Extended Robust Support Vector Machine Based on Financial Risk Minimization

Akiko Takeda
takeda@mist.i.u-tokyo.ac.jp
Shuhei Fujiwara
shuhei_fujiwara@mist.i.u-tokyo.ac.jp
Department of Mathematical Informatics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

Takafumi Kanamori
kanamori@is.nagoya-u.ac.jp
Department of Computer Science and Mathematical Informatics, Nagoya University, Chikusa-ku, Nagoya-shi, Aichi 464-8603, Japan

Financial risk measures have been used recently in machine learning. For example, $\nu$-support vector machine ($\nu$-SVM) minimizes the conditional value at risk (CVaR) of margin distribution. The measure is popular in finance because of the subadditivity property, but it is very sensitive to a few outliers in the tail of the distribution. We propose a new classification method, extended robust SVM (ER-SVM), which minimizes an intermediate risk measure between the CVaR and value at risk (VaR) by expecting that the resulting model becomes less sensitive than $\nu$-SVM to outliers. We can regard ER-SVM as an extension of robust SVM, which uses a truncated hinge loss. Numerical experiments imply the ER-SVM's possibility of achieving a better prediction performance with proper parameter setting.

1 Introduction

One important goal of classification methods is to construct classifiers with high prediction accuracy, that is, good generalization ability. The support vector machine (SVM) is designed to reduce the risk of misprediction for test data by means of the combination of a regularization term and a loss function that evaluates the fitting to training data. Risk minimization techniques are studied in financial mathematics field as well. Especially when considering long-term contracts, various risks are present in practice, and an effective way of hedging those risks is needed. One of the most widely used risk measures in finance is value at risk (VaR), a quantile at a pre-defined probability level. A coherent risk measure such as conditional VaR (CVaR) (Rockafellar & Uryasev, 2002; Artzner, Delbaen, Eber, & Heath, 1999)

Neural Computation 26, 2541–2569 (2014) © 2014 Massachusetts Institute of Technology
doi:10.1162/NECO_a_00647
Figure 1: The relation of ER-SVM to existing learning models shown in solid boxes.

became popular because of the subadditivity property, which basically means that “a merger does not create extra risk,” though the CVaR is very sensitive to a few outliers in the tail of the distribution.

Several works have studied financial risk measures from a machine learning perspective. For example, Xu, Caramanis, Mannor, and Yun (2009) proposed a comprehensive robust classification model that uses a discounted loss function depending on the data and investigated the relationship between comprehensive robustness and convex risk measures. Takeda and Sugiyama (2008) showed that \( \nu \)-SVM (Schölkopf, Smola, Williamson, & Bartlett, 2000) and its extended model, \( \nu \)-SVM (Perez-Cruz, Weston, Hermann, & Schölkopf, 2003), minimize the CVaR of margin distribution.

In this letter, we propose a new classification method, which we call ER-SVM, that minimizes a truncated CVaR by using VaR by expecting that the resulting model becomes less sensitive to outliers than \( \nu \)-SVM. Our model is closely related to robust SVM (Shen, Tseng, Zhang, & Wong, 2003; Xu, Crammer, & Schuurmans, 2006; Wu & Liu, 2007; Brooks, 2011), which minimizes a truncated hinge loss combined with a regularization term. We can say that ER-SVM is an extended variant of robust SVM (see Figure 1). More precisely, the permissible parameter range of robust SVM is included in that of ER-SVM, and therefore, ER-SVM may achieve a better prediction performance with proper parameter setting.

ER-SVM is formulated as a nonconvex problem and difficult to solve exactly, as well as existing robust SVMs. We propose a heuristic algorithm for solving ER-SVM approximately, which finds a feasible solution of ER-SVM in every iteration. Furthermore, when the algorithm is running, a hyperparameter \( \nu_O \) of ER-SVM is set to an appropriate value. If \( \nu_O \) of ER-SVM is set to the ratio of outliers, it achieves good prediction performance, but it is hard to predict the ratio in practical problem setting. Therefore, the proposed algorithm, where the parameter \( \nu_O \) is automatically tuned, is very practical. In the algorithm, we repeat two steps: solving optimization problems of \( \nu \)-SVM or \( \nu \)-SVM and removing training samples with large losses. The solution method is easy to implement because we can use existing tools and software for solving \( \nu \)-SVM or \( \nu \)-SVM several times in order to solve ER-SVM.
Numerical experiments show the superior performance of ER-SVM over robust SVM, C-SVM, and $\nu$-SVM in the presence of outliers. Indeed, Figure 2 shows that CVaR minimization, which is equivalent to $\nu$-SVM, is sensitive to outliers, whereas our model, ER-SVM, is not sensitive. The feature of our model that ignores samples with large losses contributes to the superior performance over $\nu$-SVM.

