entropy

Article

Breakdown Point of Robust Support Vector Machines

Takafumi Kanamori [1,4,*], Shuhei Fujiwara [2] and Akiko Takeda [3,4]

[1] Department of Computer Science and Mathematical Informatics, Nagoya University, Nagoya 464-8601, Japan
[2] TOPGATE Co. Ltd., Bunkyo-ku, Tokyo 113-0033, Japan; shuhei.fujiwara@gmail.com
[3] Institute of Statistical Mathematics, Tokyo 190-8562, Japan; atakeda@ism.ac.jp
[4] RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan
* Correspondence: kanamori@is.nagoya-u.ac.jp; Tel.: +81-52-789-4598

Academic Editor: Kevin H. Knuth
Received: 15 January 2017; Accepted: 16 February 2017; Published: 21 February 2017

Abstract: Support vector machine (SVM) is one of the most successful learning methods for solving classification problems. Despite its popularity, SVM has the serious drawback that it is sensitive to outliers in training samples. The penalty on misclassification is defined by a convex loss called the hinge loss, and the unboundedness of the convex loss causes the sensitivity to outliers. To deal with outliers, robust SVMs have been proposed by replacing the convex loss with a non-convex bounded loss called the ramp loss. In this paper, we study the breakdown point of robust SVMs. The breakdown point is a robustness measure that is the largest amount of contamination such that the estimated classifier still gives information about the non-contaminated data. The main contribution of this paper is to show an exact evaluation of the breakdown point of robust SVMs. For learning parameters such as the regularization parameter, we derive a simple formula that guarantees the robustness of the classifier. When the learning parameters are determined with a grid search using cross-validation, our formula works to reduce the number of candidate search points. Furthermore, the theoretical findings are confirmed in numerical experiments. We show that the statistical properties of robust SVMs are well explained by a theoretical analysis of the breakdown point.

Keywords: support vector machine; breakdown point; outlier; kernel function

1. Introduction

1.1. Background

Support vector machine (SVM) is a highly developed classification method that is widely used in real-world data analysis [1,2]. The most popular implementation is called C-SVM, which uses the maximum margin criterion with a penalty for misclassification. The positive parameter C tunes the balance between the maximum margin and penalty, and the resulting classification problem can be formulated as a convex quadratic problem based on training data. A separating hyper-plane for classification is obtained from the optimal solution of the problem. Furthermore, complex non-linear classifiers are obtained by using the reproducing kernel Hilbert space (RKHS) as a statistical model of the classifiers [3]. There are many variants of SVM for solving binary classification problems, such as ν-SVM, Eν-SVM and least squares SVM [4–6]. Moreover, the generalization ability of SVM has been analyzed in many studies [7–9].

In practical situations, however, SVM has drawbacks. The remarkable feature of the SVM is that the separating hyperplane is determined mainly from misclassified samples. Thus, the most misclassified samples significantly affect the classifier, meaning that the standard SVM is extremely susceptible to outliers. In C-SVM, the penalties of sample points are measured in terms of the hinge loss, which is a convex surrogate of the 0-1loss for misclassification. The convexity of the hinge loss causes SVM to be unstable in the presence of outliers, since the convex function is unbounded and puts an extremely large penalty on outliers. One way to remedy the instability is to replace the convex loss with a non-convex bounded loss to suppress outliers. Loss clipping is a simple method to obtain a bounded loss from a convex loss [10,11]. For example, clipping the hinge loss leads to the ramp loss [12,13], which is a loss function used in robust SVMs. Yu et al. [11,14] showed a convex loss clipping that yields a non-convex loss function and proposed a convex relaxation of the resulting non-convex optimization problem to obtain a computationally-efficient learning algorithm. The SVM using the ramp loss is regarded as a robust variant of L_1-SVM. Recently, Feng et al. [15] also proposed a robust variant of L_2-SVM.

1.2. Our Contribution

In this paper, we provide a detailed analysis on the robustness of SVMs. In particular, we deal with a robust variant of kernel-based ν-SVM. The standard ν-SVM [5] has a regularization parameter ν, and it is equivalent to C-SVM; i.e., both methods provide the same classifier for the same training data if the regularization parameters, ν and C, are properly tuned. We generate a robust variant of ν-SVM by clipping the loss function of ν-SVM, called robust (ν, μ)-SVM, with another learning parameter $\mu \in [0, 1)$. The parameter μ denotes the ratio of samples to be removed from the training dataset as outliers. When the ratio of outliers in the training dataset is bounded above by μ, robust (ν, μ)-SVM is expected to provide a robust classifier.

Robust (ν, μ)-SVM is closely related to other robust SVMs, such as CVaR-(α_L, α_U)-SVM [16], the robust outlier detection (ROD) algorithm [17] and extended robust SVM (ER-SVM) [18,19]. In particular, it is equivalent to the CVaR-(α_L, α_U)-SVM. In this paper, the learning algorithm we consider is referred to as robust (ν, μ)-SVM to emphasize that it is a robust variant of ν-SVM. On the other hand, ROD is to robust (ν, μ)-SVM what C-SVM is to ν-SVM. ER-SVM is another robust extension of ν-SVM, and it includes robust (ν, μ)-SVM as a special case. Both ROD and ER-SVM have a parameter corresponding to μ; i.e., the ratio of outliers to be removed from the training samples. The above learning algorithms share almost the same learning model. Here, the main concern of the past studies was to develop computationally-efficient learning algorithms and to confirm the robustness property in numerica experiments.

In this paper, our purpose is a theoretical investigation of the statistical properties of robust SVMs. In particular, we derive the exact finite-sample breakdown point of robust (ν, μ)-SVM. The finite-sample breakdown point indicates the largest amount of contamination such that the estimator still gives information about the non-contaminated data [20] (Chapter 3.2). In order to investigate the breakdown point, we present that the robustness of the learning method is closely related to the dual representation of the optimization problem in the learning algorithm. Indeed, the dual representation provides an intuitive picture on how each sample affects the estimated classifier. Based on such an intuition, we calculate the exact breakdown point. This is a new approach to the theoretical analysis of robust statistics.

In the detailed analysis of the breakdown point, we reveal that the finite-sample breakdown point of robust (ν, μ)-SVM is equal to μ if ν and μ satisfy a simple condition. Conversely, we prove that the finite-sample breakdown point is strictly less than μ, if the condition is violated. An important point is that our findings provide a way to specify a region of the learning parameters (ν, μ), such that robust (ν, μ)-SVM has the desired robustness property. As a result, one can reduce the number of candidate learning parameters (ν, μ) when the grid search of the learning parameters is conducted with cross-validation.

Some of the previous studies are related to ours. In particular, the breakdown point was used to assess the robustness of kernel-based estimators in [14]. In that paper, the influence of a single outlier is considered for a general class of robust estimators in regression problems. In contrast, we focus on a variant of SVM and provide a detailed analysis of the robustness property based on the breakdown point. Our analysis takes into account an arbitrary number of outliers.

The paper is organized as follows. In Section 2, we introduce the problem setup and briefly review the topic of learning algorithms using the standard SVM. Section 3 introduces the robust variant of ν-SVM. We propose a modified learning algorithm of robust (ν, μ)-SVM in order to guarantee the robustness property of local optimal solutions. We show that the dual representation of robust (ν, μ)-SVM has an intuitive interpretation that is of great help for evaluating the breakdown point. In Section 4, we introduce a finite-sample breakdown point as a measure of robustness. Then, we evaluate the breakdown point of robust (ν, μ)-SVM. The robustness of other SVMs is also considered. In Section 5, we discuss a method of tuning the learning parameters ν and μ on the basis of the robustness analysis in Section 4. Section 6 examines the generalization performance of robust (ν, μ)-SVM via numerical experiments. The conclusion is in Section 7. Detailed proofs of the theoretical results are presented in the Appendix.

2. Brief Introduction to Learning Algorithms

First of all, we summarize the notation used throughout this paper. Let $\mathbb{N}$ be the set of positive integers, and let $[m]$ for $m \in \mathbb{N}$ denote a finite set of $\mathbb{N}$ defined as $\{1, \dots, m\}$. The set of all real numbers is denoted as $\mathbb{R}$. The function $[z]_+$ is defined as $\max\{z, 0\}$ for $z \in \mathbb{R}$. For a finite set A, the size of A is expressed as $|A|$. For a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, the norm on $\mathcal{H}$ is denoted as $\|\cdot\|_{\mathcal{H}}$. See [3] for a description of RKHS.

Next, let us introduce the classification problem with an input space $\mathcal{X}$ and binary output labels $\{+1, -1\}$. Given i.i.d. training samples $D = \{(x_i, y_i) : i \in [m]\} \subset \mathcal{X} \times \{+1, -1\}$ drawn from a probability distribution over $\mathcal{X} \times \{+1, -1\}$, a learning algorithm produces a decision function $g : \mathcal{X} \to \mathbb{R}$ such that its sign predicts the output labels for input points in test samples. The decision function $g(x)$ predicts the correct label on the sample (x, y) if and only if the inequality $yg(x) > 0$ holds. The product $yg(x)$ is called the margin of the sample (x, y) for the decision function g [21]. To make an accurate decision function, the margins on the training dataset should take large positive values.

In kernel-based ν-SVM [5], an RKHS $\mathcal{H}$ endowed with a kernel function $k : \mathcal{X}^2 \to \mathbb{R}$ is used to estimate the decision function $g(x) = f(x) + b$, where $f \in \mathcal{H}$ and $b \in \mathbb{R}$. The misclassification penalty is measured by the hinge loss. More precisely, ν-SVM produces a decision function $f(x) + b$ as the optimal solution of the convex problem,

$$\begin{aligned} &\min_{f,b,\rho} \ \frac{1}{2}\|f\|_{\mathcal{H}}^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} [\rho - y_i(f(x_i) + b)]_+ \\ &\text{subject to } \ f \in \mathcal{H},\ b, \rho \in \mathbb{R}, \end{aligned} \tag{1}$$

where $[\rho - y_i(f(x_i) + b)]_+$ is the hinge loss of the margin with the threshold ρ. The second term $-\nu\rho$ is the penalty for the threshold ρ. The parameter ν in the interval $(0, 1)$ is the regularization parameter. Usually, the range of ν that yields a meaningful classifier is narrower than the interval $(0, 1)$, as shown in [5]. The first term in (1) is a regularization term to avoid overfitting to the training data. A large positive margin is preferable for each training data. The optimal ρ of ν-SVM is non-negative. Indeed, the optimal solution $f \in \mathcal{H}, b, \rho \in \mathbb{R}$ satisfies:

$$\begin{aligned} -\nu\rho &\le \frac{1}{2}\|f\|_{\mathcal{H}}^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} \big[\rho - y_i(f(x_i) + b)\big]_+ \\ &\le \frac{1}{2}\|0\|_{\mathcal{H}}^2 - \nu \cdot 0 + \frac{1}{m}\sum_{i=1}^{m} \big[0 - y_i(0 + 0)\big]_+ = 0. \end{aligned}$$

The representer theorem [22,23] indicates that the optimal decision function of (1) is of the form,

$$g(x) = \sum_{j=1}^{m} \alpha_j k(x, x_j) + b \tag{2}$$

for $\alpha_j \in \mathbb{R}$. Thanks to this theorem, even when $\mathcal{H}$ is an infinite dimensional space, the above optimization problem can be reduced to a finite dimensional quadratic convex problem. This is the great advantage of using RKHS for non-parametric statistical inference [5]. The input point x_j with a non-zero coefficient α_j is called a support vector. A remarkable property of ν-SVM is that the regularization parameter ν provides a lower bound on the fraction of support vectors.

As pointed out in [24], ν-SVM is closely related to a financial risk measure called conditional value at risk (CVaR) [25]. Suppose that $\nu m \in \mathbb{N}$ holds for a parameter $\nu \in (0, 1)$. Then, the CVaR of samples $r_1, \ldots, r_m \in \mathbb{R}$ at level ν is defined as the average of its ν-tail, i.e., $\frac{1}{\nu m} \sum_{i=1}^{\nu m} r_{\sigma(i)}$, where σ is a permutation on $[m]$ such that $r_{\sigma(1)} \geq \cdots \geq r_{\sigma(m)}$ holds. The definition of CVaR for general random variables is presented in [25].

In the literature, r_i is defined as the negative margin $r_i = -y_i g(x_i)$. For a regularization parameter ν satisfying $\nu m \in \mathbb{N}$ and a fixed decision function $g(x) = f(x) + b$, the objective function in (1) is expressed as:

$$\min_{\rho \in \mathbb{R}} \frac{1}{2} \|f\|_{\mathcal{H}}^2 - \nu\rho + \frac{1}{m} \sum_{i=1}^{m} [\rho - y_i(f(x_i) + b)]_+ = \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \frac{1}{m} \sum_{i=1}^{\nu m} r_{\sigma(i)}. \tag{3}$$

The proof is presented in Theorem 10 of [25]. Hence, ν-SVM yields a decision function that minimizes the sum of the regularization term and the CVaR of the negative margins at level ν.

In C-SVM [1], the decision function is obtained by solving:

$$\begin{aligned} &\min_{f,b} \ \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C \sum_{i=1}^{m} [1 - y_i(f(x_i) + b)]_+ \\ &\text{subject to } \ f \in \mathcal{H},\ b \in \mathbb{R}, \end{aligned} \tag{4}$$

in which the hinge loss $[1 - y_i(f(x_i) + b)]_+$ with the fixed threshold $\rho = 1$ is used. A positive regularization parameter $C > 0$ is used instead of ν. For each training data, ν-SVM and C-SVM can be made to provide the same decision function by appropriately tuning ν and C. In this paper, we focus on ν-SVM and its robust variants rather than C-SVM. The parameter ν has the explicit meaning shown above, and this interpretation will be significant when we derive the robustness property of our method.

