

---

# A Unified Robust Classification Model

---

**Akiko Takeda, Hiroyuki Mitsugi**

Keio University, 3-14-1 Hiyoshi, Kouhoku, Yokohama, Kanagawa 223-8522, Japan

TAKEDA@AE.KEIO.AC.JP, KIYUROHI7@Z2.KEIO.JP

**Takafumi Kanamori**

Nagoya University, Furocho, Chikusa-ku, Nagoya-shi, Aichi 464-8603, Japan

KANAMORI@IS.NAGOYA-U.AC.JP

## Abstract

A wide variety of machine learning algorithms such as support vector machine (SVM), minimax probability machine (MPM), and Fisher discriminant analysis (FDA), exist for binary classification. The purpose of this paper is to provide a unified classification model that includes the above models through a robust optimization approach. This unified model has several benefits. One is that the extensions and improvements intended for SVM become applicable to MPM and FDA, and vice versa. Another benefit is to provide theoretical results to above learning methods at once by dealing with the unified model. We give a statistical interpretation of the unified classification model and propose a non-convex optimization algorithm that can be applied to non-convex variants of existing learning methods.

## 1. Introduction

There are a wide variety of machine learning algorithms for binary classification. *Support vector machine* (SVM) is one of the most successful classification algorithms in modern machine learning (Schölkopf & Smola, 2002). The minimax probability machine (MPM) (Lanckriet et al., 2002) and Fisher discriminant analysis (FDA) (Fukunaga, 1990) also address the binary classification problem. Their problem settings assume that only the mean and covariance matrix of each class are known. The optimal hyperplane of MPM is determined by minimizing the worst-case (maximum) probability of misclassification of unseen test samples over all possible class-conditional distributions. FDA is to find a direction which maximizes

the projected class means while minimizing the class variance in this direction.

The purpose of this paper is to provide a unified framework for learning algorithms, including SVM, MPM, and FDA, from the viewpoint of robust optimization (Ben-Tal et al., 2009). Robust optimization is an approach that handles optimization problems defined by uncertain inputs. A simple example of robust optimization is

$$\max_{\mathbf{w} \in \mathcal{W}} \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}, \quad (1)$$

where  $\mathbf{w}$  is the parameter to be optimized under the constraint  $\mathbf{w} \in \mathcal{W}$  and  $\mathbf{x}$  is an uncertain input in the problem. The *uncertainty set*  $\mathcal{U}$  represents the uncertainty of the input. (1) determines the decision making parameter  $\mathbf{w}$  which maximizes the benefit  $\mathbf{x}^\top \mathbf{w}$  for the worst-case setup among  $\mathbf{x} \in \mathcal{U}$ .

For binary classification, we regard the means  $\mathbf{x}_+$  and  $\mathbf{x}_-$  of the data points of each class as uncertain inputs and prepare uncertainty sets  $\mathcal{U}_+$  and  $\mathcal{U}_-$  of those uncertain inputs. We assume that  $\mathbf{x}$  of (1) exists in the Minkowski difference  $\mathcal{U}$  of  $\mathcal{U}_+$  and  $\mathcal{U}_-$ , i.e.,

$$\mathcal{U} = \mathcal{U}_+ \ominus \mathcal{U}_- := \{\mathbf{x}_+ - \mathbf{x}_- \mid \mathbf{x}_+ \in \mathcal{U}_+, \mathbf{x}_- \in \mathcal{U}_-\},$$

and define  $\mathcal{W}$  by  $\{\mathbf{w} \mid \|\mathbf{w}\|^2 = 1\}$ , where  $\|\cdot\|$  is the Euclidean norm. Then we transform (1) into

$$\max_{\mathbf{w}: \|\mathbf{w}\|^2 = 1} \min_{\mathbf{x}_+ \in \mathcal{U}_+, \mathbf{x}_- \in \mathcal{U}_-} (\mathbf{x}_+ - \mathbf{x}_-)^\top \mathbf{w}. \quad (2)$$

We call it *robust classification model* (RCM)<sup>1</sup>. This problem always seems to be non-convex because of  $\mathcal{W}$ . However, it reduces to a convex problem that includes a constraint  $\|\mathbf{w}\|^2 \leq 1$  instead of  $\|\mathbf{w}\|^2 = 1$  when  $\mathcal{U}_+$  and  $\mathcal{U}_-$  do not intersect.

---

<sup>1</sup> Here we used the terminology of “robust” for the model (2) from the notion of “robust optimization”, not from the notion of “robust statistics”. The aim of the RCM is in providing a unified framework to existing learning methods, not in providing a learning method with better tolerance to outliers.

In this paper, we show that RCM (2) reduces to the learning methods mentioned above, depending on a prescribed uncertainty set  $\mathcal{U}$ . For example, we show that MPM is a special case of (2) with an ellipsoidal uncertainty set  $\mathcal{U}$ . When  $\mathcal{U}_+$  and  $\mathcal{U}_-$  are defined as reduced convex hulls (Bennett & Bredensteiner, 2000), (2) reduces to  $\nu$ -SVM (Schölkopf et al., 2000) if  $\mathcal{U}_+ \cap \mathcal{U}_- = \emptyset$  and reduces to E $\nu$ -SVM (Perez-Cruz et al., 2003), otherwise. The difference between these learning methods turns out only to be in the definition of  $\mathcal{U}$  of (2).

The first contribution of handling the unified model (2) is to obtain new learning methods. For example, we can obtain non-convex variants of MPM and FDA by mimicking Perez-Cruz et al.’s extension (Perez-Cruz et al., 2003) from convex  $\nu$ -SVM to non-convex E $\nu$ -SVM.

The second contribution is to provide theoretical results to above learning methods at once by dealing with the unified model (2). Indeed, we provide statistical interpretation for (2) on the basis of the conventional statistical learning theory. We show that (2) with some corresponding uncertainty set is a good approximation for the worst-case minimization of expected loss functions under uncertain probabilities.

We also provide a generalized local optimum search algorithm, that is applicable to non-convex variants of learning models. We prove theoretical results on the local optimum search algorithm.

The paper is organized as follows. In Section 2, we elucidate the unified model, RCM (2), for classification problems. In Section 3, we show RCM’s connection with existing learning algorithms and obtain non-convex variants for MPM and FDA in the same way as non-convex E $\nu$ -SVM. In Section 4, we give a statistical interpretation of RCM in terms of minimizing the upper and lower bounds of the worst-case expected loss. In Section 5, we describe a local optimum search algorithm for non-convex RCM. We summarize our contributions and future work in Section 6.

## 2. Unified Robust Classification Model

### 2.1. Problem Settings

We shall start by introducing the problem setting and the notations. The observed training samples are denoted as  $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{+1, -1\}$ ,  $i \in M := \{1, \dots, m\}$ . Let  $M_+$  be the set of indices of training samples with the label +1; likewise for  $M_-$ . Let  $|M_+| = m_+$  and  $|M_-| = m_-$ , where  $|\cdot|$  shows the size of the set.

