Technical Reports on Mathematical and Computing Sciences: TR-C163
title: On the Influence of Outliers in the Support Vector Machine Framework
author: Yang Dai[1], Junichi Tanaka[2], and Osamu Watanabe[3]
affiliation:
1. Dept. of Bioengineering (MC063), Univ. Illionis at Chicago
   851 S. Morgan Str, Chicago, IL 60607-7052, USA
2. Dept. of Industrial Engineering, Tokyo Institute of Technology
   Meguro-ku Ookayama, Tokyo 152-8552, Japan
3. Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology
   Meguro-ku Ookayama, Tokyo 152-8552, Japan

email: watanabe@is.titech.ac.jp
acknowledgments to financial supports:
The first author conducted this research while she is with Dept. of Mathematical and Computing Sciences, Tokyo Institue of Technology, and is supported by a Grant-in-Aid (C-13650444) from Japanese Goverment.
The third author is supported in part by a Grant-in-Aid for Scientific Research on Priority Areas "Discovery Science" from Japanese Goverment.

Abstract. A Support Vector Machine (SVM in short) is a mechanism for data analysis such as classification, regression, and clustering. In this paper, we consider the soft margin formulation of SVM for two-class classification problems, and discuss a way to choose an appropriate parameter that determines the influence of misclassified examples. Based on the observation in [3], we propose one criterion for choosing the parameter, and give some fact justifying this criterion.

1 Introduction

The *Support Vector Machine* (SVM in short) is a modern mechanism for two-class classification, regression, and clustering problems. Since the present form of SVM was proposed [5], SVMs have been used in various application areas, and their classification power has been investigated in depth from both experimental and theoretical points of view. (See, e.g., [6].) An important feature is that their way of working, by identifying the so-called support vectors among the data, offers important contributions to a number of problems related to Data Mining.

Here we consider using SVM for the two-class classification problem. A basic formulation for the problem is presented as follows. Suppose that we are given a set of m examples $\boldsymbol{x}_i$, $1 \leq i \leq m$, in an n-dimensional space, say $\mathbb{R}^n$. Each example $\boldsymbol{x}_i$ is labeled by $y_i \in \{1, -1\}$ denoting the classification of the example. Then the *SVM training problem* is essentially to solve the following optimization problem (P1). (Here we follow [3] and use their formulation. The problem can be restated with a single threshold parameter as given in the original paper [5].)

Max Margin (P1)

$$\begin{array}{ll} \text{min.} & \frac{1}{2}\|\boldsymbol{w}\|^2 - (\theta_+ - \theta_-) \\ \text{w.r.t.} & \boldsymbol{w} = (w_1, ..., w_n), \theta_+, \text{ and } \theta_-, \\ \text{s.t.} & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \geq\ \theta_+, \quad \text{if } y_i = 1, \\ & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \leq\ \theta_-, \quad \text{if } y_i = -1. \end{array}$$

In this formulation, it is assumed that the data set is linearly separable; that is, a hyperplane separating the two classes of examples exists. By the *solution* of (P1), we mean the hyperplane that achieves the minimum cost. Given a solution, its support vectors are the data points $\boldsymbol{x}_i$ for which, at the solution, the corresponding inequality is tight; that is, $\boldsymbol{w} \cdot \boldsymbol{x}_i = \theta_+$ if $y_i = 1$, $\boldsymbol{w} \cdot \boldsymbol{x}_i = \theta_-$ if $y_i = -1$.

Remarks on Notations. Throughout this paper, we use X to denote the set of examples, and let n and m denote the dimension of the example space and the number of examples respectively. Also we use i for indexing examples (and their labels), and $\boldsymbol{x}_i$ and y_i to denote the ith example and its label respectively. The range of i is always $\{1, ..., m\}$.

An important feature of SVM is that it is also applicable for the nonseparable case. More precisely, for nonseparable data we can take two positions: (i) the case where we consider that a hyperplane is too weak to be a classifier for our given examples, and that we should be able to fit them better nonlinearly; and (ii) the case where we consider that there are some erroneous examples or exceptions, i.e., "outliers", which should be somehow identified and allowed to be misclassified. Of course, it would be better if we can use a "reasonable" nonlinear classifier. Nevertheless, the second approach is important as well if we suspect that outliers exist in a given set of examples. The usability of SVM is due to the fact that we can use both.