The hinge loss in (4) is replaced with the so-called ramp loss:

$$\min\{1,\ [1 - y_i(f(x_i) + b)]_+\}$$

in the robust C-SVM proposed in [10,13,17]. By truncating the hinge loss, the influence of outliers is suppressed, and the estimated classifier is expected to be robust against outliers in the training data.

3. Robust Variants of SVM

3.1. Outlier Indicators for Robust Learning Methods

Here, we introduce robust (ν, μ)-SVM, which is a robust variant of ν-SVM. To remove the influence of outliers, an outlier indicator, $\eta_i \in [0, 1], i \in [m]$, is assigned for each training sample, where $\eta_i = 0$ is intended to indicate that the sample (x_i, y_i) is an outlier. The same idea is used in ROD [17]. Assume that the ratio of outliers is less than or equal to μ. For ν and μ such that $0 \leq \mu < \nu < 1$; robust (ν, μ)-SVM can be formalized using RKHS $\mathcal{H}$ as follows:

$$\begin{aligned} &\min_{f,b,\rho,\eta} \ \frac{1}{2}\|f\|_{\mathcal{H}}^2 - (\nu-\mu)\rho + \frac{1}{m}\sum_{i=1}^{m}\eta_i\big[\rho - y_i\big(f(x_i)+b\big)\big]_+, \\ &\text{subject to } \ f\in\mathcal{H},\ b,\rho\in\mathbb{R}, \\ &\qquad\qquad \eta=(\eta_1,\ldots,\eta_m)\in[0,1]^m,\ \sum_{i=1}^{m}\eta_i \ge m(1-\mu). \end{aligned} \tag{5}$$

The optimal solution, $f \in \mathcal{H}$ and $b \in \mathbb{R}$, provides the decision function $g(x) = f(x) + b$ for classification. The optimal ρ is non-negative, the same as with ν-SVM. Influence from samples with large negative margins can be removed by setting η_i to zero.

The representer theorem ensures that the optimal decision function of (5) is represented by (2). Suppose that the decision function $g(x) = f(x) + b$ of the form (2), threshold ρ and outlier indicator η satisfy the KKT (Karush–Kuhn–Tucker) condition [26] (Chapter 5) of (5). As in the case of the standard ν-SVM, the number of support vectors in $f(x)$ is bounded below by $(\nu-\mu)m$. In addition, the margin error on the training samples with $\eta_i = 1$ is bounded above by $\nu - \mu$; i.e.,

$$\frac{1}{m}|\{i\in[m] : \eta_i = 1,\ y_i g(x_i) < \rho\}| \le \nu-\mu$$

holds.

In sequel sections, we develop a learning algorithm and investigate its robustness property against outliers. In order to avoid technical difficulties in the theoretical analysis of robust (μ,ν)-SVM, we assume that νm and μm are positive integers throughout this paper. This is not a severe limitation unless the sample size is extremely small. This assumption ensures that the optimal solution of η in (5) lies in the binary product set $\{0,1\}^m$.

Now, let us show the equivalence of robust (ν,μ)-SVM and CVaR-(α_L,α_U)-SVM [16]. Given ν and μ, the optimization problem (5) can be represented as:

$$\min_{f\in\mathcal{H},b\in\mathbb{R}} \frac{1}{2}\|f\|_{\mathcal{H}}^2 + (\nu-\mu)\cdot\frac{1}{(\nu-\mu)m}\sum_{i=\mu m+1}^{\nu m} r_{\sigma(i)}, \tag{6}$$

where $r_i = -y_i(f(x_i)+b)$ is the negative margin and $\sigma(i), i \in [m]$ is the permutation such that $r_{\sigma(1)} \ge \cdots \ge r_{\sigma(m)}$ as defined in Section 2. The second term in (6) is the average of the negative margins included in the middle interval presented in Figure 1, and it is expressed by the difference of CVaRs at levels ν and μ. A learning algorithm based on this interpretation is proposed in [16] under the name CVaR-(α_L,α_U)-SVM with $\alpha_L = 1-\nu$ and $\alpha_U = 1-\mu$.

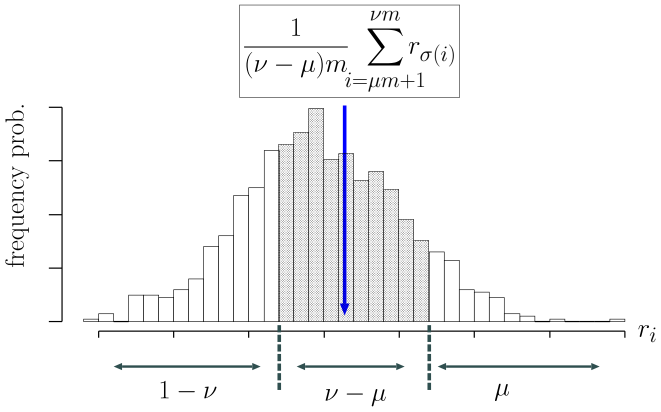


Figure 1. Distribution of negative margins $r_i = -y_i(f(x_i)+b), i \in [m]$ for a fixed decision function $f(x)+b$.

Robust (ν, μ)-SVM is also closely related to the robust outlier detection (ROD) algorithm [17], which is a robust variant of C-SVM. In ROD, the classifier is given by the optimal solution of:

$$\begin{aligned} &\min_{f,b,\eta} \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 + \sum_{i=1}^{m} \eta_i[1 - y_i(f(x_i) + b)]_+, \\ &\text{subject to } \ f \in \mathcal{H},\ \ b \in \mathbb{R}, \\ &\qquad\qquad \eta = (\eta_1, \ldots, \eta_m) \in [0,1]^m,\ \sum_{i=1}^{m} \eta_i \geq m(1-\mu), \end{aligned} \tag{7}$$

where $\lambda > 0$ is a regularization parameter and η is an outlier indicator. The linear kernel is used in the original ROD [17]. To obtain a classifier, the ROD solves a semidefinite relaxation of (7). In [18], it is proven that a KKT point of (7) with the learning parameter (λ, μ) corresponds to that of robust (ν, μ)-SVM for some parameter ν.

3.2. Learning Algorithm

It is hard to obtain a global optimal solution of (5), since the objective function is non-convex. The difference of convex functions algorithm (DCA) [27] and concave-convex programming (CCCP) [28] are popular methods to efficiently obtain practical numerical solutions of non-convex optimization problems. Indeed, DCA is used in robust C-SVM using the ramp loss [12] and ER-SVM [18].

Let us show an expression of the objective function in (5) as a difference of convex functions. The set of feasible outlier indicators is denoted as:

$$E_\mu = \left\{ (\eta_1, \ldots, \eta_m)^T \in [0,1]^m \ : \ \sum_{i=1}^{m} \eta_i \geq m(1-\mu) \right\}.$$

For the negative margin $r_i = -y_i(f(x_i) + b)$, the objective function in robust (ν, μ)-SVM is then represented as:

$$\begin{aligned} &\min_{\rho\in\mathbb{R},\eta\in E_\mu} \frac{1}{2}\|f\|_{\mathcal{H}}^2 - (\nu-\mu)\rho + \frac{1}{m}\sum_{i=1}^{m}\eta_i[\rho + r_i]_+ \\ &= \min_{\rho\in\mathbb{R}} \frac{1}{2}\|f\|_{\mathcal{H}}^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}[\rho + r_i]_+ - \max_{\eta\in E_\mu}\frac{1}{m}\sum_{i=1}^{m}(1-\eta_i)r_i, \end{aligned} \tag{8}$$

which is derived from (3) and (6).

We derive the DCA using the decomposition (8). The optimization algorithm is a simplified variant of the learning algorithm proposed in [18]. The representer theorem ensures that the optimal decision function is represented by $g(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i) + b$ when the kernel function of the RKHS $\mathcal{H}$ is $k(x, x')$. From (8), the objective function of the robust (ν, μ)-SVM is expressed as:

$$\Phi(\alpha, b, \rho) = \psi_0(\alpha, b, \rho) - \psi_1(\alpha, b) \tag{9}$$

using the convex functions ψ_0 and ψ_1 defined as:

$$\psi_0(\alpha, b, \rho) = \frac{1}{2}\alpha^T K\alpha - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}[\rho + r_i],$$

$$\psi_1(\alpha, b) = \max_{\eta\in E_\mu}\frac{1}{m}\sum_{i=1}^{m}(1-\eta_i)r_i,$$

where α is the column vector $(\alpha_1, \ldots, \alpha_m)^T \in \mathbb{R}^m$ and $K \in \mathbb{R}^{m\times m}$ is the Gram matrix defined by $K_{ij} = k(x_i, x_j)$, $i, j \in [m]$. Let $\alpha_t \in \mathbb{R}^m$, $b_t, \rho_t \in \mathbb{R}$ be the solution obtained after t iterations of the DCA. Next, the solution is updated to the optimal solution of:

$$\min_{\alpha,b,\rho} \psi_0(\alpha, b, \rho) - u^T\alpha - vb, \tag{10}$$

where $(u, v) \in \mathbb{R}^{m+1}$ with $u \in \mathbb{R}^m, v \in \mathbb{R}$ is an element of the subgradient of ψ_1 at (α_t, b_t). Let conv$S$ be the convex hull of the set S, and let $a \circ b$ denote component-wise multiplication of two vectors a and b. Accordingly, the subgradient of ψ_1 can be expressed as:

$$\begin{aligned} &\partial\psi_1(\alpha_t, b_t) \\ &= \mathrm{conv}\Big\{ (u, v) \,:\, u = -\frac{1}{m}K(y \circ (1_m - \eta)),\ v = -\frac{1}{m}y^T(1_m - \eta), \\ &\qquad\qquad \text{where } \eta \in E_\mu \text{ is a maximum solution of the problem in } \psi_1(\alpha_t, b_t) \Big\}, \end{aligned}$$

where 1_m denotes an m-dimensional vector of all ones. A parameter $\eta \in E_\mu$ that meets the condition in the above subgradient is obtained by sorting the negative margin r_i, $i \in [m]$ at $(\alpha, b) = (\alpha_t, b_t)$.

Let us describe the learning algorithm for robust (ν, μ)-SVM. We propose a modification of DCA to guarantee the robustness of the local optimal solution. The DCA for robust (ν, μ)-SVM based on Expression (9) is used to obtain a good numerical solution of the outlier indicator. Let Loss(f, b, ρ, η) be the objective function of robust (ν, μ)-SVM:

$$\mathrm{Loss}(f, b, \rho, \eta) = \frac{1}{2}\|f\|^2_{\mathcal{H}} - (\nu - \mu)\rho + \frac{1}{m}\sum_{i=1}^{m} \eta_i[\rho - y_i(f(x_i) + b)]_+.$$

The learning algorithm is presented in Algorithm 1. Given training samples $D = \{(x_i, y_i) : i \in [m]\}$, the learning algorithm outputs the decision function $f_D + b_D \in \mathcal{H} + \mathbb{R}$. The dual problem of (10) is presented as (11) in Algorithm 1.

The numerical solution given by DCA is modified in Steps 7 and 8. Step 7 of Algorithm 1 is equivalent to solving (5) with the additional equality constraint $\eta = \bar{\eta} \in E_\mu$. This is almost the same as the standard ν-SVM using the training samples with $\bar{\eta}_i = 1$ and the regularization parameter $(\nu - \mu)/(1 - \mu) \in (0, 1)$ instead of ν. Hence, the optimal solution f_D is efficiently obtained. In Step 8, the problem is reduced to the optimization of the one-dimensional piecewise linear function of b. This fact is shown in Appendix C, when we prove the robustness property of b_D in Section 4.2. Hence, finding a local optimal solution of the problem in Step 8 is tractable.

Throughout the learning algorithm, the objective value monotonically decreases. Indeed, the DCA has the monotone decreasing property of the objective value [27]. Let $\bar{f}, \bar{b}, \bar{\rho}, \bar{\eta}$ be the numerical solution obtained at the last iteration of DCA. Then, we have:

$$\begin{aligned} \mathrm{Loss}(\bar{f}, \bar{b}, \bar{\rho}, \bar{\eta}) &\geq \min_{f\in\mathcal{H}, b,\rho\in\mathbb{R}} \mathrm{Loss}(f, b, \rho, \bar{\eta}) \\ &= \min_{b,\rho\in\mathbb{R}} \mathrm{Loss}(f_D, b, \rho, \bar{\eta}) \\ &\geq \min_{b,\rho\in\mathbb{R},\eta\in E_\mu} \mathrm{Loss}(f_D, b, \rho, \eta) \\ &= \min_{\rho\in\mathbb{R},\eta\in E_\mu} \mathrm{Loss}(f_D, b_D, \rho, \eta). \end{aligned}$$

It is straightforward to guarantee the monotone decrease of the objective value even if b_D is a local optimal solution.