The superior performance over robust SVM implies the effectiveness of ER-SVM’s extended parameter range relative to the range of robust SVM. Both problems, ER-SVM and robust SVM, can be regarded as the minimum distance problem to a set $\mathcal{U}$ that consists of training samples, and hyperparameters of those problems control the size of $\mathcal{U}$. Robust SVM limits the range of the hyperparameter so that $\mathcal{U}$ does not contain 0 but ER-SVM removes the limitation.

The letter is organized as follows. Section 2 reviews several related support vector machine classifiers and risk measure minimization models in financial engineering. Section 3 presents the formulation of our model, ER-SVM, and a heuristic algorithm for ER-SVM. Section 4 provides geometric interpretations for ER-SVM by showing the dual formulation of ER-SVM. Those interpretations help us to understand ER-SVM from the geometric viewpoint and recognize the difference of ER-SVM from VaR or CVaR-based methods. In section 5, our model is compared to several related models such as robust SVM (Xu et al., 2006; Wu & Liu, 2007), $\nu$-SVM (Schölkopf...
et al., 2000), \(\nu\)-SVM (Perez-Cruz et al., 2003), and VaR-SVM (Tsyurmasto, Zabarankin, & Uryasev, 2014). Section 6 concludes the letter.

2 Background and Related Work

2.1 Support Vector Machine. The SVM has been widely used for classification in machine learning. Let us address the binary classification problem of learning a linear function \(h(x) = x^\top \mathbf{w} + b\) based on training samples \((x_i \in \mathcal{X}, y_i \in \{-1, 1\}), i \in I\). We assume that the training samples are independent and identically distributed following the unknown probability distribution on \(\mathcal{X} \times \{-1, 1\}\). For simplicity, we generally focus on linear functions \(h(x)\), but the discussions in this letter can be directly applicable to nonlinear kernel classifiers (see Schölkopf & Smola, 2002).

2.1.1 Equivalence Between C-SVM and \(\nu\)-SVM. One important goal of learning methods is to construct classifiers with high prediction accuracy, that is, good generalization ability. For that purpose, SVMs minimize the objective function, which consists of a loss function \(\sum_{i \in I} \ell(\mathbf{w}, b; x_i, y_i)\) and regularization term (such as \(\|\mathbf{w}\|^2\)). Many learning methods use convex surrogate losses (e.g., hinge loss):

\[
\ell_\zeta(\mathbf{w}, b; x_i, y_i) := [-y_i(x_i^\top \mathbf{w} + b) + \zeta]^+, \tag{2.1}
\]

where \([X]^+ := \max\{X, 0\}\) for \(\zeta \geq 0\). A representative model using the hinge loss is C-SVM (Cortes & Vapnik, 1995),

\[
\min_{\mathbf{w}, b} \frac{C}{2} \|\mathbf{w}\|^2 + \sum_{i \in I} [\mathbf{y}_i(x_i^\top \mathbf{w} + b) + 1]^+, \quad \text{C-SVM}
\]

where \(C \in (0, \infty)\) is a user-specified hyperparameter. Moreover, \(\nu\)-SVM (Schölkopf et al., 2000) is another formulation with the hinge loss,

\[
\min_{\mathbf{w}, b, \rho} \frac{1}{2} \|\mathbf{w}\|^2 - \nu \rho + \frac{1}{|I|} \sum_{i \in I} [-y_i(w^\top x_i + b) + \rho]^+, \quad \text{\(\nu\)-SVM}
\]

where \(\nu \in (0, 1]\) is a hyperparameter and \(|I|\) shows the cardinality of the set \(I\). The margin \(\rho\) in the hinge loss is determined by an optimal solution of \(\nu\)-SVM, and the resulting margin is nonnegative (see Crisp & Burges, 2000). \(\nu\)-SVM and C-SVM have the same optimal solution if we appropriately set two parameters \(\nu\) and \(C\) by \(C = |I|\rho^*\) using the optimal solution of \(\nu\)-SVM (see Schölkopf et al., 2000). It is said that setting appropriate \(\nu\) for \(\nu\)-SVM is easier and more intuitive than \(C\) of C-SVM because of \(\nu\)-properties of
ν-SVM, implying that ν is an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors.