The goal of the classification task is to obtain a classi-

fier that minimizes the prediction error rate for unseen test samples. For the sake of simplicity, we shall focus on linear classifiers, i.e.,  $\mathbf{x}^\top \mathbf{w} + b$  where  $\mathbf{w} (\in \mathbb{R}^d)$  is a vector and  $b (\in \mathbb{R})$  is a bias parameter. Most of the discussions in this paper can be directly applied to kernel classifiers (Schölkopf & Smola, 2002). Concretely, the change from  $\mathbf{x} \in \mathcal{X}$  to the kernel function  $k(\cdot, \mathbf{x})$  makes statements of Sections 2-4 hold for kernel classifiers, while the algorithm in Section 5 needs small modification.

We shall assume that the training samples are not reliable because of noise or measurement errors. To make a classification model less sensitive to noise in the training samples, we shall focus on representative points of each class, denoted by  $\mathbf{x}_+$  and  $\mathbf{x}_-$ . These points are not necessarily individual samples, but may be means of the data points of each class. Since the training samples are not reliable, it is reasonable to assume that  $\mathbf{x}_+$  and  $\mathbf{x}_-$  will involve some uncertainty. The largest possible sets of  $\mathbf{x}_+$  and  $\mathbf{x}_-$  are denoted by  $\mathcal{U}_+$  and  $\mathcal{U}_-$ , respectively, and these sets are defined on the basis of training samples. Throughout this paper, we will assume that both  $\mathcal{U}_+$  and  $\mathcal{U}_-$  are convex and compact and that they have interior points. Then, their Minkowski difference  $\mathcal{U}$  is convex and has a nonempty interior.

The way of constructing the uncertainty set  $\mathcal{U}_\pm$  is a very important issue in practice. If we set  $\mathcal{U}$  too large in (2), the optimal decision is very robust to uncertain data  $\mathbf{x}$  but too conservative. Moreover, if we define  $\mathcal{U}$  with complicated functions, we cannot easily solve (2). Many robust optimization studies have used polyhedral sets and ellipsoidal sets as  $\mathcal{U}$  for the sake of computational tractability. We show examples of  $\mathcal{U}_+$  and  $\mathcal{U}_-$  in Section 3. We might possibly deal with more complicated problem setting beyond convex  $\mathcal{U}_+$  and  $\mathcal{U}_-$  by using kernelization techniques.

### 2.2. Properties of RCM

To geometrically interpret RCM (2), Figure 1 shows the ellipsoidal uncertainty sets  $\mathcal{U}_+$ ,  $\mathcal{U}_-$  and their Minkowski difference. We can separate the problem (2) into two cases, i.e., whether  $\mathcal{U}_+$  and  $\mathcal{U}_-$  have an intersection or not, which is equivalent to whether  $\mathcal{U}$  includes  $\mathbf{0}$  or not. As shown in Theorem 2.2, there is a large difference in computational effort between the two cases. Before giving an intuitive geometric interpretation of RCM in Theorem 2.2, we introduce Lemma 2.1 that further separates the case  $\mathbf{0} \in \mathcal{U}$  into two cases:  $\mathcal{U}$  includes  $\mathbf{0}$  in its interior,  $\text{int}(\mathcal{U})$ , or on its boundary,  $\text{bd}(\mathcal{U})$ . In the geometric sense,  $\mathbf{0} \notin \mathcal{U}$  holds when  $\mathcal{U}_+$  and  $\mathcal{U}_-$  are disjoint.  $\mathbf{0} \in \mathcal{U}$  implies that  $\mathcal{U}_+$

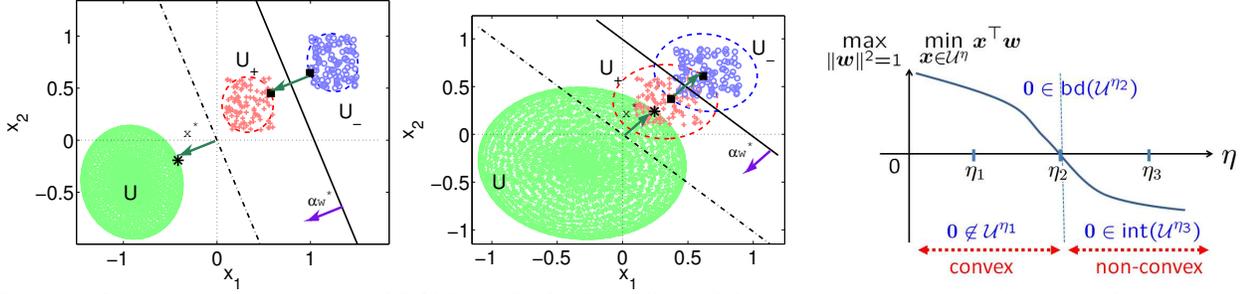


Figure 1. Geometric interpretation of RCM (2). **Left:** Two ellipsoidal uncertainty sets,  $\mathcal{U}_+$  and  $\mathcal{U}_-$ , are disjoint ( $\mathbf{0} \notin \mathcal{U}$ ). **Middle:**  $\mathcal{U}_+$  and  $\mathcal{U}_-$  are joint ( $\mathbf{0} \in \mathcal{U}$ ). The asterisk shows the optimal point  $\mathbf{x}^*$  in  $\mathcal{U}$ , and the dash-dot line depicts the hyperplane  $\mathbf{x}^{*\top} \mathbf{w}^* = 0$ . The squares are the solutions  $\mathbf{x}_+^*$  and  $\mathbf{x}_-^*$  of the inner-minimization on  $\mathcal{U}_+ \times \mathcal{U}_-$ , and the solid line stands for the optimal hyperplane,  $(\mathbf{x}_+^* - \mathbf{x}_-^*)^\top \mathbf{w}^* + b = 0$ . The bias term  $b$  in the decision function is defined such that the decision boundary passes through the mid-point of the squares. The green arrows indicate the optimal solution  $\mathbf{x}^*$ , and the purple arrows indicate the normal direction  $\alpha \mathbf{w}^*$  of the hyperplane for some positive  $\alpha$ . **Right:** Optimal value of RCM (2) with uncertainty set  $\mathcal{U}^\eta$ .

and  $\mathcal{U}_-$  are joint. In particular,  $\mathbf{0} \in \text{bd}(\mathcal{U})$  implies that  $\mathcal{U}_+$  and  $\mathcal{U}_-$  touch externally.

**Lemma 2.1.** *The optimal value of RCM (2) is positive if and only if  $\mathbf{0} \notin \mathcal{U}$ . It is zero if and only if  $\mathbf{0} \in \text{bd}(\mathcal{U})$ , and it is negative if and only if  $\mathbf{0} \in \text{int}(\mathcal{U})$ .*

We can prove “if” parts by using the supporting hyperplane theorem to three cases ( $\mathbf{0} \notin \mathcal{U}$ ,  $\mathbf{0} \in \text{bd}(\mathcal{U})$  and  $\mathbf{0} \in \text{int}(\mathcal{U})$ ). By taking the contrapositive of all the “if” parts, we also can prove “only if” parts.