Algorithm 1 Learning Algorithm of Robust (ν, μ)-SVM

```
Input: Training dataset $D = \{(x_i, y_i) : i \in [m]\}$, Gram matrix $K \in \mathbb{R}^{m\times m}$ defined as $K_{ij} = k(x_i, x_j), i, j \in [m]$, and training labels $y = (y_1, \ldots, y_m)^T \in \{+1, -1\}^m$. The matrix $\widetilde{K} \in \mathbb{R}^{m\times m}$ is defined as $\widetilde{K}_{ij} = y_i y_j K_{ij}$. Let $g(x) = f(x) + b \in \mathcal{H} + \mathbb{R}$ be the initial decision function.
repeat
    Compute the sort $r_{\sigma(1)} \geq \cdots \geq r_{\sigma(m)}$ of $r_i = -y_i g(x_i)$, and set
        $\bar{\eta}_{\sigma(i)} \leftarrow \begin{cases} 0, & 1 \leq i \leq \mu m, \\ 1, & \text{otherwise}, \end{cases}$
    for $i \in [m]$. Let $\bar{\eta}$ be $(\bar{\eta}_1, \ldots, \bar{\eta}_m)^T \in E_\mu$.
    Set $c \leftarrow -\widetilde{K}(1_m - \bar{\eta})/m$ and $d \leftarrow y^T(1_m - \bar{\eta})/m$. Compute the optimal solution $\beta_{\text{opt}}$ of the problem
        $\min_{\beta\in\mathbb{R}^m} \frac{1}{2}\beta^T\widetilde{K}\beta + c^T\beta \text{ subject to } 0_m \leq \beta \leq 1_m/m,\ \beta^T y = d,\ \beta^T 1_m = \nu.$   (11)
    Set $\alpha \leftarrow y \circ (\beta_{\text{opt}} - (1_m - \bar{\eta})/m)$. Compute $\rho$ and $b$ using the relation obtained from the KKT condition,
        $0 < (\beta_{\text{opt}})_i < 1/m \implies \rho = y_i g(x_i),$
    where $g(x_i) = \sum_{j=1}^m K_{ij}\alpha_j + b$.
until the objective value of the robust $(\nu, \mu)$-SVM is unchanged.
Let $\bar{\eta} = (\bar{\eta}_1, \ldots, \bar{\eta}_m)$ be the outlier indicator obtained by DCA.
Let $f_D$ be the optimal solution of $f$ in the following convex optimization problem,
    $\min_{f,b,\rho} \text{Loss}(f, b, \rho, \bar{\eta}), \quad f \in \mathcal{H},\ b, \rho \in \mathbb{R},$
where $\bar{\eta}$ is fixed.
Let $b_D$ be (local) optimal solution of $b$ in the following problem,
    $\min_{b,\rho,\eta} \text{Loss}(f_D, b, \rho, \eta), \quad b, \rho \in \mathbb{R},\ \eta \in E_\mu,$
where $f_D$ is fixed.
Output: the decision function $g(x) = f_D(x) + b_D$.
```

3.3. Dual Problem and Its Interpretation

The partial dual problem of (5) with a fixed outlier indicator $\eta \in [0, 1]^m$ has an intuitive geometric picture. Some variants of ν-SVM can be geometrically interpreted on the basis of the dual form [29–31]. Substituting (2) into the objective function in (5), we obtain the Lagrangian of problem (5) with a fixed $\eta \in E_\mu$ as:

$$L_\eta(\alpha,b,\rho,\xi;\beta,\gamma)=\frac{1}{2}\sum_{i,j=1}^m \alpha_i\alpha_j k(x_i,x_j)-(\nu-\mu)\rho+\frac{1}{m}\sum_{i=1}^m \eta_i\xi_i-\sum_{i=1}^m \beta_i\xi_i$$
$$+\sum_{i=1}^m \gamma_i\bigg(\rho-\xi_i-y_i\bigg(\sum_j k(x_i,x_j)\alpha_j+b\bigg)\bigg),$$

where non-negative slack variables ξ_i, $i\in[m]$ are introduced to represent the hinge loss. Here, the parameters β_i and γ_i for $i\in[m]$ are non-negative Lagrange multipliers. For a fixed $\eta\in E_\mu$, the Lagrangian is convex in the parameters α,b,ρ and ξ and concave in $\beta=(\beta_1,\ldots,\beta_m)$ and $\gamma=(\gamma_1,\ldots,\gamma_m)$. Hence, the min-max theorem [32] (Proposition 6.4.3) yields:

$$\begin{aligned}
&\inf_{\alpha,b,\rho,\xi}\sup_{\beta,\gamma\geq 0} L_\eta(\alpha,b,\rho,\xi;\beta,\gamma)\\
&=\sup_{\beta,\gamma\geq 0}\inf_{\alpha,b,\rho,\xi} L_\eta(\alpha,b,\rho,\xi;\beta,\gamma)\\
&=\sup_{\beta,\gamma\geq 0}\inf_{\alpha,b,\rho,\xi}\rho\bigg(\sum_i\gamma_i-(\nu-\mu)\bigg)+\sum_i\xi_i\bigg(\frac{\eta_i}{m}-\beta_i-\gamma_i\bigg)\\
&\qquad+\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j k(x_i,x_j)-\sum_i\gamma_i y_i\sum_j k(x_i,x_j)\alpha_j-b\sum_i y_i\gamma_i\\
&=\max_\gamma\bigg\{-\frac{1}{2}\bigg\|\sum_i\gamma_i y_i k(\cdot,x_i)\bigg\|_{\mathcal{H}}^2\;:\;\sum_{i:y_i=+1}\gamma_i=\sum_{i:y_i=-1}\gamma_i=\frac{\nu-\mu}{2},\;0\leq\gamma_i\leq\frac{\eta_i}{m}\bigg\}.
\end{aligned}$$

The last equality comes from the optimality condition with respect to the variables α,b,ρ,ξ. Given the optimal solution γ_i, $i\in[m]$ of the dual problem, the optimal coefficient α_i in the primal problem is given by $\alpha_i=\gamma_i y_i$, and the bias term b is obtained from the complementary slackness of γ_i such that $0<\gamma_i<\eta_i/m$ and $\eta_i=1$.

Let us give a geometric interpretation of the above expression. For the training data $D=\{(x_i,y_i):i\in[m]\}$, the convex sets, $\mathcal{U}_\eta^+[\nu,\mu;D]$ and $\mathcal{U}_\eta^-[\nu,\mu;D]$, are defined as the reduced convex hulls of the data points for each label, i.e.,

$$\begin{aligned}
&\mathcal{U}_\eta^\pm[\nu,\mu;D]\\
&=\bigg\{\sum_{i:y_i=\pm1}\gamma_i' k(\cdot,x_i)\in\mathcal{H}:\sum_{i:y_i=\pm1}\gamma_i'=1,\;0\leq\gamma_i'\leq\frac{2\eta_i}{(\nu-\mu)m}\text{ for } i \text{ such that } y_i=\pm1\bigg\}.
\end{aligned}$$

The coefficients γ_i', $i\in[m]$ in $\mathcal{U}_\eta^\pm[\nu,\mu;D]$ are bounded above by a non-negative real number. Hence, the reduced convex hull is a subset of the convex hull of the data points in the RKHS $\mathcal{H}$. Let $\mathcal{V}_\eta[\nu,\mu;D]$ be the Minkowski difference of two subsets,

$$\mathcal{V}_\eta[\nu,\mu;D]=\mathcal{U}_\eta^+[\nu,\mu;D]\ominus\mathcal{U}_\eta^-[\nu,\mu;D],$$

where $A\ominus B$ of subsets A and B denotes $\{a-b:a\in A,\,b\in B\}$. We obtain:

$$\inf_{\alpha,b,\rho,\xi}\sup_{\beta,\gamma\geq 0} L_\eta(\alpha,b,\rho,\xi;\beta,\gamma)\;=\;-\frac{(\nu-\mu)^2}{8}\min\Big\{\|f\|_{\mathcal{H}}^2\;:\;f\in\mathcal{V}_\eta[\nu,\mu;D]\Big\}$$

for each $\eta\in E_\mu$. As a result, the optimal value of (5) is given as $-(\nu-\mu)^2/8\times\mathrm{opt}(\nu,\mu;D)$, where:

$$\mathrm{opt}(\nu,\mu;D)=\max_{\eta\in E_\mu}\min_{f\in\mathcal{V}_\eta[\nu,\mu;D]}\|f\|_{\mathcal{H}}^2. \tag{12}$$

Therefore, the dual form of robust (ν,μ)-SVM can be expressed as the maximization of the minimum distance between two reduced convex hulls, $\mathcal{U}_\eta^+[\nu,\mu;D]$ and $\mathcal{U}_\eta^-[\nu,\mu;D]$. The estimated decision function in robust (ν, μ)-SVM is provided by the optimal solution of (12) up to a scaling factor depending on $\nu - \mu$. Moreover, the optimal value is proportional to the squared RKHS norm of $f \in \mathcal{H}$ in the decision function $g(x) = f(x) + b$.

4. Breakdown Point of Robust SVMs

4.1. Finite-Sample Breakdown Point

Let us describe how to evaluate the robustness of learning algorithms. There are a number of robustness measures for evaluating the stability of estimators as discussed later in Section 4.3. In this paper, we use the finite-sample breakdown point, and it will be referred to as the breakdown point for short. The breakdown point quantifies the degree of impact that the outliers have on the estimators when the contamination ratio is not necessarily infinitesimal [33]. In this section, we present an exact evaluation of the breakdown point of robust SVMs.

The breakdown point indicates the largest amount of contamination such that the estimator still gives information about the non-contaminated data [20] (Chapter 3.2). More precisely, for an estimator θ_D based on a dataset D of size m that takes a value in a normed parameter space, the finite-sample breakdown point is defined as:

$$\varepsilon^* = \max_{\kappa=0,1,\ldots,m} \{\, \kappa/m \,:\, \theta_{D'} \text{ is uniformly bounded for } D' \in \mathcal{D}_\kappa \,\}, \tag{13}$$

where $\mathcal{D}_\kappa$ is the family of datasets of size m including at least $m - \kappa$ elements in common with the non-contaminated dataset D, i.e.,

$$\mathcal{D}_\kappa = \left\{\, D' \,:\, |D'| = m,\, |D' \cap D| \geq m - \kappa \,\right\}.$$

For simplicity, the dependency of $\mathcal{D}_\kappa$ on the dataset D is dropped. The condition of the breakdown point ε^* can be rephrased as:

$$\sup_{D' \in \mathcal{D}_\kappa} \|\theta_{D'}\| < \infty,$$

where $\|\cdot\|$ is the norm on the parameter space. In most cases of interest, ε^* does not depend on the dataset D. For example, the breakdown point of the one-dimensional median estimator is $\varepsilon^* = \lfloor (m-1)/2 \rfloor / m$.

4.2. Breakdown Point of Robust (ν, μ)-SVM

The parameters of robust (ν, μ)-SVM have a clear meaning unlike those of robust C-SVM and ROD. In fact, $\nu - \mu$ is a lower bound of the number of support vectors and an upper bound of the margin error, as mentioned in Section 3.1. In addition, we show that the parameter μ is exactly equal to the breakdown point of the decision function under a mild assumption. Such an intuitive interpretation will be of great help in tuning the parameters in the learning algorithm. Section 5 describes how to tune the learning parameters.

To start with, let us derive a lower bound of the breakdown point for the optimal value of Problem (5) that is expressed as $\mathrm{opt}(\nu, \mu; D)$ up to a constant factor. As shown in Section 3.3, the boundedness of $\mathrm{opt}(\nu, \mu; D)$ is equivalent to the boundedness of the RKHS norm of $f \in \mathcal{H}$ in the estimated decision function $g(x) = f(x) + b$. Given a labeled dataset $D = \{(x_i, y_i) : i \in [m]\}$, let us define the label ratio r as:

$$r = \frac{1}{m} \min\{\, |\{i : y_i = +1\}|, |\{i : y_i = -1\}| \,\}.$$

In what follows, we assume $m\nu, m\mu \in \mathbb{N}$ to avoid technical difficulty.

Theorem 1. *Let D be a labeled dataset of size m with a label ratio $r > 0$. For the parameters ν, μ such that $0 \le \mu < \nu < 1$ and $\nu m, \mu m \in \mathbb{N}$, we assume $\mu < r/2$. Then, the following two conditions are equivalent:*

(i) *The inequality*

$$\nu - \mu \le 2(r - 2\mu) \tag{14}$$

holds.

(ii) *Uniform boundedness,*

$$\sup\{\mathrm{opt}(\nu, \mu; D') \,:\, D' \in \mathcal{D}_{\mu m}\} < \infty$$

holds, where $\mathcal{D}_{\mu m}$ is the family of contaminated datasets defined from D.

The proof of the above theorem is given in Appendix A. The inequality $\mu < r/2$ has an intuitive interpretation. If $\mu < r/2$ is violated, the majority of, say, positive labeled samples in the non-contaminated training dataset can be replaced with outliers. In such a situation, the statistical features in the original dataset will not be retained.

Remark 1. *The condition (14) has an intuitive interpretation. Assume that $m_+ < m/2$. After removing some training samples due to the optimal outlier indicator η, there exist at least $m_+ - m\mu - m\mu = m_+ - 2m\mu$ positive training samples for any $D' \in \mathcal{D}_{m\mu}$. In the standard ν-SVM, the condition $\nu \le 2r$ guarantees the boundedness of the optimal value, $\mathrm{opt}(\nu, 0; D)$, for a non-contaminated dataset D [29]. For the robust (ν, μ)-SVM, ν and r are replaced with $\nu - \mu$ and $(m_+ - 2m\mu)/m$, respectively. As a result, the inequality (14) is obtained as a sufficient condition of $\mathrm{opt}(\nu, \mu; D') < \infty$ for each $D' \in \mathcal{D}_{m\mu}$. This implies the pointwise boundedness of $\mathrm{opt}(\nu, \mu; D')$. However, this interpretation does not prove the uniform boundedness of $\mathrm{opt}(\nu, \mu; D')$ for any $D' \in \mathcal{D}_{m\mu}$. In the proof in Appendix A, we prove the uniform boundedness over $\mathcal{D}_{m\mu}$.*

The inequality (14) indicates the trade-off between the ratio of outliers μ and the ratio of support vectors $\nu - \mu$. This result is reasonable. The number of support vectors corresponds to the dimension of the statistical model. When the ratio of outliers is large, a simple statistical model should be used to obtain robust estimators.