However, ν of ν-SVM has the permissible range as ν ∈ (ν̄, νmax] (see Crisp & Burges, 2000 for details). ν-SVM is unbounded when ν is larger than νmax := 2 min(|I+|, |I−|)/|I|, where I+ (resp. I−) is the index set of the samples with the positive (resp. negative) label. ν-SVM produces a trivial solution (w = 0 and b = 0) when ν is smaller than some threshold ν̄.

2.1.2 Extension from ν-SVM to Eν-SVM. ν-SVM was extended to Eν-SVM (Perez-Cruz et al., 2003) by allowing the margin ρ to be negative and enforcing the norm of w to be unity,

$$\min_{w \in W, b, \rho} -\nu \rho + \frac{1}{|I|} \sum_{i \in I} [-y_i (w^\top x_i + b) + \rho]^+, \quad (2.2)$$

where W = {w : ||w||^2 = 1}. It should be noted that the convex relaxation problem, which replaces w ∈ W of Eν-SVM (see equation 2.2) with

$$w \in \text{conv}(W) := \{w : \|w\|^2 \leq 1\},$$

is equivalent to ν-SVM. Indeed, dual problems of ν-SVM and the relaxation problem with w ∈ conv(W) coincide (up to a scaling of variables). The nonconvex constraint w ∈ W prevents w from being zero for smaller ν than ν̄, and Eν-SVM gives a nontrivial solution satisfying ν-properties for ν ∈ (0, νmax].

The classifiers of Eν-SVM and ν-SVM are the same up to a scaling factor when ν > ν̄ because the convex relaxation problem of Eν-SVM attains an optimal solution at ||w||^2 = 1. In other words, w ∈ W can be relaxed to w ∈ conv(W) without changing the optimal solution. On the other hand, in the case of ν < ν̄, the relaxation problem gives a trivial solution (w = 0 and b = 0) as well as ν-SVM. In that case, w ∈ W of Eν-SVM can be relaxed to ||w||^2 ≥ 1, but it is still nonconvex. Therefore, a nonconvex optimization method (Perez-Cruz et al., 2003; Takeda & Sugiyama, 2008) needs to be applied to Eν-SVM.

2.1.3 Robust SVM. These classification models deteriorate the prediction performance in the presence of outliers. To overcome such drawbacks, various papers (e.g., Shen et al., 2003; Xu et al., 2006; Wu & Liu, 2007; Brooks, 2011) proposed robust SVM, which uses a truncated hinge loss. The truncated loss implies that the loss does not increase the penalty after a certain
point. Such a loss can improve the robustness to outliers at the expense of convexity. Using the \( \eta \)-hinge loss\(^1\),
\[
\ell_{\eta, \zeta}^s (w, b; x, y) := \eta[-y(x^T w + b) + \zeta]^+ + s(1 - \eta)
\] (2.3)

with parameters \( \eta \in [0, 1] \) and \( s \geq 0 \) in addition to \( \zeta \geq 0 \), Xu et al. (2006) formulated robust SVM as
\[
\min_{w, b, \eta \in [0, 1]} \frac{C}{2} \| w \|^2 + \sum_{i \in I} \ell_{\eta, i}^s (w, b; x_i, y_i).
\] (2.4)

Here \( C (> 0) \) is a user-specified positive parameter. The optimized \( \eta \)-hinge function is shown to be equivalent to the truncated hinge loss in (Xu et al., 2006):
\[
\min_{\eta \in [0, 1]} \ell_{\eta, i}^1 (w, b; x, y) = \min \{ s, [-y(x^T w + b) + 1]^+ \}.
\]

Xu et al. (2006) solved equation 2.4 by using semidefinite relaxation. Wu and Liu (2007) also described the truncated hinge loss by the difference of two hinge loss functions and solved the resulting SVM problem by applying the difference convex (d.c.) algorithm through a sequence of convex subproblems.

### 2.2 Risk Measure Minimization.

We define financial risk measures using the losses of samples; \( f(w, b; x_i, y_i) := -y_i(x_i^T w + b) \), \( i \in I \) and show classification methods that minimize risk measures with respect to \( w \in W \) and \( b \).