Let  $\mathcal{U}^\eta$  be a parametrized uncertainty set for RCM (2) such that  $\mathcal{U}^{\eta_1} \subset \mathcal{U}^{\eta_2}$  holds for  $\eta_1 \leq \eta_2$ . Then the following inequality holds:

$$\max_{\mathbf{w}: \|\mathbf{w}\|^2=1} \min_{\mathbf{x} \in \mathcal{U}^{\eta_1}} \mathbf{x}^\top \mathbf{w} \geq \max_{\mathbf{w}: \|\mathbf{w}\|^2=1} \min_{\mathbf{x} \in \mathcal{U}^{\eta_2}} \mathbf{x}^\top \mathbf{w}.$$

This indicates that the optimal value of (2) is non-increasing with respect to the inclusion relation of uncertainty sets. Figure 1 (right) plots the non-increasing optimal value of RCM (2) with respect to  $\eta$ . An uncertainty set  $\mathcal{U}^{\eta_2}$  might exist such that the optimal value of (2) becomes zero.

The following theorem shows that when  $\mathbf{0} \notin \mathcal{U}$ , the equality constraint  $\|\mathbf{w}\|^2 = 1$  in (2) can be replaced by  $\|\mathbf{w}\|^2 \leq 1$  without changing the solution. Moreover,  $\|\mathbf{w}\|^2 = 1$  can be replaced by  $\|\mathbf{w}\|^2 \geq 1$  when  $\mathbf{0} \in \mathcal{U}$ . Figure 1 (left and middle) illustrates Theorem 2.2.

**Theorem 2.2.** *For an uncertainty set such that  $\mathbf{0} \notin \mathcal{U}$ , RCM (2) is equivalent to*

$$\max_{\mathbf{w}: \|\mathbf{w}\|^2 \leq 1} \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}. \quad (3)$$

Moreover, the problem is equivalent to

$$\min_{\mathbf{x}_\pm \in \mathcal{U}_\pm} \|\mathbf{x}_+ - \mathbf{x}_-\|, \quad \text{or equivalently,} \quad \min_{\mathbf{x} \in \mathcal{U}} \|\mathbf{x}\|. \quad (4)$$

An optimal  $\mathbf{w}$  of (3) can be obtained from  $\mathbf{x}^*/\|\mathbf{x}^*\|$  by using the optimal  $\mathbf{x}^* \in \mathcal{U}$  of (4). For an uncertainty set such that  $\mathbf{0} \in \text{int}(\mathcal{U})$ , RCM (2) is equivalent to

$$\max_{\mathbf{w}: \|\mathbf{w}\|^2 \geq 1} \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}. \quad (5)$$

Moreover, the problem is equivalent to  $\min_{\mathbf{x} \in \overline{\mathcal{U}^c}} \|\mathbf{x}\|$ , where  $\overline{\mathcal{U}^c}$  is the closure of the complement of the convex set  $\mathcal{U}$ . An optimal  $\mathbf{w}$  of (5) can be obtained from  $-\mathbf{x}^*/\|\mathbf{x}^*\|$  by using the optimal  $\mathbf{x}^* \in \overline{\mathcal{U}^c}$ .

*Proof.* Assume  $\mathbf{0} \notin \mathcal{U}$ . By applying the discussion on the minimum norm duality (Luenberger, 1969) to (3), we can confirm the equivalence of (3) and  $\min_{\mathbf{x} \in \mathcal{U}} \|\mathbf{x}\|$ , and the optimal solution  $\mathbf{w}^* = \mathbf{x}^*/\|\mathbf{x}^*\|$ . On the other hand, in the case of  $\mathbf{0} \in \text{int}(\mathcal{U})$ , the equivalence of (5) and  $\min_{\mathbf{x} \in \overline{\mathcal{U}^c}} \|\mathbf{x}\|$  is proved from Proposition 3.1 of (Briec, 1997) under the assumption that a convex  $\mathcal{U}$  has a nonempty interior. Hence, it is enough to show that there exists an optimal solution  $\mathbf{w}^*$  of (3) (or (5)) such that  $\|\mathbf{w}^*\| = 1$ , because the difference between (2) and (3) (or (5)) is only the norm constraint of  $\mathbf{w}$ .

Lemma 2.1 ensures that the optimal value of (3) is positive, because

$$\max_{\mathbf{w}: \|\mathbf{w}\|^2 \leq 1} \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w} \geq \max_{\mathbf{w}: \|\mathbf{w}\|^2=1} \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w} > 0.$$

Since the optimal solution  $\mathbf{w}^*$  of (3) satisfies  $0 < \|\mathbf{w}^*\| \leq 1$ , the following inequalities hold:

$$0 < \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}^* \leq \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}^* / \|\mathbf{w}^*\| \leq \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}^*.$$

The last inequality comes from the optimality of  $\mathbf{w}^*$ . These inequalities imply that  $\mathbf{w}^*/\|\mathbf{w}^*\|$  is also an optimal solution of (3) and that  $\|\mathbf{w}^*\| = 1$ . For the case of  $\mathbf{0} \in \text{int}(\mathcal{U})$ , we can similarly show that the optimal

Table 1. Correspondence with existing classifiers

$\mathcal{U}$	$\mathbf{0} \in \text{int}(\mathcal{U})$	$\mathbf{0} \in \text{bd}(\mathcal{U})$	$\mathbf{0} \notin \mathcal{U}$
Ellip. (11)	✓	MPM	MM-MPM
Ellip. (15)	✓	FDA	FS-FD
RCH (8)	E $\nu$ -SVM	$\nu_{\min}$	$\nu$ -SVM
CH (6)	×	×	HM-SVM

value of (5) is negative and that an optimal solution  $\mathbf{w}^*$  of (5) exists such that  $\|\mathbf{w}^*\| = 1$ .  $\square$

For  $\mathbf{0} \in \text{int}(\mathcal{U})$ , RCM (2) is essentially a non-convex problem, and we need to use non-convex optimization methods to solve it. Section 5 describes an optimization algorithm for non-convex problems of (2).

### 3. Equivalence to Existing Classifiers

We will show that RCM can be reduced to support vector machine (SVM), minimax probability machine (MPM), or Fisher discriminant analysis (FDA) depending on the prescribed uncertainty set  $\mathcal{U}$ . In Table 1, “ $\times$ ” means that the corresponding cases never happen. “ $\checkmark$ ” means that there are no corresponding existing models as far as we know. The models indicated by  $\checkmark$  are the target in this paper.

We denote an optimal solution of (2) as  $\mathbf{w}^*$  and define the bias term  $b$  such that the decision boundary passes through the mid-point of  $\mathbf{x}_+^*$  and  $\mathbf{x}_-^*$ , i.e.,  $b = -(\mathbf{x}_+^* + \mathbf{x}_-^*)^\top \mathbf{w}^*/2$ . Here,  $\mathbf{x}_+^* \in \mathcal{U}_+$  and  $\mathbf{x}_-^* \in \mathcal{U}_-$  stand for the optimal solutions of the inner-minimization in (2) for  $\mathbf{w} = \mathbf{w}^*$ .