In this paper, we focus on the second approach, and discuss how to deal with outliers. A standard SVM solution to the case where some outliers exist is to relax the constraints by introducing slack variables or "soft margin error". That is, we consider the following generalization of the problem (P1), corresponding to the soft margin hyperplane separation problem.

$$
\begin{array}{lll}
\multicolumn{3}{l}{\underline{\text{Max Soft Margin (P2)}}} \\
\text{min.} & \multicolumn{2}{l}{\frac{1}{2}\|\boldsymbol{w}\|^2 - (\theta_+ - \theta_-) + D\sum_i \xi_i} \\
\text{w.r.t.} & \multicolumn{2}{l}{\boldsymbol{w} = (w_1, ..., w_n), \theta_+, \theta_-,} \\
 & \multicolumn{2}{l}{\text{and } \xi_1, ..., \xi_m,} \\
\text{s.t.} & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \geq\ \theta_+ - \xi_i, & \text{if } y_i = 1, \\
 & \boldsymbol{w} \cdot \boldsymbol{x}_i \ \leq\ \theta_- + \xi_i, & \text{if } y_i = -1, \\
 & \text{and } \xi_i \ \geq\ 0. &
\end{array}
$$

For a given set X of examples, suppose we solve the problem (P2) and obtain the optimal hyperplane. Then an example in X is called an *outlier* if it is misclassified with this hyperplane. On the other hand, examples other than outliers are called *normal examples*. Throughout this paper, we use ℓ to denote the number of outliers. Notice that this definition of outlier is relative both to the hypothesis class and to the soft margin parameter D, which determines the degree of influence of the outliers. In this paper, we discuss a way to choose this parameter D appropriately.

Though important, there seems, at least as far as the authors know, no systematic method for choosing the influence parameter D. The most standard way [6] is to try several choices and use the one with the best performance on the training set. Note that D should be fixed in advance; that is, when solving (P2) (in other words, when training SVM), D is considered as a constant. But of course, we may be able to revise D depending the result of solving (P2). Thus, if there is a reasonable criterion for D depending on the solution of (P2), then we would have at least the following method to choose D: solve (P2) by revising D until some D satisfying the criterion is obtained. Here we propose one simple criterion for D and give a mathematical fact supporting this criterion.

2 Alternative Formulations

Bennett and Bredensteiner [3] showed an alternative formulation of (P2), giving us an intuitive geometric interpretation of the influence parameter D. Since our criterion is based on their geometric interpretation, we explain here briefly the formulations introduced in [3]. (In fact, a criterion equivalent to us has been given, though implicitly, in [3].)

First, it is shown [3] that (P2) is equivalent to the following problem (P3). (More precisely, the problem (P3) is the Wolfe dual of (P2).)

Reduced Convex Hull (P3)

$$\begin{array}{ll} \text{min.} & \dfrac{1}{2}\left\|\sum_i y_i s_i \boldsymbol{x}_i\right\|^2 \quad \text{w.r.t. } s_1, ..., s_m, \\ \text{s.t.} & \sum_{i:\, y_i=1} s_i \;=\; 1, \quad \sum_{i:\, y_i=-1} s_i \;=\; 1, \quad \text{and } 0 \le s_i \le D. \end{array}$$

Note that

$$\left\|\sum_i y_i s_i \boldsymbol{x}_i\right\|^2 \;=\; \left\|\sum_{i:\, y_i=1} s_i \boldsymbol{x}_i - \sum_{i:\, y_i=-1} s_i \boldsymbol{x}_i\right\|^2.$$

That is, this is the distance between two points in the convex hull of positive and negative examples. In the separable case, it is the distance between two closest points. On the other hand, in the nonseparable case, we give some restriction to the influence of each example and consider only points defined by examples in such a way that each example cannot contribute more than D.