Assuming that \( (x, y) \) is a random vector with discrete support \( \{(x_i, y_i), i \in I\} \) consisting of training samples, let us consider the distribution of losses over these samples:
\[
\Phi(\alpha | w, b) := \frac{1}{|I|} \left| \{ i \in I | f(w, b; x_i, y_i) \leq \alpha \} \right|.
\]

For \( \nu \in (0, 1] \), let \( \alpha_{1-\nu} (w, b) \) be the 100(1 - \( \nu \))-percentile of the loss distribution, known as the value at risk, or \((1 - \nu)\)-VaR in finance:
\[
\alpha_{1-\nu} (w, b) := \min \{ \alpha | \Phi(\alpha | w, b) \geq 1 - \nu \}.
\]

\(^1\)Xu et al. (2006) proposed the \( \eta \)-hinge loss with \( s = 1 \) and \( \zeta = 1 \), while Wu and Liu (2007) implicitly used the \( \eta \)-hinge loss with \( s > 0 \). Note that \( \eta \)-hinge (see equation 2.3) with \( \eta = 1 \) coincides with the ordinary hinge loss, equation 2.1. Yu, Yang, Xu, White, and Schuurmans (2010) discussed truncating standard loss functions as well as the hinge loss.
Figure 3: An example of the distribution of losses $f(w, b; x, y)$ over all training samples.

Thus, the fraction $\nu$ of the losses $f(w, b; x_i, y_i), i \in I$, exceeds the threshold $\alpha_{1-\nu}(w, b)$ (see Figure 3). VaR minimization is formulated as follows:

$$\begin{align*}
\min_{w, b, \alpha, J} \alpha \\
\text{s.t. } w \in W, f(w, b; x_i, y_i) \leq \alpha, \quad i \in I - J, \quad |J| \leq \nu |I|, \quad J \subseteq I.
\end{align*}$$

(2.5)

We define the $(1 - \nu)$-tail distribution as

$$\Phi_{1-\nu}(\alpha | w, b) := \begin{cases} 0, & \text{for } \alpha < \alpha_{1-\nu}(w, b), \\ \frac{\Phi(\alpha | w, b) - (1 - \nu)}{\nu}, & \text{for } \alpha \geq \alpha_{1-\nu}(w, b). \end{cases}$$

Let $\phi_{1-\nu}(w, b)$ be the mean of the $(1 - \nu)$-tail distribution of $f(w, b; x_i, y_i), i \in I$ (see Figure 3):

$$\phi_{1-\nu}(w, b) := \mathbb{E}_{\Phi_{1-\nu}}[f(w, b; x, y)],$$

where $\mathbb{E}_{\Phi_{1-\nu}}$ denotes the expectation over the distribution $\Phi_{1-\nu}$. $\phi_{1-\nu}(w, b)$ is called the conditional VaR, or $(1 - \nu)$-CVaR. By definition, the CVaR, $\phi_{1-\nu}(w, b)$, is always larger than or equal to the VaR, $\alpha_{1-\nu}(w, b)$. Indeed, Rockafellar and Uryasev (2002) proved the relation between the CVaR $\phi_{1-\nu}(w, b)$ and the VaR $\alpha_{1-\nu}(w, b)$ as

$$\begin{align*}
\mathbb{E}_{\Phi}[f(w, b; x, y) | \alpha_{1-\nu}(w, b) \leq f(w, b; x, y)] &\leq \phi_{1-\nu}(w, b) \\
&\leq \mathbb{E}_{\Phi}[f(w, b; x, y) | \alpha_{1-\nu}(w, b) < f(w, b; x, y)]
\end{align*}$$

(2.6)
Figure 4: Profiles of minimum CVaR and minimum VaR as functions of \( \nu \).

"Convex" means that the feasible region is \( \text{conv}(W) \), while "nonconvex" means \( W \).

using the expectation \( \mathbb{E}_\Phi \) over the distribution \( \Phi \). If there is no probability atom at \( \alpha_{1-\nu}(w, b) \), all three terms in equation 2.6 are equal, and \( \phi_{1-\nu}(w, b) \) is the expected loss given that the loss is greater than or equal to the VaR.