When the contamination ratio in the training dataset is greater than the parameter μ of robust (ν, μ)-SVM, the estimated decision function is not necessarily bounded.

Theorem 2. *Suppose that ν and μ are rational numbers such that $0 < \mu < 1/4$ and $\mu < \nu < 1$. Then, there exists a dataset D of size m with the label ratio r such that $\mu < r/2$ and:*

$$\sup\{\mathrm{opt}(\nu, \mu; D') \,:\, D' \in \mathcal{D}_{\mu m+1}\} = \infty$$

hold, where $\mathcal{D}_{\mu m+1}$ is defined from D.

The proof is given in Appendix B. Theorems 1 and 2 provide lower and upper bounds of the breakdown point, respectively. Hence, the breakdown point of the function part $f \in \mathcal{H}$ in the estimated decision function $g = f + b$ is exactly equal to $\varepsilon^* = \mu$, when the learning parameters of robust (ν, μ)-SVM satisfy $\mu < r/2$ and $\nu - \mu \le 2(r - 2\mu)$. Otherwise, the breakdown point of f is strictly less than μ. Note that the results in Theorems 1 and 2 hold for the global optimal solution.

Remark 2. *Let us consider the robustness of the local optimal solution f_D obtained by robust (ν, μ)-SVM. Let f_{opt} be the global optimal solution of robust (ν, μ)-SVM. For the outlier indicator $\eta = \bar{\eta} \in E_\mu$ in Algorithm 1, we have:*

$$\|f_{\mathrm{opt}}\|_{\mathcal{H}}^2 = \mathrm{opt}(\nu, \mu; D) \geq \min\{\|f\|_{\mathcal{H}}^2 \,:\, f \in \mathcal{V}_{\bar{\eta}}[\nu, \mu; D]\} = \|f_D\|_{\mathcal{H}}^2.$$

where the last equality is guaranteed by the result in Section 3.3. Therefore, f_D is less sensitive to contamination than the RKHS element of the global optimal solution.

Now, we will show the robustness of the bias term b. Let b_D be the estimated bias parameter obtained by Algorithm 1. We will derive a lower bound of the breakdown point of the bias term. Then, we will show that the breakdown point of robust (ν, μ)-SVM with a bounded kernel is given by a simple formula.

Theorem 3. *Let D be an arbitrary dataset of size m with a label ratio r that is greater than zero. Suppose that ν and μ satisfy $0 < \mu < \nu < 1$, $\nu m, \mu m \in \mathbb{N}$, and $\mu < r/2$. For a non-negative integer ℓ, we assume:*

$$0 \leq 2\left(\mu - \frac{\ell}{m}\right) < \nu - \mu < 2(r - 2\mu). \tag{15}$$

Then, uniform boundedness,

$$\sup\{\,|b_{D'}| : D' \in \mathcal{D}_{\mu m - \ell}\,\} < \infty,$$

holds, where $\mathcal{D}_{\mu m - \ell}$ is defined from D.

The proof is given in Appendix C, in which a detailed analysis is needed especially when the kernel function is unbounded. The proof shows that the uniform boundedness holds even if $b_{D'}$ is a local optimal solution in Algorithm 1. Note that the inequality (15) is a sufficient condition of Inequality (14). Theorem 3 guarantees that the breakdown point of the estimated decision function $f_D + b_D$ is not less than $\mu - \ell/m$, when (15) holds.

The robustness of b_D for a bounded kernel is considered in the theorem below.

Theorem 4. *Let D be an arbitrary dataset of size m with a label ratio r that is greater than zero. For parameters such that $0 < \mu < \nu < 1$ and $\nu m, \mu m \in \mathbb{N}$, suppose that $\mu < r/2$ and $\nu - \mu < 2(r - 2\mu)$ hold. In addition, assume that the kernel function $k(x, x')$ of the RKHS $\mathcal{H}$ is bounded, i.e., $\sup_{x \in \mathcal{X}} k(x, x) < \infty$. Then, uniform boundedness,*

$$\sup\{\,|b_{D'}| : D' \in \mathcal{D}_{\mu m}\,\} < \infty,$$

holds, where $\mathcal{D}_{\mu m}$ is defined from D.

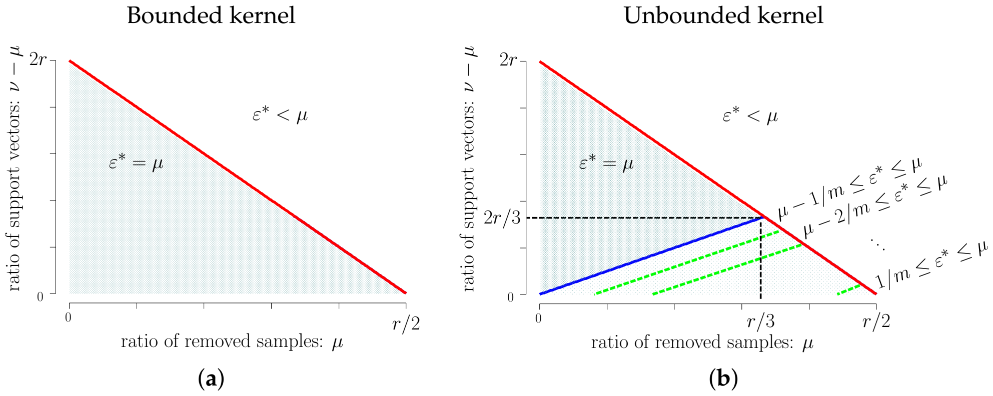
The proof is given in Appendix D. Compared with Theorem 3 in which arbitrary kernel functions are treated, Theorem 4 ensures that a tighter lower bound of the breakdown point is obtained for bounded kernels. The above result agrees with those of other studies. The authors of [14] proved that bounded kernels produce robust estimators for regression problems in the sense of bounded response, i.e., robustness against a single outlier.

Combining Theorems 1–4, we find that the breakdown point of robust (ν, μ)-SVM with $\mu < r/2$ is given as follows.

- Bounded kernel: For $\nu - \mu > 2(r - 2\mu)$, the breakdown point of $f_D \in \mathcal{H}$ is less than μ. For $\nu - \mu \leq 2(r - 2\mu)$, the breakdown point of (f_D, b_D) is equal to μ.

- Unbounded kernel: For $\nu - \mu > 2(r - 2\mu)$, the breakdown point of $f_D \in \mathcal{H}$ is less than μ. For $2\mu < \nu - \mu \leq 2(r - 2\mu)$, the breakdown point of (f_D, b_D) is equal to μ. When $0 < \nu - \mu < \min\{2\mu, 2(r - 2\mu)\}$, the breakdown point of f_D is equal to μ, and the breakdown point of b_D is bounded from below by $\mu - \ell/m$ and from above by μ, where $\ell \in \mathbb{N}$ depends on ν and μ, as shown in Theorem 3.

Figure 2 shows the breakdown point of robust (ν, μ)-SVM. The line $\nu - \mu = 2(r - 2\mu)$ is critical. For unbounded kernels, we only obtain a bound of the breakdown point.



(**a**) (**b**)

Figure 2. (**a**) breakdown point of (f_D, b_D) given by robust (ν, μ)-SVM with bounded kernel; (**b**) breakdown point of (f_D, b_D) given by robust (ν, μ)-SVM with unbounded kernel.

4.3. Breakdown Point Revisited

Let us reconsider the breakdown point of learning methods.

4.3.1. Effective Case of Breakdown Point

Suppose that the function $\widehat{f}_D \in \mathcal{H}$ is obtained by a learning method using the dataset D. Learning methods are categorized into two types according to the norm of $\widehat{f}_D$. The first type is the learning methods satisfying $\sup_{D'} \|\widehat{f}_{D'}\|_{\mathcal{H}} = \infty$, and the second type is the ones such that $\sup_{D'} \|\widehat{f}_{D'}\|_{\mathcal{H}} < \infty$, where the supremum is taken over arbitrary dataset of size m, i.e., $D' \in \mathcal{D}_m = (\mathcal{X} \times \{+1, -1\})^m$.

For learning methods of the first type, the breakdown point indicates the number of outliers such that the estimator remains in a uniformly-bounded region. This is meaningful information about the robustness of the learning method. In this case, the larger breakdown point is regarded as a more robust method. As shown in Theorems 1 and 2, the robust (ν, μ)-SVM is a learning method of the first type.

The second type implies that the hypothesis space of the learning method is bounded regardless of datasets. The C-SVM, robust C-SVM and ROD belong to learning methods of the second type. Indeed, given a labeled dataset $D = \{(x_i, y_i) : i \in [m]\}$, the non-negative property of the hinge loss in C-SVM leads to:

$$\frac{1}{2}\|\widehat{f}_D\|_{\mathcal{H}}^2 \leq \frac{1}{2}\|\widehat{f}_D\|_{\mathcal{H}}^2 + C\sum_{i=1}^{m}[1 - y_i(\widehat{f}_D(x_i) + \widehat{b}_D)]_+ \leq mC,$$

where the last inequality comes from the fact that the objective value at $f = 0$ and $b = 0$ is greater than or equal to the optimal value. Likewise, one can prove that robust C-SVM and ROD have the same property. In this case, the naive definition of the breakdown point shown in Section 4.1 is not adequate, because the boundary effect of the hypothesis set is not taken into account. In the general definition of the breakdown point, the boundary of the hypothesis space is taken into account [20] (Chapter 3.2.5).

In this paper, we focus on the breakdown point of learning algorithms of the first type. Then, the analysis based on the breakdown point suggests proper choices of hyperparameters (ν, μ) as shown in succeeding sections.

4.3.2. Other Robust Estimators

Robust statistical inference has been studied for a long time in mathematical statistics, and a number of robust estimators have been proposed for many kinds of statistical problems [20,34,35]. In mathematical analysis, one needs to quantify the influence of samples on estimators. Here, the influence function, change of variance and breakdown point are often used as measures of robustness. In the machine learning literature, these measures have been used to analyze the theoretical properties of SVM and its robust variants. In [36], the robustness of a learning algorithm using a convex loss function was investigated on the basis of an influence function defined over an RKHS. When the influence function is uniformly bounded on the RKHS, the learning algorithm is regarded to be robust against outliers. It was proven that the least squares loss provides a robust learning algorithm for classification problems in this sense [36].

From the standpoint of the breakdown point, however, convex loss functions do not provide robust estimators, as shown in [20] (Chapter 5.16). Yu et al. [14] proved that the breakdown point of a learning algorithm using clipped loss is greater than or equal to $1/m$ in regression problems. In Section 4.2, we show a detailed analysis of the breakdown point for robust (ν, μ)-SVM.

5. Admissible Region for Learning Parameters

The theoretical analysis in Section 4.2 suggests that robust (ν, μ)-SVM satisfying $0 < \nu - \mu < 2(r - 2\mu)$ is a good choice for obtaining a robust classifier, especially when a bounded kernel is used. Here, r is the label ratio of the non-contaminated original data D, and usually, it is unknown in real-world data analysis. Thus, we need to estimate r from the contaminated dataset D'.

If an upper bound of the outlier ratio is known to be $\widetilde{\mu}$, we have $D' \in \mathcal{D}_{\widetilde{\mu}m}$, where $\mathcal{D}_{\widetilde{\mu}m}$ is defined from D. Let r' be the label ratio of D'. Then, the label ratio of the original dataset D should satisfy $r_{\mathrm{low}} \le r \le r_{\mathrm{up}}$, where $r_{\mathrm{low}} = \max\{r' - \widetilde{\mu}, 0\}$ and $r_{\mathrm{up}} = \min\{r' + \widetilde{\mu}, 1/2\}$. Let Λ_{low} and Λ_{up} be:

$$\Lambda_{\mathrm{low}} = \{(\nu, \mu) \,:\, 0 \le \mu \le \widetilde{\mu},\, 0 < \nu - \mu < 2(r_{\mathrm{low}} - 2\mu)\},$$
$$\Lambda_{\mathrm{up}} = \{(\nu, \mu) \,:\, 0 \le \mu \le \widetilde{\mu},\, 0 < \nu - \mu < 2(r_{\mathrm{up}} - 2\mu)\}.$$

Robust (ν, μ)-SVM with $(\nu, \mu) \in \Lambda_{\mathrm{low}}$ reaches the breakdown point μ for any non-contaminated dataset D such that $D' \in \mathcal{D}_{\mu m}$ for given D'. On the other hand, the parameters (ν, μ) on the outside of Λ_{up} are not necessary. Indeed, for any non-contaminated data D such that $D' \in \mathcal{D}_{\widetilde{\mu}m}$ for given D', the parameters (ν, μ) satisfying $\nu - \mu > 2(r_{\mathrm{up}} - 2\mu)$ do not yield a learning method that reaches the breakdown point μ.