#### 3.1. Hard-Margin SVM, $\nu$ -SVM and E $\nu$ -SVM

Whenever a data set is linearly separable, there are many hyperplanes that correctly classify all training samples. Vapnik-Chervonenkis theory indicates that a large margin classifier has a small generalization error. The problem can be transformed into a quadratic programming problem and the classification method is called hard-margin support vector classification machine (HM-SVM). Here, we define the uncertainty set (convex hull, CH) as follows:

$$\mathcal{U}_{\pm} = \text{conv}\{\mathbf{x}_i \mid i \in M_{\pm}\}, \quad (6)$$

where  $\text{conv}$  means convex hull. By using the Wolfe duality, the equivalence of HM-SVM and RCM (4) is obvious for  $\mathcal{U}_+ \cap \mathcal{U}_- = \emptyset$ .

HM-SVM has been extended to cope with non-separable data.  $C$ -SVM (Cortes & Vapnik, 1995) and  $\nu$ -SVM (Schölkopf et al., 2000) are typical examples

of “soft-margin” SVMs. There is a correspondence between  $C$ -SVM and  $\nu$ -SVM. That is, the classifier estimated by  $C$ -SVM with  $C \in (0, \infty)$  can be obtained from  $\nu$ -SVM with a parameter  $\nu \in (\nu_{\min}, \nu_{\max}] \subset [0, 1]$ , and vice versa. Crisp and Burges (2000) showed  $\nu_{\max} = 2 \min\{m_+, m_-\}/m$  and gave a geometric interpretation for  $\nu_{\min}$ . For  $\nu \in (\nu_{\max}, 1]$ , the optimization problem of  $\nu$ -SVM is unbounded, and for  $\nu \in [0, \nu_{\min})$ ,  $\nu$ -SVM provides a trivial solution ( $\mathbf{w} = \mathbf{0}$  and  $b = 0$ ). Perez-Cruz et al. (2003) devised extended  $\nu$ -SVM (E $\nu$ -SVM) as a way of avoiding such a trivial solution:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \rho} \quad & -\nu\rho + \frac{1}{m} \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{x}_i^\top \mathbf{w} + b) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i \in M, \quad \|\mathbf{w}\|^2 = 1. \end{aligned} \quad (7)$$

By forcing the norm of  $\mathbf{w}$  to be unity, a non-trivial and meaningful solution is obtained for any  $\nu \in [0, \nu_{\min})$ , but this comes at the expense of convexity. It furthermore provides the same solution as  $\nu$ -SVM for other values of  $\nu$ . In that sense, E $\nu$ -SVM can be regarded as an extension of  $\nu$ -SVM. It was experimentally found in (Perez-Cruz et al., 2003) that E $\nu$ -SVM often has better generalization performance than  $\nu$ -SVM.

In order to connect (E) $\nu$ -SVM with RCM, we define  $\mathcal{U}_{\pm}^{\nu}$  as

$$\left\{ \sum_{i \in M_{\pm}} \lambda_i \mathbf{x}_i \mid \sum_{i \in M_{\pm}} \lambda_i = 1, \quad 0 \leq \lambda_i \leq \frac{2}{\nu m}, \quad i \in M_{\pm} \right\}. \quad (8)$$

The set (8) is essentially equal to a *reduced convex hull* (RCH) (Bennett & Bredensteiner, 2000) or *soft convex hull* (Crisp & Burges, 2000). For linearly non-separable data set,  $\mathcal{U}_+^{\nu}$  and  $\mathcal{U}_-^{\nu}$  intersect with small  $\nu$ .

Crisp and Burges (2000) showed that  $\nu_{\min}$  is the largest  $\nu$  such that two RCHs,  $\mathcal{U}_+^{\nu}$  and  $\mathcal{U}_-^{\nu}$ , intersect. The model that finds  $\nu_{\min}$  corresponds to the case of  $\mathbf{0} \in \text{bd}(\mathcal{U}^{\nu})$  in the “RCH” of Table 1. Barbero et al. (2012) transformed  $\nu$ -SVM and E $\nu$ -SVM (7) into RCM (2) with  $\mathcal{U}_{\pm}^{\nu}$  in order to give them a geometric interpretation. Using the results, we can relate  $\nu$ -SVM, E $\nu$ -SVM, and RCM (2) as shown in Table 1.

#### 3.2. Minimax Probability Machine and Its Extension

The minimax probability machine (MPM) only uses the mean and covariance matrix of each class for classification tasks (Lanckriet et al., 2002). Suppose that  $\mathbf{x}_+$  (or  $\mathbf{x}_-$ ) is a  $d$ -dimensional random vector with mean  $\bar{\mathbf{x}}_+$  (or  $\bar{\mathbf{x}}_-$ ) and covariance  $\Sigma_+$  (or  $\Sigma_-$ ). We assume that  $\bar{\mathbf{x}}_+ \neq \bar{\mathbf{x}}_-$  and that  $\Sigma_{\pm}$  are positive definite. The MPM minimizes the misclassification probabili-

ties under the worst-case setting as follows:

$$\max_{\alpha, \mathbf{w}, b} \alpha \quad \text{s.t.} \quad \inf_{\mathbf{x}_{\pm} \sim (\bar{\mathbf{x}}_{\pm}, \Sigma_{\pm})} \Pr\{\mathbf{x}_{\pm}^{\top} \mathbf{w} + b \geq 0\} \geq \alpha, \quad (9)$$

where  $\mathbf{x}_{+} \sim (\bar{\mathbf{x}}_{+}, \Sigma_{+})$  refers to the class of distributions that have mean  $\bar{\mathbf{x}}_{+}$  and covariance  $\Sigma_{+}$ , but are otherwise arbitrary; likewise for  $\mathbf{x}_{-}$ . In practice, the mean vectors and covariance matrices of each class are estimated from the training samples.

Lanckriet et al. (2002) represented problem (9) as a convex optimization problem known as a second-order cone program (SOCP) and show the dual form:

$$\min_{\kappa} \kappa \quad \text{s.t.} \quad \mathbf{0} \in \mathcal{U}^{\kappa} := \mathcal{U}_{+}^{\kappa} \ominus \mathcal{U}_{-}^{\kappa}, \quad (10)$$

$$\text{where } \mathcal{U}_{\pm}^{\kappa} = \{\bar{\mathbf{x}}_{\pm} + \Sigma_{\pm}^{1/2} \mathbf{u} \mid \|\mathbf{u}\| \leq \kappa\}. \quad (11)$$

$\alpha$  of (9) corresponds to  $\kappa$  of (10) as  $\kappa = \sqrt{\alpha/(1-\alpha)}$ . Therefore, MPM (9) is the problem to find the smallest positive  $\kappa$  (denoted by  $\kappa_{\max}$ ) such that the two ellipsoids intersect, i.e.,  $\mathbf{0} \in \text{bd}(\mathcal{U}^{\kappa_{\max}})$ .

The idea of MPM is combined with the idea of the margin maximization in (Nath & Bhattacharyya, 2007). Given acceptable false positive and negative rates,  $\eta_{+}$  and  $\eta_{-}$ , the linear classifier can be estimated by

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad \sup_{\mathbf{x}_{\pm} \sim (\bar{\mathbf{x}}_{\pm}, \Sigma_{\pm})} \Pr\{\mathbf{x}_{\pm}^{\top} \mathbf{w} + b < 0\} \leq \eta_{\pm}. \quad (12)$$

In this paper, we call this model the ‘‘margin maximized MPM’’ (MM-MPM). In the same way as in MPM, (12) can be transformed into an SOCP.