Bennett and Bredensteiner [3] gave the following intuitive interpretation of D: First assume, for simplifying our discussion, that its inverse $k = 1/D$ is integer. Then our objective can be regarded as the distance between the convex hulls of "composed examples", points that are defined as a center of k examples. Then resulting convex hulls are reduced ones as in the formulation (P3).

More formally, we can reformulate (P3) as follows. Let Z be the set of *composed examples* $\boldsymbol{z}_I$ that is defined by

$$\boldsymbol{z}_I \;=\; \frac{\boldsymbol{x}_{i_1} + \boldsymbol{x}_{i_2} + \cdots + \boldsymbol{x}_{i_k}}{k},$$

with some k distinct elements $\boldsymbol{x}_{i_1}$, $\boldsymbol{x}_{i_2}$, ..., $\boldsymbol{x}_{i_k}$ of X with the same label (i.e., $y_{i_1} = y_{i_2} = \cdots = y_{i_k}$). That is, a composed example is a mass center of all groups of k homogeneously labeled initial data points. The label y_I of the composed example $\boldsymbol{z}_I$ inherits its members'. Throughout this note, we will use I for indexing elements of Z and their labels. For each $\boldsymbol{z}_I$, we use z_I to denote the set of original examples from which $\boldsymbol{z}_I$ is composed. (For distinguishing from composed examples $\boldsymbol{z}_I$, we will call $\boldsymbol{x}_i$ an *original example.*) Also we will use M to denote the number of composed examples.

Then it is easy to see that (P3) is equivalent to the following (P4) [3].

$$\underline{\text{Convex Hull of Reduced Points (P4)}}$$

$$\begin{array}{ll} \text{min.} & \displaystyle\frac{1}{2}\left\|\sum_I y_I s_I \boldsymbol{z}_I\right\|^2 \quad \text{w.r.t. } s_1, ..., s_M, \\ \text{s.t.} & \displaystyle\sum_{I:\, y_I=1} s_I \;=\; 1, \quad \sum_{I:\, y_I=-1} s_I \;=\; 1, \quad \text{and } 0 \le s_I \le 1. \end{array}$$

As proposed in [2], we further consider the Wolfe primal of this problem again. Then we come back to the one almost identical to (P1)!

$$\underline{\text{Max Margin for Composed Examples (P5)}}$$

$$\begin{array}{lrcl} \text{min.} & \multicolumn{3}{l}{\frac{1}{2}\|\boldsymbol{w}\|^2 - (\eta_+ - \eta_-)} \\ \text{w.r.t.} & \multicolumn{3}{l}{\boldsymbol{w} = (w_1, ..., w_n), \eta_+, \text{ and } \eta_-,} \\ \text{s.t.} & \boldsymbol{w}\cdot\boldsymbol{z}_I & \ge & \eta_+, \quad \text{if } y_I = 1, \\ & \boldsymbol{w}\cdot\boldsymbol{z}_I & \le & \eta_-, \quad \text{if } y_I = -1. \end{array}$$

In general, composed examples may not be separable. But they are separable for sufficiently large k. For example, in the extreme case where $k = m_+ = m_-$ (where m_+ and m_- are respectively the number of positive and negative examples), we have only one positive and one negative composed example, which are clearly separable (unless they are the same).

Based on the results in [3], we can show the following relationship between (P2) and (P5).

Theorem 2.1 *Suppose that k is large enough so that the obtained composed examples are linearly separable. Then the objective value achieved by the optimal solution is the same between (P2) and (P5); furthermore, the weight vector for the optimal separating hyperplane, i.e., $\boldsymbol{w}^*$, coincides for (P2) and (P5). (Note that the margin parameters (θ_+^*, θ_-^*) and (η_+^*, η_-^*) are usually different.)*

3 A Criterion for the Parameter D

To simplify our discussion, we consider for D only those expressed as $D = 1/k$ for some integer k. Since we have Theorem 2.1, one natural criterion for D (or k) is the linear separability of composed examples. Note, on the other hand, this criterion is rather difficult to use when solving (P2). Our technical contribution of this paper is to prove the following characterization theorem, thereby providing a way to determine the linear separability of composed examples.