When the upper bound $\widetilde{\mu}$ is unknown, we set $\widetilde{\mu} = r/2$. As shown in the comments after Theorem 1, the outlier ratio greater than $r/2$ can totally violate the statistical features of the original dataset. In such a case, we need to reconsider the observation process. For $\widetilde{\mu} = r/2$, we obtain $\bar{r}_{\mathrm{low}} \le r \le \bar{r}_{\mathrm{up}}$, where $\bar{r}_{\mathrm{low}} = 2r'/3$ and $\bar{r}_{\mathrm{up}} = \min\{2r', 1/2\}$. Hence, in the worst case, the admissible set of learning parameters ν and μ is:

$$\begin{aligned} \overline{\Lambda}_{\mathrm{low}} &= \{(\nu, \mu) \,:\, 0 < \nu - \mu < 2(\bar{r}_{\mathrm{low}} - 2\mu)\}, \text{ or} \\ \overline{\Lambda}_{\mathrm{up}} &= \{(\nu, \mu) \,:\, 0 < \nu - \mu < 2(\bar{r}_{\mathrm{up}} - 2\mu)\}. \end{aligned} \tag{16}$$

Given contaminated training data D', for any D of size m with a label ratio $r \in [\bar{r}_{\mathrm{low}}, \bar{r}_{\mathrm{up}}]$, such that $D' \in \mathcal{D}_{\mu m}$ with $\mu < \bar{r}_{\mathrm{low}}/2$, robust (ν, μ)-SVM with $(\nu, \mu) \in \overline{\Lambda}_{\mathrm{low}}$ provides a classifier with the breakdown point μ. A parameter (ν, μ) on the outside of $\overline{\Lambda}_{\mathrm{up}}$ is not necessary, for the same reasons as for Λ_{up}.

The admissible region of (ν, μ) is useful when the parameters are determined by a grid search based on cross-validation. On the other hand, C of robust C-SVM and λ in ROD can take a wide range of positive real numbers. Hence, differently from robust (ν, μ)-SVM, these algorithms need heuristics to determine the region of the grid search for the learning parameters.

The numerical experiments presented in Section 6 applied a grid search to the region $\overline{\Lambda}_{\mathrm{up}}$.

6. Numerical Experiments

We conducted numerical experiments on synthetic and benchmark datasets to compare a number of SVMs. Algorithm 1 was used for robust (ν, μ)-SVM, and DCA in [12] was used for robust C-SVM with the ramp loss. We used CPLEX to solve the convex quadratic problems.

6.1. DCA versus Global Optimization Methods

As has been shown in many studies including [37], DCA quite often gives global optimal solutions to many different and various non-convex optimization problems. We examined how often DCA produces global optimal solutions to robust (ν, μ)-SVM with the 0-1 valued outlier indicator. Here, the numerical solution of DCA in robust (ν, μ)-SVM denotes the output of Step 5 in Algorithm 1. In these numerical experiments, the optimization problem was formulated as a mixed integer programming (MIP) problem, and the CPLEX MIP solver was used to compute the global optimal solution of robust (ν, μ)-SVM based on a relatively small dataset. The numerical solution given by DCA was compared with the global optimal solution.

In binary classification problems, positive (resp. negative) samples were generated from a multivariate normal distribution with mean $\mu_p = 1_d \in \mathbb{R}^d$ (resp. $\mu_n = -1_d \in \mathbb{R}^d$) and a variance-covariance matrix cI, where I is the identity matrix and c is a positive constant. Each class had 20 samples. For such a small dataset, the global optimal solution was obtained by the CPLEX MIP solver. Outliers were added by flipping positive labels randomly, and the outlier ratio was 10%. The DCA with the multi-start method was used to solve the robust (ν, μ)-SVM using the linear kernel. In the multi-start method, a number of initial points were randomly generated, and for each initial point, a numerical solution was obtained by DCA. Among these numerical solutions, the point that attained the smallest objective value was chosen as the output of the multi-start method. $\mathrm{opt(DCA)}$ was the objective value at the numerical solution of DCA, and $\mathrm{opt(MIP)}$ was the global optimal value. Note that the optimal value of the problem in robust (ν, μ)-SVM is non-positive, i.e., $\mathrm{opt(MIP)} \leq 0$. In addition, one can find that any numerical solution obtained by DCA satisfies $\mathrm{opt(DCA)} \leq 0$.

Table 1. Number of times that the numerical solution of difference of convex functions algorithm (DCA) satisfiesopt$(\mathrm{DCA})/\mathrm{opt(MIP)} \geq 0.97$ out of 100 trials. The number of initial points used in the multi-start method is denoted as #initial points. The“Dim.” and “Cov.” columns denote the dimension d and the covariance matrix of the input vectors in each label. The column labeled “Err.” shows the Bayes error of each problem setting.

Setting		**Err. (%)**	**(ν, μ) in Robust (ν, μ)-SVM Using Linear Kernel**								
			(0.4, 0.1)			**(0.5, 0.1)**			**(0.6, 0.1)**		
Dim.	**Cov.**		**#Initial Points**			**#Initial Points**			**#Initial Points**		
			1	5	10	1	5	10	1	5	10
2	I	7.9	87	96	97	90	99	99	93	99	99
5	I	1.3	98	99	100	100	100	100	99	100	100
10	I	0.1	100	100	100	100	100	100	100	100	100
2	$5I$	26.4	78	84	88	76	85	90	75	85	86
5	$10I$	24.0	46	84	90	53	83	90	66	90	90
10	$50I$	32.7	16	59	73	31	72	77	46	85	92

In the numerical experiments, 100 training datasets such that $\mathrm{opt}(\mathrm{MIP}) < -10^{-4}$ were randomly generated, and $\mathrm{opt}(\mathrm{DCA})$ was computed for each dataset. Table 1 shows the number of times that $\mathrm{opt}(\mathrm{DCA})/\mathrm{opt}(\mathrm{MIP}) \geq 0.97$ holds out of 100 trials. When the achievable lowest test error, i.e., the Bayes error, was large, the DCA tended to yield a local optimal solution that was not globally optimal. When the Bayes error was small, DCA produced approximately global optimal solutions in almost all trials. Even when DCA using a single initial point failed to find the global optimal solution, the multi-start method with five or 10 initial points greatly improved the quality of the numerical solutions. In numerical experiments, DCA was more than 50 times more computationally efficient than the MIP solver.

6.2. Computational Cost

We conducted numerical experiments to compare the computational cost of robust (ν, μ)-SVM with that of robust C-SVM. Both learning algorithms employed the DCA. The numerical experiments were conducted on AMD Opteron Processors 6176 (2.3 GHz) with 48 cores, running Cent OS Linux Release 6.4. We used three benchmark datasets, Sonar, BreastCancer and spam, which were also used in the experiments in Section 6.5. m training samples were randomly chosen from each dataset, and each dataset was contaminated by outliers. The outlier ratio was 5%, and outliers were added by flipping the labels randomly. Robust (ν, μ)-SVM and robust C-SVM with the linear kernel were used to obtain classifiers from the contaminated datasets. This process was repeated 20 times for each dataset. Table 2 presents the average computation time and average ratio of support vectors (SV ratio) together with standard deviations. The support vector was numerically identified as the data point x_i having the coefficient α_i such that $|\alpha_i|$ is greater than 10^{-10}. Although the SV ratio is bounded below by $\nu - \mu$, the bound was not necessarily tight. A similar tendency is often observed in ν-SVM. In terms of the computation time, two learning algorithms were not significantly different except in the case of robust C-SVM with a small C that induces a strong regularization.

Table 2. Computation time (Time) and ratio of support vectors (SV Ratio) of robust (ν, μ)-SVM and robust C-SVMwith standard deviations.

Linear Kernel	**Sonar ($m = 104$)**		**BreastCancer ($m = 350$)**		**Spam ($m = 1000$)**	
Robust (ν, μ)-SVM, (ν, μ)	**Time (s)**	**SV Ratio**	**Time (s)**	**SV Ratio**	**Time (s)**	**SV Ratio**
$(0.2, 0.10)$	1.10 (0.22)	0.79 (0.14)	1.02 (0.17)	0.21 (0.11)	13.38 (3.90)	0.27 (0.22)
$(0.2, 0.05)$	0.87 (0.15)	0.75 (0.20)	0.73 (0.13)	0.18 (0.06)	11.29 (2.41)	0.64 (0.27)
$(0.3, 0.10)$	1.17 (0.19)	0.57 (0.12)	0.80 (0.13)	0.22 (0.07)	9.65 (2.13)	0.24 (0.04)
$(0.3, 0.05)$	0.81 (0.09)	0.58 (0.16)	0.63 (0.07)	0.28 (0.05)	8.64 (2.12)	0.36 (0.21)
$(0.4, 0.10)$	1.11 (0.18)	0.49 (0.10)	0.83 (0.14)	0.30 (0.03)	8.65 (1.25)	0.30 (0.02)
$(0.4, 0.05)$	0.90 (0.15)	0.62 (0.16)	0.76 (0.12)	0.36 (0.02)	8.72 (1.77)	0.38 (0.04)
Robust C-SVM, C						
10^{-7}	0.12 (0.02)	0.00 (0.00)	0.15 (0.02)	0.00 (0.00)	1.62 (0.08)	0.00 (0.00)
1	0.61 (0.07)	0.45 (0.08)	0.60 (0.16)	0.04 (0.01)	7.38 (2.36)	0.08 (0.01)
10^{7}	1.02 (0.11)	0.54 (0.13)	0.68 (0.18)	0.03 (0.01)	10.16 (3.31)	0.11 (0.16)
10^{12}	1.07 (0.13)	0.47 (0.09)	0.63 (0.17)	0.05 (0.06)	20.98 (5.95)	0.30 (0.32)

6.3. Outlier Detection

Robust (ν, μ)-SVM uses an outlier indicator to suppress the influence of outliers. Figure 3 shows that the outlier indicator in the robust (ν, μ)-SVM using the linear kernel is able to detect outliers in a synthetic setting. Similar results have been reported for learning methods using outlier indicators such as ROD and ER-SVM. Systematic experiments using a recall-precision criterion were presented in [17,19].

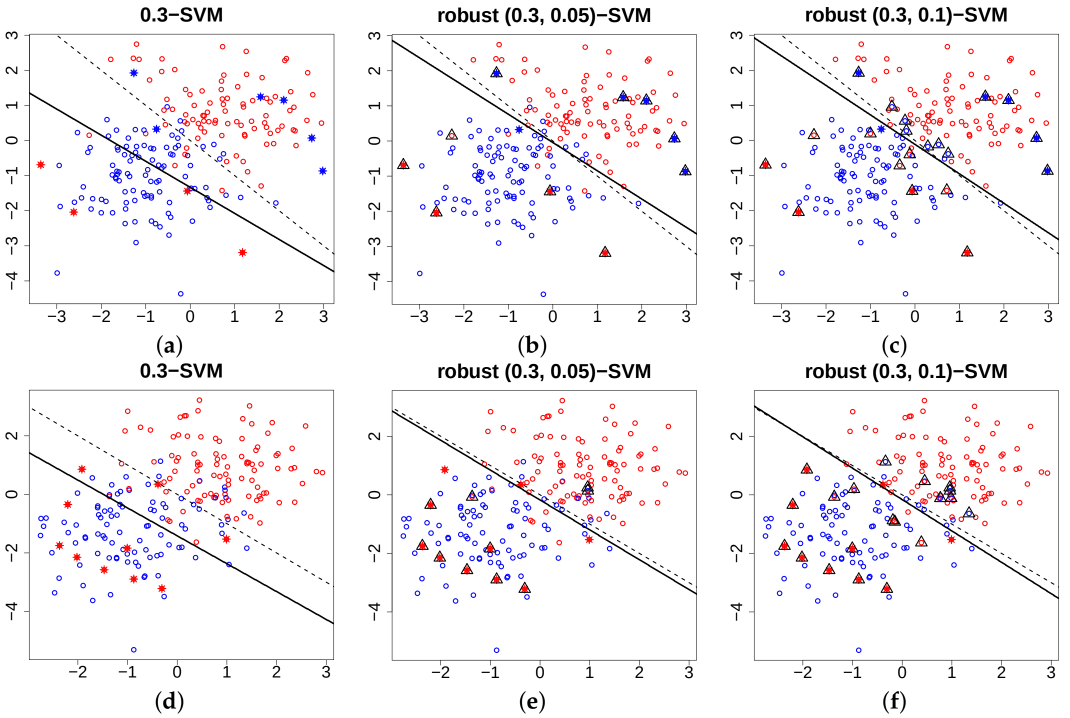
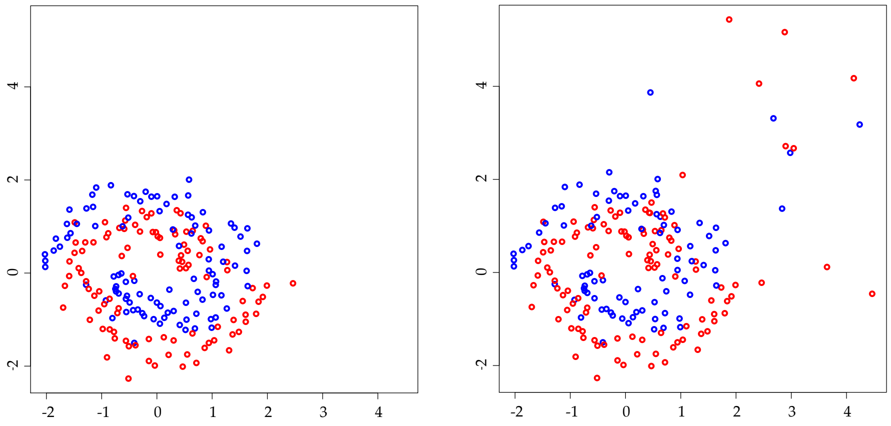


Figure 3. Plot of contaminated dataset of size $m = 200$. The outlier ratio is 0.05, and the asterisks ($*$) denote the outlier. In the panels of the upper (resp. lower) row, outliers are added by flipping labels (resp. flipping positive labels) randomly. The dashed line is the true decision boundary, and the solid line is the decision boundary estimated using ν-SVM with $\nu = 0.3$ in (**a**,**d**); robust (ν, μ)-SVM with $(\nu, \mu) = (0.3, 0.05)$ in (**b**,**e**); and $(\nu, \mu) = (0.3, 0.1)$ in (**c**,**f**). The triangles denote the samples on which $\eta_i = 0$ is assigned.