Robust optimization techniques for ellipsoidal uncertainty (Ben-Tal et al., 2009) transform RCM (2) with  $\mathcal{U}_{\pm} = \mathcal{U}_{\pm}^{\kappa_{\pm}}$  into

$$\min_{\|\mathbf{w}\|^2=1} \kappa_{+} \|\Sigma_{+}^{1/2} \mathbf{w}\| + \kappa_{-} \|\Sigma_{-}^{1/2} \mathbf{w}\| - (\bar{\mathbf{x}}_{+} - \bar{\mathbf{x}}_{-})^{\top} \mathbf{w}. \quad (13)$$

We define  $\kappa_{+}^{\max}$  and  $\kappa_{-}^{\max}$  as constants such that  $\mathcal{U}_{+}^{\kappa_{+}^{\max}}$  and  $\mathcal{U}_{-}^{\kappa_{-}^{\max}}$  touch. For  $\kappa_{\pm} \in [0, \kappa_{\pm}^{\max}]$ ,  $\mathcal{U}_{+}^{\kappa_{+}} \cap \mathcal{U}_{-}^{\kappa_{-}} = \emptyset$  holds, and RCM (13) is equivalent to MM-MPM (12) with  $\kappa_{\pm} = \sqrt{(1-\eta_{\pm})/\eta_{\pm}}$ . We can confirm this by comparing the dual form of MM-MPM and the dual of (13), that is equivalent to (4). Furthermore, (13) with  $\kappa_{\pm} = \kappa_{\max}$  coincides with MPM (9) (see Table 1).

### 3.3. Fisher Discriminant Analysis and Its Extension

In Fisher discriminant analysis (FDA) as in MPM (9), a discriminant hyperplane is computed from the means and covariances of random vectors  $\mathbf{x}_{+}$  and  $\mathbf{x}_{-}$ . The hyperplane is determined from the optimal solution  $\mathbf{w}^*$  to the following problem (Fukunaga, 1990):

$$\max_{\mathbf{w}} \frac{(\bar{\mathbf{x}}_{+} - \bar{\mathbf{x}}_{-})^{\top} \mathbf{w}}{\|(\Sigma_{+} + \Sigma_{-})^{1/2} \mathbf{w}\|}. \quad (14)$$

The problem finds a direction which maximizes the projected class means while minimizing the class variance in this direction.

Likewise for MPM, FDA has a probabilistic interpretation under the worst-case scenario. Using the ellipsoidal uncertainty set defined by

$$\mathcal{U}^{\zeta} = \{\mathbf{x} = (\bar{\mathbf{x}}_{+} - \bar{\mathbf{x}}_{-}) + (\Sigma_{+} + \Sigma_{-})^{1/2} \mathbf{u} \mid \|\mathbf{u}\| \leq \zeta\}, \quad (15)$$

FDA (14) can be represented as

$$\min_{\zeta} \zeta \quad \text{s.t.} \quad \mathbf{0} \in \mathcal{U}^{\zeta}. \quad (16)$$

FDA can be extended to RCM (2) with the uncertainty set  $\mathcal{U}^{\zeta}$  for a prescribed parameter  $\zeta > 0$ . Let  $\zeta_{\max}$  be the optimal value of (16). Then, along the same lines as the MPM in Section 3.2, we find that RCM (2) with  $\mathcal{U} = \mathcal{U}^{\zeta_{\max}}$  is equivalent to FDA.

Indeed, RCM (2) with  $\mathcal{U}^{\zeta}$  is transformed into

$$\min_{\|\mathbf{w}\|^2=1} \zeta \|(\Sigma_{+} + \Sigma_{-})^{1/2} \mathbf{w}\| - (\bar{\mathbf{x}}_{+} - \bar{\mathbf{x}}_{-})^{\top} \mathbf{w}.$$

Especially for  $\zeta \in [0, \zeta_{\max}]$ , the norm constraint is replaced with the convex constraint  $\|\mathbf{w}\|^2 \leq 1$  without changing the optimal solution. Here, MM-FDA refers to this estimator. In replacing the Euclidean norm  $\|\mathbf{w}\|$  with the  $L_1$ -norm  $\|\mathbf{w}\|_1$ , MM-FDA is equivalent to a sparse feature selection model based on FDA (FS-FD) (Bhattacharyya, 2004).

## 4. Statistical Interpretation for RCM

We can give a statistical interpretation for RCM on the basis of statistical learning theory. Let us start by introducing a loss function  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  that defines the loss of the decision function  $\mathbf{x}^{\top} \mathbf{w} + b$  regarding the sample  $(\mathbf{x}, y)$  as  $\ell(y(\mathbf{x}^{\top} \mathbf{w} + b))$ .

A goal of the classification task is to obtain an accurate classifier. For this purpose, it is reasonable to minimize the expected loss,  $\mathbb{E}[\ell(y(\mathbf{x}^{\top} \mathbf{w} + b))]$ , with respect to  $\mathbf{w}$  and  $b$ . Let us define  $p(\mathbf{x}|y)$  as the conditional probability density of  $\mathbf{x}$ , given the binary label  $y$ , and  $\pi_{+}$  and  $\pi_{-}$  as the marginal probabilities of the positive and negative labels, respectively.  $\mathbb{E}[\ell(y(\mathbf{x}^{\top} \mathbf{w} + b))]$  is computed by

$$\pi_{+} \int \ell(\mathbf{x}^{\top} \mathbf{w} + b) p(\mathbf{x}|+1) d\mathbf{x} + \pi_{-} \int \ell(-(\mathbf{x}^{\top} \mathbf{w} + b)) p(\mathbf{x}|-1) d\mathbf{x}.$$

Since the true probability distribution is unknown, we cannot minimize the expected loss directly.

Now let us consider the ambiguity of the probability distribution  $p(\mathbf{x}|y)$ . Let  $\mathcal{P}_+$  and  $\mathcal{P}_-$  be sets of probability densities. Each set of probabilities expresses the uncertainty of the conditional probabilities  $p(\mathbf{x}|+1)$  and  $p(\mathbf{x}|-1)$ , respectively. We can use the min-max decision rule for the uncertainty of  $p(\mathbf{x}|y)$  as follows:

$$\min_{\mathbf{w}: \|\mathbf{w}\|^2=1} \max_{p(\mathbf{x}|\pm 1) \in \mathcal{P}_{\pm}} \min_{b \in \mathbb{R}} \mathbb{E}[\ell(y(\mathbf{x}^{\top} \mathbf{w} + b))]. \quad (17)$$

The worst-case minimization problem is difficult to solve. Therefore, we propose to solve RCM (2), since we can prove that RCM (2) is a good approximation for minimizing the worst-case expected loss.