Theorem 3.1 *Let X be any set X of examples, and for any k, let Z be the set of composed examples made up from k examples from X. Let $(\boldsymbol{w}^*, \theta_+^*, \theta_-^*, \boldsymbol{\xi}^*)$ be a solution of (P2) on X. Then $\|\boldsymbol{w}^*\| > 0$ (or, equivalently $\boldsymbol{w}^* \neq \boldsymbol{0}$) if and only if k is large enough so that Z is linearly separable.*

Remark. *The problem (P5) does not make sense if $k > m_+$ or $k > m_-$, where m_+ and m_- are respectively the total number of positive and negative examples in X. Similarly, the problem*

(P2) is unbounded if $D < 1/m_+$ or $D < 1/m_-$. Thus, we assume below that $k \le m_+$ and $k \le m_-$.

From this theorem, if D is too large and the composed examples are not linearly separable, then we know it by obtaining the trivial answer $\boldsymbol{w}^* = \mathbf{0}$ by solving (P2); on the other hand, if D is small enough (i.e., k is large enough) so that the composed examples are linearly separable, then (P2) should have a nontrivial answer. That is, we can determine whether D is appropriate by checking whether (P2) has a nontrivial solution.

In the context of linear programming type classification, Bennet and Mangasarian [4] proposed to use this "linear separability" for choosing weights that are similar to our influence parameter. There they also provide a way to decide whether examples are linearly separable under a given choice of weights. Here we prove that a similar characterization theorem.

Let us prove our characterization theorem. Though our proof is not so difficult, it is not immediate from the equivalence between (P2) and (P5) proved in Theorem 2.1. Notice that the equivalence is guaranteed *when* the composed examples are linearly separable. Thus, we cannot use Theorem 2.1 for proving our characterization theorem.

First we note the following facts.

Fact 1 *Let $(\boldsymbol{w}^*, \theta_+^*, \theta_-^*, \boldsymbol{\xi}^*)$ be an optimal solution of (P2). Then the number of respectively positive and negative outliers w.r.t. $(\boldsymbol{w}^*, \theta_+^*, \theta_-^*)$ is at most k.*

Proof. The optimal solution $(\boldsymbol{w}^*, \theta_+^*, \theta_-^*, \boldsymbol{\xi}^*)$ satisfies the KKT-condition. In particular, the following holds for some $\boldsymbol{s}^*$ and $\boldsymbol{u}^*$. (Below we use i to denote indices of examples, and let P and N respectively denote the set of indices i of examples such that $y_i = 1$ and $y_i = -1$. We use $\boldsymbol{e}$ to denote the vector with 1 at every entry.)

$$
\begin{array}{l}
\boldsymbol{w}^* - \sum_{i\in P} s_i \boldsymbol{x}_i + \sum_{i\in N} s_i^* \boldsymbol{x}_i \ = \ 0, \\
D\boldsymbol{e} - \boldsymbol{s}^* - \boldsymbol{u}^* \ = \ 0, \\
\displaystyle\sum_{i\in P} s_i^* \ = \ \sum_{i\in N} s_i^* \ = \ 1, \\
\forall i \in P \ [\ s_i^*(\boldsymbol{w}^* \cdot \boldsymbol{x}_i - \theta_+^* + \xi_i^*) \ = \ 0\ ], \\
\forall i \in N \ [\ s_i^*(\boldsymbol{w}^* \cdot \boldsymbol{x}_i - \theta_-^* - \xi_i^*) \ = \ 0\ ], \\
\boldsymbol{u}^* \cdot \boldsymbol{\xi}^* \ = \ 0 \text{ (i.e., } (D\boldsymbol{e} - \boldsymbol{s}^*) \cdot \boldsymbol{\xi}^* = 0), \\
\boldsymbol{\xi}^*, \boldsymbol{u}^*, \boldsymbol{s}^* \ \ge \ 0.
\end{array}
$$