6.4. Breakdown Point

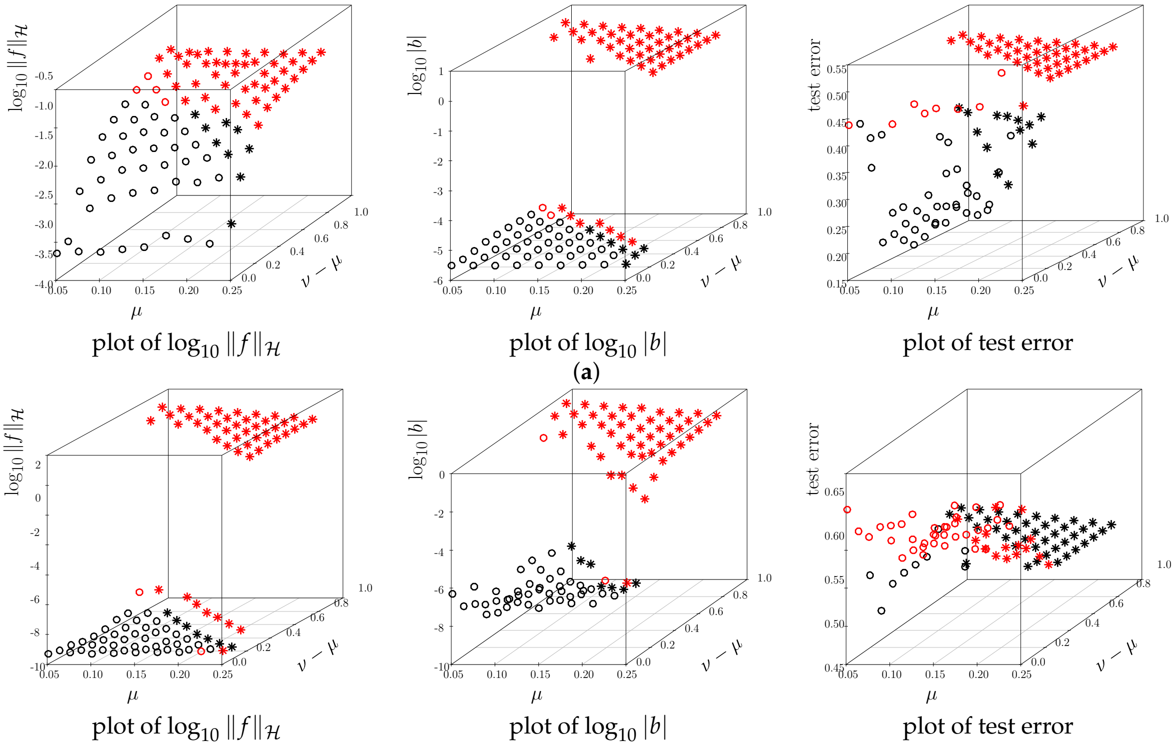
We investigate the validity of Inequality (14) in Theorem 1. In the numerical experiments, the original data D were generated using `mlbench.spirals`in the mlbench library of the R language [38]. Given an outlier ratio μ, positive samples of size μm were randomly chosen from D, and they were replaced with randomly-generated outliers to obtain a contaminated dataset $D' \in \mathcal{D}_{\mu m}$. The original data D and an example of the contaminated data $D' \in \mathcal{D}_{\mu m}$ are shown in Figure 4. The decision function $g(x) = f(x) + b$ was estimated from D' by using robust (ν, μ)-SVM. Here, the true outlier ratio μ was used as the parameter of the learning algorithm. The norms of f and b were then calculated. The above process was repeated 30 times for each pair of parameters (ν, μ), and the maximum values of $\|f\|_{\mathcal{H}}$ and $|b|$ were computed.

Figure 5 shows the results of the numerical experiments. The maximum norm of the estimated decision function is plotted for the parameter $(\mu, \nu - \mu)$ on the same axis as in Figure 2. The top (bottom) panels show the results for a Gaussian (linear) kernel. The left and middle columns show the maximum norm of f and b, respectively. The maximum test errors are presented in the right column. In all panels, the red points denote the top 50 percent of values, and the asterisks ($*$) are the point that violates the inequality $\nu - \mu \leq 2(r - 2\mu)$. In this example, the numerical results agree with the theoretical analysis in Section 4; i.e., the norm becomes large when the inequality $\nu - \mu \leq 2(r - 2\mu)$ is violated. Accordingly, the test error gets close to 0.5; no information for classification. Even when the unbounded linear kernel is used, robustness is confirmed for the parameters in the left lower region in the right panel of Figure 2.



(**a**) (**b**)

Figure 4. (**a**) original data D; (**b**) contaminated data $D' \in \mathcal{D}_{\mu m}$. In this example, the sample size is $m = 200$, and the outlier ratio is $\mu = 0.1$.



plot of $\log_{10} \|f\|_{\mathcal{H}}$ plot of $\log_{10} |b|$ plot of test error

(**a**)

plot of $\log_{10} \|f\|_{\mathcal{H}}$ plot of $\log_{10} |b|$ plot of test error

(**b**)

Figure 5. Plots of maximum norms and worst-case test errors. The top (Bottom) panels show the results for a Gaussian (linear) kernel. Red points mean the top 50 percent of values; the asterisks ($*$) are points that violate the inequality $\nu - \mu \leq 2(r - 2\mu)$. (**a**) Gaussian kernel; (**b**) linear kernel.

In the bottom right panel, the test error gets large when the inequality $\nu - \mu \leq 2(r - 2\mu)$ holds. This result comes from the problem setup. Even with non-contaminated data, the test error of the standard ν-SVM is approximately 0.5, because the linear kernel works poorly for spiral data. Thus, the worst-case test error under the target distribution can go beyond 0.5. For the parameter at which (14) is violated, the test error is always close to 0.5. Thus, a learning method with such parameters does not provide any useful information for classification.

6.5. Prediction Accuracy

As shown in Section 5, the theoretical analysis of the breakdown point yields the admissible region, such as $\overline{\Lambda}_{\mathrm{up}}$, for learning parameters in robust (ν, μ)-SVM. Learning parameters outside the admissible region produce an unstable learning algorithm. Hence, one can reduce the computational cost of tuning the learning parameters by ignoring outside of the admissible region. In this section, we verify the usefulness of the admissible region.

We compared the generalization ability of robust (ν, μ)-SVM with ν-SVM and robust C-SVM using the ramp loss. In robust (ν, μ)-SVM, a grid search of the region $\overline{\Lambda}_{\mathrm{up}}$ is used to choose the learning parameters, ν and μ.

The datasets are presented in Table 3. The datasets are from the `mlbench` and `kernlab` libraries of the R language [38]. The number of positive samples in these datasets is less than or equal to the number of negative samples. Before running the learning algorithms, we standardized each input variable to be mean zero and standard deviation one.

Table 3. Test error and standard deviation of robust (ν, μ)-SVM, robust C-SVM and ν-SVM. The dimension of the input vector, number of training samples, number of test samples and label ratio of all samples with no outliers are shown for each dataset. Linear and Gaussian kernels were used to build the classifier in each method. The outlier ratio in the training data ranged from 0% to 15%, and the test error was evaluated on the non-contaminated test data. The asterisks (*) mean the best result for a fixed kernel function in each dataset, and the double asterisks (**) mean that the corresponding method is 5% significant compared with the second best method under a one-sided t-test. The learning parameters were determined by five-fold cross-validation on the contaminated training data.

Data	**Outlier**	**Linear Kernel**			**Gaussian Kernel**		
		Robust (ν, μ)-SVM	**Robust C-SVM**	**ν-SVM**	**Robust (ν, μ)-SVM**	**Robust C-SVM**	**ν-SVM**
Sonar: $\dim x = 60$,	0%	0.258(.032)	0.270(.038)	* 0.256(.051)	* 0.179(.038)	0.188(0.043)	0.181(0.039)
$\#Train = 104$,	5%	* 0.256(0.039)	0.273(0.047)	0.258(0.046)	0.225(0.042)	0.229(0.051)	* 0.224(0.061)
$\#Test = 104$,	10%	* 0.297(0.060)	0.306(0.067)	0.314(0.060)	0.249(0.059)	** 0.230(0.046)	0.259(0.062)
$r = 0.466$.	15%	* 0.329(0.061)	0.339(0.064)	0.345(0.062)	0.280(0.053)	* 0.280(0.050)	0.294(0.064)
BreastCancer: $\dim x = 10$,	0%	0.033(.010)	0.035(0.008)	* 0.033(0.006)	* 0.032(0.008)	0.035(0.012)	0.033(0.010)
$\#train = 350$,	5%	0.034(0.009)	* 0.034(0.010)	0.043(0.015)	* 0.032(.005)	0.033(0.007)	0.033(0.006)
$\#test = 349$,	10%	0.055(0.015)	* 0.051(0.026)	0.076(0.036)	** 0.035(0.008)	0.043(0.025)	0.038(0.008)
$r = 0.345$	15%	0.136(0.058)	* 0.120(0.050)	0.148(0.058)	0.160(0.083)	* 0.145(0.070)	0.150(0.110)
PimaIndiansDiabetes:	0%	0.237(0.018)	* 0.232(0.014)	0.246(0.018)	* 0.238(0.021)	0.240(0.019)	0.243(0.022)
$\dim x = 8, \#train = 384$,	5%	0.239(0.019)	* 0.237(0.016)	0.269(0.036)	* 0.264(0.025)	0.267(0.024)	0.273(0.024)
$\#test = 384$,	10%	** 0.280(0.046)	0.299(0.042)	0.330(0.030)	0.302(0.039)	* 0.293(0.036)	0.315(0.038)
$r = 0.349$	15%	** 0.338(0.042)	0.349(0.030)	0.351(0.026)	* 0.344(0.028)	0.344(0.031)	0.353(0.016)
spam: $\dim x = 57$,	0%	0.083(0.005)	0.088(0.006)	*0.083(0.005)	0.081(0.005)	0.086(0.006)	* 0.081(0.006)
$\#train = 1000$,	5%	** 0.094(0.008)	0.104(0.013)	0.109(0.010)	0.095(0.008)	0.097(0.009)	* 0.095(0.008)
$\#test = 3601$,	10%	** 0.129(0.022)	0.152(0.020)	0.166(0.067)	* 0.129(0.015)	0.133(0.017)	0.141(.030)
$r = 0.394$	15%	** 0.201(0.029)	0.240(0.030)	0.256(0.091)	** 0.206(0.018)	0.223(0.030)	0.240(0.055)
Satellite: $\dim x = 36$,	0%	0.097(0.004)	0.096(0.003)	** 0.094(0.003)	0.069(0.031)	0.067(0.004)	** 0.063(0.004)
$\#train = 2000$,	5%	0.101(0.003)	* 0.100(0.005)	0.100(0.004)	*0.072(0.015)	0.078(0.007)	0.078(0.043)
$\#test = 4435, r = 0.234$	10%	** 0.148(0.020)	0.161(0.026)	0.161(0.019)	*0.117(0.034)	0.126(0.040)	0.137(0.027)

We randomly split the dataset into training and test sets. To evaluate the robustness, the training data were contaminated by outliers. More precisely, we randomly chose positive labeled samples in the training data and changed their labels to negative; i.e., we added outliers by flipping the labels. After that, robust (ν, μ)-SVM, robust C-SVM using the ramp loss and the standard ν-SVM were used to obtain classifiers from the contaminated training dataset. The prediction accuracy of each classifier was evaluated over test data that had no outliers. Linear and Gaussian kernels were employed for each learning algorithm. The learning parameters, such as μ, ν and C, were determined by conducting a grid search based on five-fold cross-validation over the training data. For robust (ν, μ)-SVM, the parameter (μ, ν) was selected from the admissible region $\overline{\Lambda}_{\mathrm{up}}$ in (16). For standard ν-SVM, the candidate of the regularization parameter ν was selected from the interval $(0, 2r')$, where r' is the label ratio of the contaminated training data. For robust C-SVM, the regularization parameter C was selected from the interval $[10^{-7}, 10^{7}]$. In the grid search of the parameters, 24 or 25 candidates were examined for each learning method. Thus, we needed to solve convex or non-convex optimization problems more than 24 × 5 times in order to obtain a classifier. The above process was repeated 30 times, and the average test error was calculated.

The results are presented in Table 3. For non-contaminated training data, robust (ν, μ)-SVM and robust C-SVM were comparable to the standard ν-SVM. When the outlier ratio is high, we can conclude that robust (ν, μ)-SVM and robust C-SVM tend to work better than the standard ν-SVM. In this experiment, the kernel function does not affect the relative prediction performance of these learning methods. In large datasets, such as spam and Satellite, robust (ν, μ)-SVM tends to outperform robust C-SVM. When the learning parameters, such as ν, μ and C, are appropriately chosen by using a large dataset, the learning algorithms with multiple learning parameters clearly work better than those with a single learning parameter. In addition, in robust C-SVM, there is a difficulty in choosing the regularization parameter. Indeed, the parameter C does not have a clear meaning, and thus, it is not so easy to determine its candidates in the grid search optimization. In contrast, ν in ν-SVM and its robust variant has a clear meaning, i.e., a lower bound of the ratio of support vectors and an upper bound of the margin error on the training data [5]. Such a clear meaning is helpful for choosing candidate points of regularization parameters.

7. Concluding Remarks

We have investigated the breakdown point of robust variants of SVMs. The theoretical analysis provides inequalities of learning parameters, ν and μ, in robust (ν, μ)-SVM that guarantee the robustness of the learning algorithm. Numerical experiments showed that the inequalities are critical to obtaining a robust classifier. The exact evaluation of the breakdown point for robust (ν, μ)-SVM enables us to restrict the range of the learning parameters and to increase the chance of finding a robust classifier with good performance for the same computational cost.