To relate (17) and RCM, we firstly give an equivalent formulation for RCM. Here, we define  $\mathbf{x}_+$  and  $\mathbf{x}_-$  as the mean of the input vector  $\mathbf{x}$  under the conditional probabilities  $p(\mathbf{x}|+1)$  and  $p(\mathbf{x}|-1)$ , respectively, i.e.,  $\mathbf{x}_{\pm} = \int \mathbf{x}p(\mathbf{x}|\pm 1)d\mathbf{x}$ . Here, we assume that all probability distributions in  $\mathcal{P}_{\pm}$  have the mean vector. Let  $\mathcal{U}_+$  and  $\mathcal{U}_-$  be

$$\mathcal{U}_{\pm} = \left\{ \int \mathbf{x}p(\mathbf{x}|\pm 1)d\mathbf{x} \mid p(\mathbf{x}|\pm 1) \in \mathcal{P}_{\pm} \right\}. \quad (18)$$

Suppose that the uncertainty sets of probability densities,  $\mathcal{P}_{\pm}$ , are both convex; i.e., a mixture of two probability densities also lies in the uncertainty set. Then  $\mathcal{U}_+$  and  $\mathcal{U}_-$  are convex sets.

**Theorem 4.1.** *Suppose that  $\ell(z)$  is a non-increasing function. An optimal solution of the RCM with the uncertainty sets  $\mathcal{U}_+$  and  $\mathcal{U}_-$  in (18) is also optimal to*

$$\min_{\mathbf{w}: \|\mathbf{w}\|^2=1} \max_{\mathbf{x}_{\pm} \in \mathcal{U}_{\pm}} \min_{b \in \mathbb{R}} J_{\ell}(\mathbf{w}, b; \mathbf{x}_+, \mathbf{x}_-), \quad (19)$$

where

$$J_{\ell}(\mathbf{w}, b; \mathbf{x}_+, \mathbf{x}_-) = \pi_+ \ell(\mathbf{x}_+^{\top} \mathbf{w} + b) + \pi_- \ell(-\mathbf{x}_-^{\top} \mathbf{w} - b).$$

*Proof.* For a fixed  $\mathbf{w}$  and  $\mathbf{x}_{\pm} \in \mathcal{U}_{\pm}$ , minimizing  $J_{\ell}(\mathbf{w}, b; \mathbf{x}_+, \mathbf{x}_-)$  respect to  $b$  is equivalent to

$$\min_{b'} \pi_+ \ell((\mathbf{x}_+ - \mathbf{x}_-)^{\top} \mathbf{w} - b') + \pi_- \ell(b').$$

Since the objective function above is non-increasing in  $(\mathbf{x}_+ - \mathbf{x}_-)^{\top} \mathbf{w}$ , there exists a non-increasing function  $\phi(z)$  such that

$$\phi((\mathbf{x}_+ - \mathbf{x}_-)^{\top} \mathbf{w}) = \min_b J_{\ell}(\mathbf{w}, b; \mathbf{x}_+, \mathbf{x}_-).$$

Hence, one has

$$\begin{aligned} & \min_{\mathbf{w}: \|\mathbf{w}\|^2=1} \max_{\mathbf{x}_{\pm} \in \mathcal{U}_{\pm}} \min_b J_{\ell}(\mathbf{w}, b; \mathbf{x}_+, \mathbf{x}_-) \\ &= \phi\left( \max_{\mathbf{w}: \|\mathbf{w}\|^2=1} \min_{\mathbf{x}_{\pm} \in \mathcal{U}_{\pm}} (\mathbf{x}_+ - \mathbf{x}_-)^{\top} \mathbf{w} \right). \end{aligned}$$

As a result, the optimal solution of the RCM is also optimal for problem (19).  $\square$

**Theorem 4.2.** *We assume that i)  $\ell(z)$  is convex, decreasing, and second-order differentiable, and that ii)  $0 \leq \ell''(z) \leq L \in \mathbb{R}$  holds for all  $z$ . Suppose that  $\mathbf{x}_{\pm} \in \mathcal{U}_{\pm}$  is in the ball with the radius  $c$ , i.e.,  $\|\mathbf{x}_{\pm}\| \leq c$ . Then, for the optimal value  $J^*$  of (19), one has*

$$\min_{\mathbf{w}; \|\mathbf{w}\|^2=1} \max_{p(\mathbf{x}|\pm 1) \in \mathcal{P}_{\pm}} \min_{b \in \mathbb{R}} \mathbb{E}[\ell(y(\mathbf{x}^{\top} \mathbf{w} + b))] \in [J^*, J^* + \frac{Lc^2}{2}].$$

*Proof.* The convexity of  $\ell(z)$  leads to a lower bound,  $J_{\ell}(\mathbf{w}, b; \mathbf{x}_+, \mathbf{x}_-)$ , and the Taylor expansions of  $\ell(z)$  around  $z = \mathbf{x}_+^{\top} \mathbf{w} + b$  and  $z = \mathbf{x}_-^{\top} \mathbf{w} + b$  yield an upper bound,  $J_{\ell}(\mathbf{w}, b; \mathbf{x}_+, \mathbf{x}_-) + \frac{Lc^2}{2}$ , of  $\mathbb{E}[\ell(y(\mathbf{x}^{\top} \mathbf{w} + b))]$ . Even when the min-max operation is applied,  $J^*$  and  $J^* + \frac{Lc^2}{2}$  remain bounds for the worst-case expected loss (17).  $\square$

The theorem implies that problem (19) minimizes the bounds of (17). Noticing that the optimal solution of problem (19) is available by solving RCM as shown in Theorem 4.1, Theorem 4.2 implies that RCM minimizes the upper and lower bounds of the worst-case expected loss (17) at the same time.

There are various ways to estimate the bias term  $b$  for RCM. The simplest way is to use  $b^* = -(\mathbf{x}_+^* + \mathbf{x}_-^*)^{\top} \mathbf{w}^*/2$ . Another promising method is to construct an appropriate statistical model for the projected samples  $(\mathbf{x}_i^{\top} \mathbf{w}^*, y_i)$ ,  $i \in M$ . The projected samples,  $\mathbf{x}_i^{\top} \mathbf{w}^*$ ,  $i \in M$ , are scattered in one-dimensional space, from which we can estimate  $b$  on the basis of the statistical model.

## 5. Solution Method for RCM

The RCM has a significantly larger range of parameter  $\kappa$  or  $\zeta$  than an existing convex model such as MPM, MM-MPM, FDA or FS-FD (see Table 1). Therefore, the RCM enhances a possibility of improving these existing classification models. Indeed, Perez-Cruz et al. (2003) experimentally showed that the generalization performance of  $\nu$ -SVM is often better than that of original  $\nu$ -SVM. In this section, we propose a solution method that is generalized from the local algorithms of (Perez-Cruz et al., 2003; Takeda & Sugiyama, 2008).

### 5.1. Two-stage Optimization Strategy

Suppose that we solve RCM (2) with the uncertainty set  $\mathcal{U}^{\eta}$  with one parameter  $\eta$  and that  $\mathcal{U}^{\eta_1} \subset \text{int}(\mathcal{U}^{\eta_2})$  holds for  $\eta_1 < \eta_2$ . Let us define  $\eta_{\max}$  such that the optimal value of (2) with  $\mathcal{U} = \mathcal{U}^{\eta_{\max}}$  is zero.

First, we need to compute  $\eta_{\max}$  in order to confirm that the given problem (2) is essentially convex or not.

**Algorithm 5.1.**

**Step 1:** Choose  $\tilde{\mathbf{w}}_0$  satisfying  $\|\tilde{\mathbf{w}}_0\| = 1$  and  $\epsilon > 0$ .  
Let  $t \leftarrow 0$ .