Now outliers are those $\boldsymbol{x}_i$ with $\xi_i^* > 0$. Then from the above, we have $D - s_i = 0$, i.e., $s_i = 1/k$, for every outliner $\boldsymbol{x}_i$. On the other hand, $\boldsymbol{s}$ must satisfy $\sum_{i\in P} s_i^* = \sum_{i\in N} s_i^* = 1$. Thus, there are at most k outliers among positive and negative examples respectively. □

Fact 2 *Let $(\boldsymbol{w}, \theta_+, \theta_-, \boldsymbol{\xi})$ be any feasible solution of (P2) such that $\boldsymbol{w} = \mathbf{0}$. Then the objective value corresponding to this solution is more than or equal to 0.*

Proof. The objective value is clearly positive if $\theta_+ \le \theta_-$. Thus, we consider here the case that $\theta_+ > \theta_-$.

Since $\boldsymbol{w} = \mathbf{0}$, it follows from the constraints of (P2) that $\theta_+ \le \xi_i$ for every positive $\boldsymbol{x}_i$, and $-\theta_- \le \xi_i$ for every negative $\boldsymbol{x}_i$. Then by using the assumption (see the remark of Theorem 3.1) that $Dm_+ \ge 1$ and $Dm_- \ge 1$, we have

$$
\begin{aligned}
\theta_+ \quad &\le \min_{i\colon y_i=+1} \xi_i \;\le\; Dm_+\left(\min_{i\colon y_i=+1} \xi_i\right) \;\le\; D \sum_{i\colon y_i=+1} \xi_i, \\
-\theta_- \quad &\le \min_{i\colon y_i=-1} \xi_i \;\le\; Dm_-\left(\min_{i\colon y_i=-1} \xi_i\right) \;\le\; D \sum_{i\colon y_i=-1} \xi_i.
\end{aligned}
$$

Hence, by adding both sides, we have $\theta_+ - \theta_- \le D\sum_i \xi_i$; therefore, the objective value of (P2) for this solution is nonnegative. □

Proof of Theorem 3.1. Let Z_+ and Z_- respectively denote the set of all positive and negative composed examples of Z. By "Z is linearly separable", we formally mean that the following holds for *some* $\boldsymbol{w}$.

$$
\min_{\boldsymbol{z}_I \in Z_+} \boldsymbol{w} \cdot \boldsymbol{z}_I \;>\; \min_{\boldsymbol{z}_J \in Z_-} \boldsymbol{w} \cdot \boldsymbol{z}_J. \tag{1}
$$

(The if direction) Since Z is separable, there must be some $\boldsymbol{w}$ satisfying (1). Clearly $\boldsymbol{w} \ne \boldsymbol{0}$. We construct a feasible solution from $\boldsymbol{w}$ with the negative objective value, which, with Fact 2, proves that $\boldsymbol{w}^* \ne \boldsymbol{0}$.

First note that for any θ_+ and θ_-, there exists some $\boldsymbol{\xi}$ such that $(\boldsymbol{w}, \theta_+, \theta_-, \boldsymbol{\xi})$ satisfies the constraints of (P2). In particular, we can define θ_+ and θ_- so that
(a) $\ell_+ \le k-1$, and there are $k - \ell_+$ positive examples $\boldsymbol{x}_i$ such that $\boldsymbol{w} \cdot \boldsymbol{x}_i = \theta_+$; and
(b) $\ell_- \le k-1$, and there are $k - \ell_-$ negative examples $\boldsymbol{x}_i$ such that $\boldsymbol{w} \cdot \boldsymbol{x}_i = \theta_-$,
where ℓ_+ and ℓ_- are respectively the number of positive and negative examples $\boldsymbol{x}_i$ such that $\xi_i > 0$. (We also use $X_{\mathrm{err},+}$ and $X_{\mathrm{err},-}$ to denote respectively the set of positive and negative examples $\boldsymbol{x}_i$ such that $\xi_i > 0$.)