In our paper, the dual representation of robust SVMs is applied to the calculation of the breakdown point. Theoretical analysis using the dual representation can be a powerful tool for the detailed analysis of other learning algorithms.

On the theoretical side, it is interesting to establish the relationship between the robustness, say breakdown point, and the convergence speed of learning algorithms, as presented for the parametric inference in mathematical statistics [34] (Chapter 2.4). Furthermore, it is important to determine the optimal parameter choice of (ν, μ) in robust (ν, μ)-SVM as an extension of the parameter choice for ν-SVM [39]. Another important issue is to develop efficient optimization algorithms. Although the DC algorithm [12,27] and convex relaxation [14,17] are promising methods, more scalable algorithms will be required to deal with massive datasets that are often contaminated by outliers. Recently, a computationally-efficient algorithm, called iteratively weighted SVM (IWSVM), was developed to solve optimization problems in the robust C-SVM and its variants [40]. Moreover, a fixed point of IWSVM is assured to be a local optimal solution obtained by the DC algorithm. It will be worthwhile to investigate the applicability of IWSVM to robust (ν, μ)-SVM.

Acknowledgments: This work was supported by JSPS KAKENHI, Grant Number 16K00044 and 15K00031.

Author Contributions: Takafumi Kanamori and Akiko Takeda contributed the theoretical analysis; Takafumi Kanamori and Shuhei Fujiwara performed the experiments; Takafumi Kanamori and Akiko Takeda wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Theorem 1

The proof is decomposed into two lemmas. Lemma A1 shows that Condition (i) is sufficient for Condition (ii), and Lemma A2 shows that Condition (ii) does not hold if Inequality (14) is violated. For the dataset $D = \{(x_i, y_i) : i \in [m]\}$, let I_+ and I_- be the index sets defined as $I_\pm = \{i : y_i = \pm 1\}$. When the parameter μ is equal to zero, the theorem holds according to the argument on the standard ν-SVM [29]. Below, we assume $\mu > 0$.

Lemma A1. *Under the assumptions of Theorem 1, Condition (i) leads to Condition (ii).*

Proof of Lemma A1. We will show that $\mathcal{V}_\eta[\nu, \mu; D']$ is not empty for any $D' \in \mathcal{D}_{\mu m}$. For a contaminated dataset $D' = \{(x'_i, y'_i) : i \in [m]\} \in \mathcal{D}_{\mu m}$, let us define $\widetilde{I}_+ \subset I_+$ as an index set, such that the sample $(x_i, y_i) \in D$ for $i \in \widetilde{I}_+$ is replaced with $(x'_i, y'_i) \in D'$ as an outlier. In the same way, $\widetilde{I}_- \subset I_-$ is defined for negative samples in D. Therefore, for any index i in $I_+ \setminus \widetilde{I}_+$ or $I_- \setminus \widetilde{I}_-$, we have $(x_i, y_i) = (x'_i, y'_i)$. The assumptions of the theorem ensure $|\widetilde{I}_+| + |\widetilde{I}_-| \le \mu m$. Let us define $J_{\eta,+} = \{i \in I_+ \setminus \widetilde{I}_+ : \eta_i = 1\}$ and $J_{\eta,-} = \{i \in I_- \setminus \widetilde{I}_- : \eta_i = 1\}$. These sets are not empty. Indeed, we have:

$$|J_{\eta,+}| \ge m_+ - m\mu - m\mu \ge \frac{(\nu - \mu)m}{2} > 0, \tag{A1}$$

where Condition (i) in Theorem 1 is used in the second inequality. Likewise, we have $|J_{\eta,-}| > 0$.

We define two points in $\mathcal{H}$ as:

$$\begin{aligned} f_{\eta,+} &= \frac{1}{|J_{\eta,+}|} \sum_{i \in J_{\eta,+}} k(\cdot, x'_i) = \frac{1}{|J_{\eta,+}|} \sum_{i \in J_{\eta,+}} k(\cdot, x_i), \\ f_{\eta,-} &= \frac{1}{|J_{\eta,-}|} \sum_{i \in J_{\eta,-}} k(\cdot, x'_i) = \frac{1}{|J_{\eta,-}|} \sum_{i \in J_{\eta,-}} k(\cdot, x_i). \end{aligned}$$

Then, we have:

$$\begin{aligned} f_{\eta,+} &\in \mathcal{U}_\eta^+[\nu, \mu; D'] \cap \mathrm{conv}\{k(\cdot, x_i) \ : \ i \in I_+\}, \\ f_{\eta,-} &\in \mathcal{U}_\eta^-[\nu, \mu; D'] \cap \mathrm{conv}\{k(\cdot, x_i) \ : \ i \in I_-\}. \end{aligned}$$

Because $1/|J_{\eta,+}|$ and $1/|J_{\eta,-}|$ are both less than or equal to $\frac{2}{(\nu-\mu)m}$ due to (A1), $\eta_i = 1$ holds for all $i \in J_{\eta,+} \cup J_{\eta,-}$.

Now, let us prove the inequality,

$$\sup_{D' \in \mathcal{D}_{\mu m}} \max_{\eta \in E_\mu} \inf_{f \in \mathcal{V}_\eta[\nu,\mu;D']} \|f\|_{\mathcal{H}}^2 < \infty. \tag{A2}$$

The above argument leads to:

$$\min_{f \in \mathcal{V}_\eta[\nu,\mu;D']} \|f\|_{\mathcal{H}}^2 \le \|f_{\eta,+} - f_{\eta,-}\|_{\mathcal{H}}^2$$

for any $\eta \in E_\mu$. Let us define:

$$\mathcal{C}[D] = \mathrm{conv}\{k(\cdot, x_i) \ : \ i \in I_+\} \ominus \mathrm{conv}\{k(\cdot, x_i) \ : \ i \in I_-\}$$

for the original dataset D. Then, we obtain:

$$
\begin{aligned}
\mathrm{opt}(\nu,\mu;D') &= \max_{\eta\in E_\mu}\ \min_{f\in\mathcal{V}_\eta[\nu,\mu;D']} \|f\|_{\mathcal{H}}^2 \\
&\le \max_{\eta\in E_\mu} \|f_{\eta,+}-f_{\eta,-}\|_{\mathcal{H}}^2 \\
&\le \max_{\eta\in E_\mu}\ \max_{f\in\mathcal{C}[D]} \|f\|_{\mathcal{H}}^2 \\
&= \max_{f\in\mathcal{C}[D]} \|f\|_{\mathcal{H}}^2 \\
&\le 4\max_{i\in[m]} k(x_i,x_i) < \infty. \qquad\qquad \text{(triangle inequality)}
\end{aligned}
$$

The upper bound does not depend on the contaminated dataset $D' \in \mathcal{D}_{\mu m}$. Thus, the inequality (A2) holds. □

Lemma A2. *Under the condition of Theorem 1, we assume $\nu-\mu > 2(r-2\mu)$. Then, we have:*

$$
\sup\{\mathrm{opt}(\nu,\mu;D') : D' \in \mathcal{D}_{\mu m}\} = \infty.
$$

Proof of Lemma A2. We will use the same notation as in the proof of Lemma A1. Without loss of generality, we can assume $r = |I_-|/m$. We will prove that there exists a feasible parameter $\eta' \in E_\mu$ and a contaminated training set $D' = \{(x_i', y_i') : i \in [m]\} \in \mathcal{D}_{\mu m}$ such that $\mathcal{U}_{\eta'}^-[\nu,\mu;D']$ becomes empty. The construction of the dataset D' is illustrated in Figure A1. Suppose that $|\widetilde{I}_+| = 0$ and $|\widetilde{I}_-| = \mu m$ and that $y_i' = +1$ holds for all $i \in \widetilde{I}_-$, meaning that all outliers in D' are made by flipping the labels of the negative samples in D. This is possible, because $\mu m < |I_-|/2 < |I_-|$ holds. The outlier indicator $\eta' = (\eta_1', \ldots, \eta_m') \in E_\mu$ is defined by $\eta_i' = 0$ for μm samples in $I_- \setminus \widetilde{I}_-$, and $\eta_i' = 1$ otherwise. This assignment is possible because $|I_- \setminus \widetilde{I}_-| = |I_-| - \mu m > \mu m$. Then, we have:

$$
|J_{\eta',-}| = |\{i \in I_- \setminus \widetilde{I}_- \,:\, \eta_i' = 1\}| = |I_- \setminus \widetilde{I}_-| - \mu m = |I_-| - 2\mu m < \frac{(\nu-\mu)m}{2},
$$

where $\nu-\mu > 2(r-2\mu)$ is used in the last inequality. In addition, $y_i' = -1$ holds only when $i \in I_- \setminus \widetilde{I}_-$. Therefore, we have $\mathcal{U}_{\eta'}^-[\nu,\mu;D'] = \emptyset$. The infeasibility of the dual problem means that the primal problem is unbounded or infeasible. In this case, the infeasibility of the primal problem is excluded. Hence, a contaminated dataset $D' \in \mathcal{D}_{\mu m}$ and an outlier indicator $\eta' \in E_\mu$ exist such that:

$$
\mathrm{opt}(\nu,\mu;D') \ge \min_{f\in\mathcal{V}_{\eta'}[\nu,\mu;D']} \|f\|_{\mathcal{H}}^2 = \infty
$$

holds. □

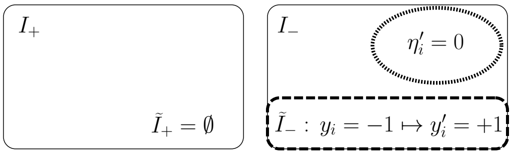

I_+
$\widetilde{I}_+ = \emptyset$
I_-
$\eta_i' = 0$
$\widetilde{I}_- : y_i = -1 \mapsto y_i' = +1$

Figure A1. Index sets $\widetilde{I}_\pm$ and value of η_i' defined in the proof of Lemma A2.

Appendix B. Proof of Theorem 2

Proof. For a rational number $\mu \in (0, 1/4)$, there exists an $m \in \mathbb{N}$ such that $\mu m \in \mathbb{N}$ and $2\mu m + 1 \le m - (2\mu m + 1)$ hold. For such m, let $D = \{(x_i, y_i) : i \in [m]\}$ be training data such that $|I_-| = 2\mu m + 1$ and $|I_+| = m - (2\mu m + 1)$, where the index sets $I_\pm$ are defined in the proof of Appendix A. Since the label ratio of D is $r = \min\{|I_-|, |I_+|\}/m = 2\mu + 1/m$, and we have $\mu < r/2$. For $\mathcal{D}_{\mu m+1}$ defined from D, let $D' = \{(x_i', y_i') : i \in [m]\} \in \mathcal{D}_{\mu m+1}$ be a contaminated dataset of D such that $\mu m + 1$ outliers are made by flipping the labels of the negative samples in D. Thus, there are μm negative samples in D'. Let us define the outlier indicator $\eta' = (\eta_1', \ldots, \eta_m') \in E_\mu$ such that $\eta_i' = 0$ for μm negative samples in D'. Then, any sample in D' with $\eta_i' = 1$ should be a positive one. Hence, we have $\mathcal{U}_{\eta'}^{-}[\nu, \mu; D'] = \emptyset$. The infeasibility of the dual problem means that the primal problem is unbounded. Thus, we obtain $\mathrm{opt}(\nu, \mu; D') = \infty$. □

Appendix C. Proof of Theorem 3

Let us define $f_D + b_D$ with $f_D \in \mathcal{H}$, $b_D \in \mathbb{R}$ as the decision function estimated using robust (ν, μ)-SVM based on the dataset D.

Proof. The non-contaminated dataset is denoted as $D = \{(x_i, y_i) : i \in [m]\}$. For the dataset D, let I_+ and I_- be the index sets defined by $I_\pm = \{i : y_i = \pm 1\}$. Inequality (14) holds under the conditions of Theorem 3. Given a contaminated dataset $D' = \{(x_i', y_i') : i \in [m]\} \in \mathcal{D}_{\mu m-\ell}$, let $r_i'(b)$ be the negative margin of $f_{D'} + b$, i.e., $r_i'(b) = -y_i'(f_{D'}(x_i') + b)$ for $(x_i', y_i') \in D'$. For $b \in \mathbb{R}$, the function $\zeta(b)$ is defined as:

$$\zeta(b) = \frac{1}{m} \sum_{i \in T_b} r_i'(b),$$

where the index set T_b is defined by the sorted negative margins as follows:

$$T_b = \{\sigma(j) \in [m] \,:\, \mu m + 1 \le j \le \nu m,\ r_{\sigma(1)}'(b) \ge \cdots \ge r_{\sigma(m)}'(b)\}.$$

The estimated bias term $b_{D'}$ is a local optimal solution of $\zeta(b)$ because of (6). The function $\zeta(b)$ is continuous. In addition, $\zeta(b)$ is linear on the interval such that T_b is unchanged. Hence, $\zeta(b)$ is a continuous piecewise linear function. Below, we prove that local optimal solutions of $\zeta(b)$ are uniformly bounded regardless of the contaminated dataset $D' \in \mathcal{D}_{\mu m-\ell}$. To prove the uniform boundedness, we control the slope of $\zeta(b)$.