**Step 2:** Solve the following program:

$$\max_{\mathbf{w}} g(\mathbf{w}) \quad \text{s.t.} \quad \tilde{\mathbf{w}}_t^\top \mathbf{w} = 1, \quad (21)$$

where  $g(\mathbf{w}) = \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}$ , and let the optimal solution be  $\hat{\mathbf{w}}_t^*$ .

**Step 3:** If  $\|\tilde{\mathbf{w}}_t - \hat{\mathbf{w}}_t^*\| \leq \epsilon$ , terminate and output  $\tilde{\mathbf{w}}^* \leftarrow \tilde{\mathbf{w}}_t$ .

**Step 4:** Otherwise, let  $\tilde{\mathbf{w}}_{t+1} \leftarrow \hat{\mathbf{w}}_t^* / \|\hat{\mathbf{w}}_t^*\|$ . Let  $t \leftarrow t + 1$ . Repeat Steps 2–4.

Figure 2. Local optimum search algorithm for non-convex RCM with  $\eta > \eta_{\max}$ , that is,  $\mathbf{0} \in \text{int}(\mathcal{U})$ .

The parameter  $\eta_{\max}$  is obtained as the optimal solution of the convex problem:

$$\min_{\eta} \eta \quad \text{s.t.} \quad \mathbf{0} \in \mathcal{U}^\eta. \quad (20)$$

When  $\mathcal{U}_\pm^\eta$  are ellipsoidal sets of (11) (or (15)), the problem reduces to MPM (10) (or FDA (16)). When  $\mathcal{U}_\pm^\eta$  are RCHs, the problem reduces to a linear programming problem and gives us  $\nu_{\min}$ .

If the input parameter  $\eta$  is equal to  $\eta_{\max}$ , we have already obtained an optimal solution from (20). If  $\eta < \eta_{\max}$ , we next solve the convex problem (4) by using a standard optimization software.

## 5.2. Local Optimization Algorithm for Non-convex RCM

For  $\eta > \eta_{\max}$ , RCM (2) is essentially equivalent to (5) that includes a non-convex constraint,  $\|\mathbf{w}\|^2 \geq 1$ . We next need to solve (2) as a non-convex problem.

In the area of global optimization, non-convex RCM (5) (precisely, a problem constructed by taking dual for the inner-minimization in (5)) is known as a *reverse convex program* (RCP), or *canonical d.c. programming*. This differs from a conventional convex program only by the presence of a reverse convex constraint ( $\|\mathbf{w}\|^2 \geq 1$  in the current case). When all functions are linear except for the reverse convex constraint, the RCP problem is especially called *linear reverse convex program* (LRCP).  $E\nu$ -SVM is an LRCP, for which Perez-Cruz et al. (2003) proposed a local optimum search algorithm and Takeda and Sugiyama (2008) proposed a global optimum search algorithm.

Here, we show a local optimum search algorithm (Algorithm 5.1) that is generalized from the local algorithms (Perez-Cruz et al., 2003; Takeda & Sugiyama,

2008) of  $E\nu$ -SVM for non-convex RCM. It is essentially the same as the local algorithm, Algorithm 7, in (Takeda & Sugiyama, 2008) when  $\mathcal{U}$  of  $g(\mathbf{w})$  is an RCH (8) and  $\epsilon = 0$ .

RCM (2) requires maximizing  $g(\mathbf{w}) = \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}$  subject to a non-convex constraint,  $\mathbf{w}^\top \mathbf{w} = 1$ . Instead of solving the non-convex problem directly, we can iteratively solve the relaxation problems (21) (in Algorithm 5.1). Since  $g(\mathbf{w})$  is concave, (21) can be solved by using convex minimization techniques.

The non-convex constraint of (2) is linearized at  $\tilde{\mathbf{w}}_t$  in the algorithm, and the linear constraint  $\tilde{\mathbf{w}}_t^\top \mathbf{w} = 1$  is updated every iteration. Note that the negativity of the optimal value of (21) is guaranteed because of  $\mathbf{0} \in \text{int}(\mathcal{U})$ . As the algorithm proceeds, the solutions  $\tilde{\mathbf{w}}_t$  improve, i.e.,

$$g(\tilde{\mathbf{w}}_t) \leq g(\hat{\mathbf{w}}_t^*) < g(\hat{\mathbf{w}}_t^* / \|\hat{\mathbf{w}}_t^*\|) = g(\tilde{\mathbf{w}}_{t+1}) < 0, \quad (22)$$

because  $\tilde{\mathbf{w}}_t^\top \hat{\mathbf{w}}_t^* = 1$  together with  $\tilde{\mathbf{w}}_t^\top \tilde{\mathbf{w}}_t = 1$  implies  $\|\hat{\mathbf{w}}_t^*\| > 1$ . Note that  $\tilde{\mathbf{w}}_t$  is a feasible solution for (21). Hence, if (21) has no better solutions than  $\tilde{\mathbf{w}}_t$ ,  $\tilde{\mathbf{w}}_t$  is returned as an optimal solution  $\hat{\mathbf{w}}_t^*$  of (21). The algorithm terminates after that.

The computation of  $g(\mathbf{w})$  may be difficult for general uncertainty sets. However, we do not need an explicit formula for  $g(\mathbf{w})$  in (21). If  $\mathcal{U}$  is a convex set, we can obtain a dual formulation (max-problem) for  $g(\mathbf{w})$  and replace the max-min problem (21) with a simple max-problem, that is, a one-level convex problem. Indeed, when the uncertainty set is an RCH (8) of data points, we can take the dual for  $g(\mathbf{w}) = \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \mathbf{w}$  and change (21) into a linearized  $E\nu$ -SVM (7) whose constraint is  $\tilde{\mathbf{w}}_t^\top \mathbf{w} = 1$  instead of  $\|\mathbf{w}\|^2 = 1$ . When the algorithm is applied to the RCM having ellipsoidal uncertainty, we analytically obtain the optimal value  $g(\mathbf{w})$  for any  $\mathbf{w}$ . Indeed, for ellipsoidal uncertainty (11),  $g(\mathbf{w})$  is equal to the one derived by multiplying the objective function of (13) by -1.