Then we have

$$
\begin{aligned}
\min_{\boldsymbol{z}_I \in Z_+} \boldsymbol{w} \cdot \boldsymbol{z}_I \;&=\; \frac{(k-\ell_+)\theta_+ + \sum_{i \in X_{\mathrm{err},+}} (\theta_+ - \xi_i)}{k} \\
&=\; \theta_+ - \frac{1}{k} \sum_{i \in X_{\mathrm{err},+}} \xi_i \;=\; \theta_+ - D \sum_{i \in X_{\mathrm{err},+}} \xi_i.
\end{aligned}
$$

Similarly, we have

$$
\max_{\boldsymbol{z}_J \in Z_+} \boldsymbol{w} \cdot \boldsymbol{z}_J \;=\; \theta_- + D \sum_{i \in X_{\mathrm{err},+}} \xi_i.
$$

Then it follows from (1) that

$$
\theta_+ - D \sum_{i \in X_{\mathrm{err},+}} \xi_i \;>\; \theta_- + D \sum_{i \in X_{\mathrm{err},+}} \xi_i,
$$

and hence,

$$
(\theta_+ - \theta_-) - D \sum_i \xi_i \;>\; 0.
$$

Now since the above value is not zero, there must be some $b > 0$ that satisfies

$$(\theta_+ - \theta_-) - D\sum_i \xi_i \;=\; b\|\boldsymbol{w}\|^2.$$

On the other hand, it is clear that $(a\boldsymbol{w}, a\theta_+, a\theta_-, a\boldsymbol{\xi})$ is also feasible for any $a > 0$, and the objective value of this solution is now computed as

$$\frac{1}{2}\|a\boldsymbol{w}\|^2 - (a\theta_+ - a\theta_-) + D\sum_i a\xi_i \;=\; \frac{a^2}{2}\|\boldsymbol{w}\|^2 - ab\|\boldsymbol{w}\|^2 \;=\; a^2\|\boldsymbol{w}\|^2\left(\frac{1}{2} - \frac{b}{a}\right).$$

Thus, by choosing a appropriately, we can make the objective value negative.

(The only-if direction) Suppose that an optimal solution $(\boldsymbol{w}^*, \theta_+^*, \theta_-^*, \boldsymbol{\xi}^*)$ of (P2) satisfies $\boldsymbol{w}^* \neq 0$. We show that $\min_{\boldsymbol{z}_I \in Z_+} \boldsymbol{w}^* \cdot \boldsymbol{z}_I > \max_{\boldsymbol{z}_J \in Z_-} \boldsymbol{w}^* \cdot \boldsymbol{z}_J$.

Consider a composed example $\boldsymbol{z}_+$ with the smallest $\boldsymbol{w}^* \cdot \boldsymbol{z}_+$. Note that for any outlier $\boldsymbol{x}_i$ and any normal example $\boldsymbol{x}_j$, we have $\boldsymbol{w}^* \cdot \boldsymbol{x}_i < \boldsymbol{w}^* \cdot \boldsymbol{x}_j$. Furthermore, the number of positive outliers ℓ_+ is at most k (Fact 1). Hence, the value $\boldsymbol{w}^* \cdot \boldsymbol{z}_+$ is minimized if z_+ contains all positive outliers and the other (normal) elements $\boldsymbol{x}_j$ of z_+ satisfies $\boldsymbol{w}^* \cdot \boldsymbol{x}_j = \theta_+^*$. Thus, letting $X_{\mathrm{err},+}$ be the set of positive outliers, we have

$$\begin{aligned}\min_{\boldsymbol{z}_+ \in Z_+} \boldsymbol{w}^* \cdot \boldsymbol{z}_+ \;&\geq\; \frac{(k - \ell_+)\theta_+^* + \sum_{\boldsymbol{x}_i \in X_{\mathrm{err},+}} (\theta_+^* - \xi_i^*)}{k} \\ &=\; \theta_+^* - \frac{1}{k}\sum_{\boldsymbol{x}_i \in X_{\mathrm{err},+}} \xi_i^* \;=\; \theta_+^* - D\sum_{\boldsymbol{x}_i \in X_{\mathrm{err},+}} \xi_i^*.\end{aligned}$$

