Ciocco (Ed.), *Algorithms for Continuous Optimization*. Dordrecht: Kluwer Acad. Publ.

Friess, T.-T., N. Cristianini, and C. Campbell (1998). The kernel adatron algorithm: a fast and simple learning procedure for support vector machines. In *Proceeding of 15th Intl. Conf. Machine Learning*. Morgan Kaufman Publishers.

Girsanov, I. V. (1972). *Lectures on Mathematical Theory of Extremum Problems*, Volume 67 of *Lecture Notes in Economics and Mathematical Systems*. Berlin: Springer-Verlag.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Schölkopf, B., C. J. C. Burges, and A. J. Smola (Eds.) (1998). *Advances in Kernel Methods - Support Vector Learning.* Cambridge, MA: MIT Press.

Schölkopf, B., A. Smola, R. C. Williamson, and P. L. Bartlett (2000). New support vector algorithms. *Neural Computation 12*, 1083 – 1121.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York, NY: Springer-Verlag.

Vapnik, V. (1998). *Statistical Learning Theory.* New York, NY: John Wiley.