**Theorem 5.2.** *For any  $\epsilon > 0$ , Algorithm 5.1 terminates in a finite number of iterations.*

*Proof.* Let the negative value  $g_{\text{opt}}$  be the optimal value of RCM (2). Suppose  $\|\hat{\mathbf{w}}_t^*\| > 1$  for all  $t = 1, 2, \dots$ . Otherwise, the algorithm terminates. By evaluating  $g(\tilde{\mathbf{w}}_{t+1}) - g(\tilde{\mathbf{w}}_t)$ , we have

$$\sum_{t=0}^{\infty} (g(\tilde{\mathbf{w}}_{t+1}) - g(\tilde{\mathbf{w}}_t)) \geq \sum_{t=0}^{\infty} \left( \frac{1}{\|\hat{\mathbf{w}}_t^*\|} - 1 \right) g_{\text{opt}} > 0.$$

The above inequality and the boundedness of  $g(\tilde{\mathbf{w}}_t)$  lead to  $\lim_{t \rightarrow \infty} \frac{1}{\|\hat{\mathbf{w}}_t^*\|} - 1 = 0$ . Therefore,  $\gamma_t$  exists such that  $\|\hat{\mathbf{w}}_t^*\| = 1 + \gamma_t$ ,  $0 < \gamma_t = o(1)$ , that leads

to  $\|\tilde{\mathbf{w}}_t - \hat{\mathbf{w}}_t^*\| = \sqrt{2\gamma_t + \gamma_t^2}$ . Since  $\gamma_t \rightarrow 0$  holds, the stopping rule  $\|\tilde{\mathbf{w}}_t - \hat{\mathbf{w}}_t^*\| \leq \epsilon$  with positive  $\epsilon$  is satisfied in a finite number of iterations.  $\square$

We can show that Algorithm 5.1 with  $\epsilon = 0$  terminates within a finite number of iterations when the uncertainty set of RCM (2) is represented by a convex polyhedron by mimicking the proof of Theorem 8 for  $E\nu$ -SVM in (Takeda & Sugiyama, 2008).

Here, suppose that  $g(\tilde{\mathbf{w}}^*)$  is differentiable, i.e.,  $g(\mathbf{w})$  has a unique subgradient at  $\tilde{\mathbf{w}}^*$  as  $\partial g(\tilde{\mathbf{w}}^*) = \arg \min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \tilde{\mathbf{w}}^* = \{\nabla g(\tilde{\mathbf{w}}^*)\}$ . For example,  $g(\mathbf{w})$  is differentiable under ellipsoidal uncertainty (11) (or (15)). Then Theorem 5.3 shows a sufficient condition for the local optimality of the solution  $\tilde{\mathbf{w}}^*$  if it is obtained by Algorithm 5.1 with  $\epsilon = 0$ .

**Theorem 5.3.** *Suppose that  $g(\tilde{\mathbf{w}}^*)$  is differentiable. Algorithm 5.1 that terminates with  $\epsilon = 0$  provides a local solution  $\tilde{\mathbf{w}}^*$  to RCM (2) when the maximum eigenvalue of  $\nabla^2 g(\tilde{\mathbf{w}}^*)$  is less than  $g(\tilde{\mathbf{w}}^*)$ .*

*Proof.* Note that  $\tilde{\mathbf{w}}^*$  is the optimal solution of (21) at the final iteration. Therefore,  $\tilde{\mathbf{w}}^*$  satisfies the first- and second-order necessary conditions:

$$\nabla g(\tilde{\mathbf{w}}^*) + \eta \tilde{\mathbf{w}}^* = \mathbf{0}, \quad (23)$$

$$\mathbf{d}^\top \nabla^2 g(\tilde{\mathbf{w}}^*) \mathbf{d} \leq 0, \quad \forall \mathbf{d} \text{ such that } \tilde{\mathbf{w}}^{*\top} \mathbf{d} = 0, \quad (24)$$

where  $\eta$  is a Lagrange multiplier. We can show  $\eta = -g(\tilde{\mathbf{w}}^*) > 0$  by noticing that  $\nabla g(\tilde{\mathbf{w}}^*)$  is a minimizer to  $\min_{\mathbf{x} \in \mathcal{U}} \mathbf{x}^\top \tilde{\mathbf{w}}^* = g(\tilde{\mathbf{w}}^*)$  and using (23). When the maximum eigenvalue of  $\nabla^2 g(\tilde{\mathbf{w}}^*)$  is less than  $-\eta$ ,  $\tilde{\mathbf{w}}^*$  satisfying (23) and (24) also satisfies the second-order sufficient conditions for the local optimality of (2):

$$\begin{aligned} \nabla g(\mathbf{w}) + 2\zeta \mathbf{w} = \mathbf{0}, \quad \mathbf{d}^\top (\nabla^2 g(\mathbf{w}) + 2\zeta \mathbf{I}) \mathbf{d} < 0, \\ \forall \mathbf{d} \neq \mathbf{0} \text{ such that } \mathbf{w}^\top \mathbf{d} = 0, \end{aligned}$$

where  $\mathbf{I}$  is the identity matrix and  $\zeta (\geq 0)$  is a multiplier. This shows that  $\tilde{\mathbf{w}}^*$  is a local solution of (2).  $\square$

Theorem 5.3 may be extendable to the non-differentiable case of  $g(\tilde{\mathbf{w}}^*)$ , though more assumptions are necessary (see Theorems 3.2.16, 3.2.20 and 3.2.21 in (Polak, 1997)).

## 6. Conclusions

We developed the robust classification model (RCM), a model which includes SVM, MPM and FDA for specific uncertainty sets. The choice of uncertainty set is significant in this model. This model enables extensions and improvements to SVM to be applied to MPM and FDA, and vice versa.

The unified model will be of help in clarifying relationships among existing models and in finding new classifiers and new algorithms. That is, we might be able to devise a new classifier by finding a reasonable uncertainty set for RCM. It will be important to see how the learning algorithm, uncertainty set, and prediction accuracy relate to each other.

## References

- Barbero, Á., Takeda, A., and López, J. Geometric intuition and algorithms for  $E\nu$ -svm. *in preparation*, X(X):X–X, 2012.
- Ben-Tal, A., El-Ghaoui, L., and Nemirovski, A. *Robust Optimization*. Princeton University Press, Princeton, 2009.
- Bennett, K. P. and Bredensteiner, E. J. Duality and geometry in SVM classifiers. In *ICML*, pp. 57–64, 2000.
- Bhattacharyya, C. Second order cone programming formulations for feature selection. *Journal of Machine Learning Research*, 5:1417–1433, 2004.
- Briec, W. Minimum distance to the complement of a convex set: Duality result. *Journal of Optimization Theory and Applications*, 93(2):301–319, 1997.
- Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20:273–297, 1995.
- Crisp, D. J. and Burges, C. J. C. A geometric interpretation of  $\nu$ -SVM classifiers. In *NIPS*, pp. 244–250, 2000.
- Fukunaga, K. *Introduction to statistical pattern recognition*. Academic Press, Boston, 1990.
- Lanckriet, G. R. G., Ghaoui, L. El, Bhattacharyya, C., and Jordan, M. I. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.
- Luenberger, D.G. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, 1969.
- Nath, J. S. and Bhattacharyya, C. Maximum margin classifiers with specified false positive and false negative error rates. In *Proceedings of the seventh SIAM International Conference on Data mining*, pp. 35–46, 2007.
- Perez-Cruz, F., Weston, J., Hermann, D. J. L., and Schölkopf, B. Extension of the  $\nu$ -SVM range for classification. In *Advances in Learning Theory: Methods, Models and Applications 190*, pp. 179–196, 2003.
- Polak, E. *Optimization: Algorithms and Consistent Approximations*. Springer, New York, 1997.
- Schölkopf, B. and Smola, A. J. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- Schölkopf, B., Smola, A., Williamson, R., and Bartlett, P. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.
- Takeda, A. and Sugiyama, M.  $\nu$ -support vector machine as conditional value-at-risk minimization. In *ICML*, pp. 1056–1063, 2008.