Similarly, letting $X_{\mathrm{err},-}$ be the set of negative outliers, we have

$$\max_{\boldsymbol{z}_- \in Z_-} \boldsymbol{w}^* \cdot \boldsymbol{z}_- \;\leq\; \theta_-^* + D\sum_{\boldsymbol{x}_i \in X_{\mathrm{err},-}} \xi_i^*.$$

Thus,

$$\min_{\boldsymbol{z}_+ \in Z_+} \boldsymbol{w}^* \cdot \boldsymbol{z}_+ - \max_{\boldsymbol{z}_- \in Z_-} \boldsymbol{w}^* \cdot \boldsymbol{z}_- \;\geq\; (\theta_+^* - \theta_-^*) - D\sum_i \xi_i^*.$$

On the other hand, since $\boldsymbol{w}^* \neq 0$ is an optimal solution and the objective value gets 0 by the solution $(\boldsymbol{w}, \theta_+, \theta_-, \boldsymbol{\xi}) = (\boldsymbol{0}, 0, 0, \boldsymbol{0})$, the objective value corresponding to $\boldsymbol{w}^* \neq 0$ must be at most 0. Thus, we have

$$\frac{1}{2}\|\boldsymbol{w}^*\|^2 - (\theta_+^* - \theta_+^*) + D\sum_i \xi_i \;\leq\; 0,$$

which implies (noting $\|\boldsymbol{w}^*\|^2 > 0$)

$$(\theta_+^* - \theta_-^*) - D\sum_i \xi_i \;>\; 0.$$

Therefore, we have

$$\min_{\boldsymbol{z}_I \in Z_+} \boldsymbol{w}^* \cdot \boldsymbol{z}_I \;>\; \max_{\boldsymbol{z}_J \in Z_-} \boldsymbol{w}^* \cdot \boldsymbol{z}_J.$$

□

4 Concluding Remarks

Assume, as our working hypothesis, that the separability of composed examples is an appropriate criterion for choosing the influence parameter D. Then from our theorem one can easily think of an algorithmic way to determine D under this criterion. The simplest and straightforward way is to use the binary search method to find the largest D $(= 1/k)$ such that (P2) has a nontrivial solution $\boldsymbol{w}^*$, i.e., $\boldsymbol{w}^* \neq \mathbf{0}$. For this, we need to solve (P2) several times with different D, which may be computationally hard if there are huge number of examples. We point out here that a random sampling approach proposed in [2] could be used to solve this hardness.

Based on the random sampling technique, Balcázar etal proposed a randomized algorithm solving (P2) by solving (P2) several times on a small subset of examples randomly chosen according to their weights. If D is not small enough, it may be found in an early stage of this randomized computation, and we can revise D (e.g., make it half) on the course of the randomized computation. By this way, we may be able to combine the search for an appropriate choice of D with the random sampling iteration.

Notice further that it is not necessary to use the same parameter D for all examples. For example, one can use small D_i for an example $\boldsymbol{x}_i$ that seems like an outlier. It would be very interesting if this adjustment of D_i's can be coordinated with the modification of the weight in the randomized algorithm.

References

[1] J. L. Balcázar, Y. Dai, and O. Watanabe, A Random sampling technique for training support vector machines: for a primal-form maximal-margin classifiers, in *Proc. 12th Int'l Conf. on Algorithmic Learning Theory* (ALT'01) LNAI 2225, 119−134, 2001.

[2] J. L. Balcázar, Y. Dai, and O. Watanabe, Provably fast training algorithms for support vector machines, in *Proc. First IEEE International Conference on Data Mining* (ICDM'01), IEEE, 43−50, 2001.

[3] K.P. Bennett and E.J. Bredensteiner, Duality and geometry in SVM classifiers, in *Proc. 17th Int'l Conf. on Machine Learning* (ICML'2000), 57−64, 2000.

[4] K.P. Bennett and O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, *Optimization Methods and Software* 1, 23−34, 1992.

[5] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning* 20, 273−297, 1995.

[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press 2000.