For the non-contaminated data D, let R be a positive real number such that:

$$\sup\{|f_{D''}(x)| : (x, y) \in D,\ D'' \in \mathcal{D}_{\mu m-\ell}\} \le R.$$

The existence of R is guaranteed. Indeed, one can choose:

$$R = \sup_{D'' \in \mathcal{D}_{\mu m-\ell}} \|f_{D''}\|_{\mathcal{H}} \cdot \max_{(x,y) \in D} \sqrt{k(x, x)} < \infty,$$

because the RKHS norm of $f_{D''}$ is uniformly bounded above for $D'' \in \mathcal{D}_{\mu m-\ell}$ and D is a finite set. For the contaminated dataset $D' = \{(x_i', y_i') \,:\, i \in [m]\} \in \mathcal{D}_{\mu m-\ell}$, let us define the index sets $I_\pm'$, $I_{\mathrm{in},\pm}'$ and $I_{\mathrm{out},\pm}'$ for each label by:

$$\begin{aligned} I_\pm' &= \{i \in [m] \,:\, y_i' = \pm 1\}, \\ I_{\mathrm{in},\pm}' &= \{i \in I_\pm' \,:\, |f_{D'}(x_i')| \le R\}, \\ I_{\mathrm{out},\pm}' &= \{i \in I_\pm' \,:\, |f_{D'}(x_i')| > R\}. \end{aligned}$$

For any non-contaminated sample $(x_i, y_i) \in D$, we have $|f_{D'}(x_i)| \le R$. Hence, $(x_i', y_i') \in D'$ for $i \in I_{\mathrm{out},\pm}'$ should be an outlier that is not included in D. This fact leads to:

$$\begin{aligned} |I'_{\text{out},+}| + |I'_{\text{out},-}| &\le \mu m - \ell, \\ |I'_{\text{in},\pm}| &\ge |I_\pm| - (\mu m - \ell) \ge (r - \mu)m + \ell. \end{aligned}$$

On the basis of the argument above, we can prove two propositions:

1. The function $\zeta(b)$ is increasing for $b > R$.
2. The function $\zeta(b)$ is decreasing for $b < -R$.

In addition, for any $D' \in \mathcal{D}_{\mu m - \ell}$, the Lipschitz constant of $\zeta(b)$ is greater than or equal to $1/m$ for $R < |b|$.

Let us prove the first statement. If $b > R$ holds, we have:

$$R - b < \min\{r'_i(b) : i \in I'_{\text{in},-}\} \tag{A3}$$

from the definition of the index set $I'_{\text{in},-}$. Let us consider two cases:

(i) for all $i \in T_b$, $R - b < r'_i(b)$ holds and
(ii) there exists an index $i \in T_b$ such that $r'_i(b) \le R - b$.

For a fixed b such that $b > R$, let us assume (i) above. Then, for any index i in $I'_+ \cap T_b$, we have $R < -f_{D'}(x'_i)$, meaning that $i \in I'_{\text{out},+}$. Hence, the size of the set $I'_+ \cap T_b$ is less than or equal to $\mu m - \ell$. Therefore, the size of the set $I'_- \cap T_b$ is greater than or equal to $(\nu - \mu)m - (\mu m - \ell) = (\nu - 2\mu)m + \ell$. The first inequality of (15) leads to $(\nu - 2\mu)m + \ell > \mu m - \ell$. Therefore, in the set T_b, the number of negative samples is more than the number of positive samples.

For a fixed b such that $b > R$, let us assume (ii) above. Due to the inequality (A3), for any index $i \in I'_{\text{in},-}$, the negative margin $r'_i(b)$ is at the top νm of those ranked in the descending order. Hence, the size of the set $I'_- \cap T_b$ is greater than or equal to $|I'_{\text{in},-}| - \mu m \ge (r - 2\mu)m$. Therefore, the size of the set $I'_+ \cap T_b$ is less than or equal to $(\nu - \mu)m - (r - 2\mu)m = (\nu - r + \mu)m$. The second inequality of (15) leads to $(\nu - r + \mu)m < (r - 2\mu)m$. Furthermore, in the case of (ii), the negative label dominates the positive label in the set T_b.

For negative (resp. positive) samples, the negative margin is expressed as $r'_i(b) = u_i + b$ (resp. $r'_i(b) = u_i - b$) with a constant $u_i \in \mathbb{R}$. Thus, the continuous piecewise linear function $\zeta(b)$ is expressed as:

$$\zeta(b) = \frac{c_b + b \cdot a_b}{m},$$

where $a_b, c_b \in \mathbb{R}$ are constants as long as T_b is unchanged. As proven above, a_b is a positive integer, since negative samples outnumber positive samples in T_b when $b > R$. As a result, local optimal solutions of the bias term should satisfy:

$$\sup\{b_{D'} : D' \in \mathcal{D}_{\mu m - \ell}\} \le R.$$

In the same manner, we can prove the second statement by using the fact that $b < -R$ is a sufficient condition of:

$$R + b < \min\{r'_i(b) : i \in I'_{\text{in},+}\}.$$

Then, we have:

$$\inf\{b_{D'} : D' \in \mathcal{D}_{\mu m - \ell}\} \ge -R.$$

In summary, we obtain:

$$\sup\{|b_{D'}| : D' \in \mathcal{D}_{\mu m-\ell}\} \le R < \infty.$$

□

Appendix D. Proof of Theorem 4

Proof. We will use the same notation as in the proof of Theorem 3 in Appendix C. Note that Inequality (14) holds under the assumption of Theorem 4. The reproducing property of the RKHS inner product yields:

$$|f_{D'}(x'_i)| \le \|f_{D'}\|_{\mathcal{H}}\sqrt{k(x'_i,x'_i)} \le \sup_{D''\in\mathcal{D}_{\mu m}} \|f_{D''}\|_{\mathcal{H}} \cdot \sup_{x\in\mathcal{X}} \sqrt{k(x,x)} < \infty$$

for any $D' = \{(x'_i, y'_i) : i \in [m]\} \in \mathcal{D}_{\mu m}$ due to the boundedness of the kernel function and Inequality (14). Hence, for a sufficiently large $R \in \mathbb{R}$, the sets $I'_{\text{out},+}$ and $I'_{\text{out},-}$ become empty for any $D' \in \mathcal{D}_{\mu m}$.

Under Inequality (A3), suppose that $R - b < r'_i(b)$ holds for all $i \in T_b$. Then, for $i \in I'_+ \cap T_b$, we have $R < -f_{D'}(x'_i)$. Thus, $i \in I'_{\text{out},+}$ holds. Since $I'_{\text{out},+}$ is the empty set, $I'_+ \cap T_b$ is also the empty set. Therefore, T_b has only negative samples. Let us consider the other case; i.e., there exists an index $i \in T_b$, such that $r'_i(b) \le R - b$. Assuming that $\nu - \mu < 2(r - 2\mu)$, we can prove that the negative labels dominate the positive labels in T_b in the same manner as the proof of Theorem 3. For any $D' \in \mathcal{D}_{\mu m}$, the function $\zeta(b)$ is strictly increasing for $b > R$. In the same way, we can prove that $\zeta(b)$ is strictly decreasing for $b < -R$. Moreover, for any $D' \in \mathcal{D}_{\mu m}$ and for $|b| > R$, one can prove that the absolute value of the slope of $\zeta(b)$ is bounded below by $1/m$ according to the argument in the proof of Theorem 3. As a result, we obtain $\sup\{|b_{D'}| : D' \in \mathcal{D}_{\mu m}\} \le R$. □

References

1. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
2. Schölkopf, B.; Smola, A.J. *Learning with Kernels*; MIT Press: Cambridge, MA, USA, 2002.
3. Berlinet, A.; Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*; Kluwer Academic: Boston, MA, USA, 2004.
4. Perez-Cruz, F.; Weston, J.; Hermann, D.J.L.; Schölkopf, B. Extension of the ν-SVM Range for Classification. In *Advances in Learning Theory: Methods, Models and Applications 190*; IOS Press: Amsterdam, The Netherlands, 2003; pp. 179–196.
5. Schölkopf, B.; Smola, A.; Williamson, R.; Bartlett, P. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207–1245.
6. Suykens, J.A.K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300.
7. Bartlett, P.L.; Jordan, M.I.; McAuliffe, J.D. Convexity, classification, and risk bounds. *J. Am. Stat. Assoc.* **2006**, *101*, 138–156.
8. Steinwart, I. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* **2001**, *2*, 67–93.
9. Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.* **2004**, *32*, 56–134.
10. Shen, X.; Tseng, G.C.; Zhang, X.; Wong, W.H. On ψ-learning. *J. Am. Stat. Assoc.* **2003**, *98*, 724–734.
11. Yu, Y.; Yang, M.; Xu, L.; White, M.; Schuurmans, D. Relaxed Clipping: A Global Training Method for Robust Regression and Classification. In *Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2010; pp. 2532–2540.

12. Collobert, R.; Sinz, F.; Weston, J.; Bottou, L. Trading Convexity for Scalability. In Proceedings of the ICML06, 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; ACM Press: New York, NY, USA, 2006; pp. 201–208.
13. Wu, Y.; Liu, Y. Robust truncated hinge loss support vector machines. *J. Am. Stat. Assoc.* **2007**, *102*, 974–983.
14. Yu, Y.; Aslan, Ö.; Schuurmans, D. A Polynomial-Time Form of Robust Regression. In *Advances in Neural Information Processing Systems 25*; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 2483–2491.
15. Feng, Y.; Yang, Y.; Huang, X.; Mehrkanoon, S.; Suykens, J.A. Robust Support Vector Machines for Classification with Nonconvex and Smooth Losses. *Neural Comput.* **2016**, *28*, 1217–1247.
16. Tsyurmasto, P.; Uryasev, S.; Gotoh, J. *Support Vector Classification with Positive Homogeneous Risk Functionals*; Technical Report, Research Report 2013-4; Department of Industrial and Systems Engineering, University of Florida: Gainesville, FL, USA, 2013.
17. Xu, L.; Crammer, K.; Schuurmans, D. Robust Support Vector Machine Training Via Convex Outlier Ablation. In Proceedings of the AAAI, Boston, MA, USA, 16–20 July 2006; pp. 536–542.
18. Fujiwara, S.; Takeda, A.; Kanamori, T. *DC Algorithm for Extended Robust Support Vector Machine*; Technical Report METR 2014–38; The University of Tokyo: Tokyo, Japan, 2014.
19. Takeda, A.; Fujiwara, S.; Kanamori, T. Extended robust support vector machine based on financial risk minimization. *Neural Comput.* **2014**, *26*, 2541–2569.
20. Maronna, R.; Martin, R.D.; Yohai, V. *Robust Statistics: Theory and Methods*; Wiley: Hoboken, NJ, USA, 2006.
21. Schapire, R.E.; Freund, Y.; Bartlett, P.L.; Lee, W.S. Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Stat.* **1998**, *26*, 1651–1686.
22. Kimeldorf, G.S.; Wahba, G. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **1971**, *33*, 82–95.
23. Wahba, G. *Advances in Kernel Methods;* Chapter Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV; MIT Press: Cambridge, MA, USA, 1999; pp. 69–88.
24. Takeda, A.; Sugiyama, M. ν-Support Vector Machine as Conditional Value-at-Risk Minimization. In Proceedings of the ICML, ACM International Conference Proceeding Series, Yokohama, Japan, 3–5 December 2008; Cohen, W.W., McCallum, A., Roweis, S.T., Eds.; ACM: New York, NY, USA, 2008; Volume 307, pp. 1056–1063.
25. Rockafellar, R.T.; Uryasev, S. Conditional value-at-risk for general loss distributions. *J. Bank. Financ.* **2002**, *26*, 1443–1472.
26. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
27. Le Thi, H.A.; Dinh, T.P. Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Math. Vietnam.* **1997**, *22*, 289–355.
28. Yuille, A.L.; Rangarajan, A. The concave-convex procedure. *Neural Comput.* **2003**, *15*, 915–936.
29. Crisp, D.J.; Burges, C.J.C. A Geometric Interpretation of ν-SVM Classifiers. In *Advances in Neural Information Processing Systems 12*; Solla, S.A., Leen, T.K., Müller, K.-R., Eds.; MIT Press: Cambridge, MA, USA, 2000; pp. 244–250.
30. Kanamori, T.; Takeda, A.; Suzuki, T. Conjugate relation between loss functions and uncertainty sets in classification problems. *J. Mach. Learn. Res.* **2013**, *14*, 1461–1504.
31. Takeda, A.; Mitsugi, H.; Kanamori, T. A Unified Robust Classification Model. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), ICML'12, Edinburgh, Scotland, 26 June–1 July 2012; Langford, J., Pineau, J., Eds.; Omnipress: New York, NY, USA, 2012; pp. 129–136.
32. Bertsekas, D.; Nedic, A.; Ozdaglar, A. *Convex Analysis and Optimization*; Athena Scientific: Belmont, MA, USA, 2003.
33. Donoho, D.; Huber, P. The Notion of Breakdown Point. In *A Festschrift for Erich L. Lehmann*; CRC Press: Boca Raton, FL, USA, 1983; pp. 157–184.
34. Hampel, F.R.; Rousseeuw, P.J.; Ronchetti, E.M.; Stahel, W.A. *Robust Statistics. The Approach Based on Influence Functions;* John Wiley and Sons, Inc.: Hoboken, NJ, USA, 1986.
35. Huber, P.J.; Ronchetti, E.M. *Robust Statistics*, 2nd ed.; Wiley: New York, NY, USA, 2009.
36. Christmann, A.; Steinwart, I. On robustness properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.* **2004**, *5*, 1007–1034.
37. Le Thi, H.A.; Dinh, T.P. The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems. *Ann. Oper. Res.* **2005**, *133*, 23–46.

38. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
39. Steinwart, I. On the optimal parameter choice for ν-support vector machines. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1274–1284.
40. Wu, Y.; Liu, Y. Adaptively weighted large margin classifiers. *J. Comput. Graph. Stat.* **2013**, *22*, 416–432.