Minimizing the CVaR, \( \phi_{1-\nu}(w, b) \), is shown in Rockafellar and Uryasev (2002) to be equivalent to

\[
\min_{w \in W, b, \alpha} \alpha + \frac{1}{\nu|I|} \sum_{i \in I} [f(w, b; x_i, y_i) - \alpha]^+.
\]  

(2.7)

\( \alpha^* \) in its optimal solution \( (w^*, b^*, \alpha^*) \) is almost equal to \( (1 - \nu)\)-VaR, \( \alpha_{1-\nu}(w^*, b^*) \) (see Rockafellar & Uryasev, 2002).

Note that equation 2.7 reduces to \( \nu \)-SVM equation 2.2, by changing a variable as \( \alpha = -\rho \). Let \( (w_\rho^*, b_\rho^*, \alpha_\rho^*) \) be an optimal solution for equation 2.7, with \( \nu \leq \nu_{\text{max}} \). As Figure 4 shows, \( \phi_{1-\nu}(w_\rho^*, b_\rho^*) \) is decreasing with respect to \( \nu \). We can compute the threshold of convexity, \( \nu_{\phi} \), such that \( \phi_{1-\nu}(w_\rho^*, b_\rho^*) = 0 \) by solving a linear programming problem as shown in Takeda, Mitsugi, and Kanamori (2013). For \( \nu \in (\nu_{\phi}, \nu_{\text{max}}) \), equation 2.7 reduces to a convex problem with \( \text{conv}(W) \) instead of \( W \). As discussed before, the convex problem is equivalent to \( \nu \)-SVM (Schölkopf et al., 2000). A classification model minimizing VaR with a convex constraint, \( w \in \text{conv}(W) \), (i.e., the convex relaxation problem of equation 2.5 with respect to \( w \)) was studied in Tsyurmasto et al. (2014) as VaR-SVM.

3 Extended Robust SVM

The CVaR and VaR measures have strong points and weak points when applied to classification problems. Note that the VaR measure ignores \( \lfloor \nu|I| \rfloor \) samples with large losses. Minimizing the VaR gives us a robust estimate of
Extended Robust Support Vector Machine 2549

Extended Robust Support Vector Machine \((w, b)\) that is insensitive to outliers in the data. However, the performance is sensitive to the input parameter \(v\), which controls the ratio of ignored samples. We can easily discard essential training samples by regarding them as outliers and achieve worse prediction results. The CVaR minimization model, including \(\nu\)-SVM, has an excellent reputation in its performance in standard setup. However, the estimator is sensitive to outliers.

3.1 ER-SVM Formulation. We propose a new classification model that uses both CVaR and VaR measures for their advantages,

\[
\min_{w \in W, b, \alpha, J \subseteq I, |J| \leq \nu O |I|} \alpha + \frac{1}{\nu I \tilde{m}_{\nu O}} \sum_{i \in J} [ -y_i (x_i^T w + b) - \alpha ]^+ ,
\]

where \(\tilde{m}_{\nu O} = |I| - |\nu O |I||. We call it extended robust SVM (ER-SVM). It can be reformulated as

\[
\min_{w, b, \alpha, \eta} \alpha + \frac{1}{\nu I \tilde{m}_{\nu O}} \sum_{i \in I} \eta_i [ -y_i (x_i^T w + b) - \alpha ]^+
\]

\[
\text{s.t. } w \in W, \eta \in \{0, 1\}^{|I|}, \sum_{i \in I} (1 - \eta_i) \leq \nu O |I|.
\]

Note that optimal solutions \(J^*\) for equation 3.1 and \(\eta^*_i = 0\) for equation 3.2 correspond to each other so that \(\{i \in I : \eta^*_i = 0\} = J^*\).

We can regard ER-SVM as a mixture model of CVaR and VaR minimization. The model has two input parameters, \(\nu O\) and \(\nu I\), in the range \((0, 1)\). When \(\nu O\) becomes small enough so that \(\nu O < \frac{1}{|I|}\), \(J\) becomes empty, and equation 3.1 is the same as CVaR minimization, equation 2.2, with \(v = \nu I\). When \(\nu I\) becomes small enough so that \(\nu I < \frac{1}{|I||J|}\), the objective function in equation 3.1 reduces to \(\max_{i \in J} -y_i (x_i^T w + b)\), and therefore equation 3.1 is the same as VaR minimization, equation 2.5, with \(v = \nu O\) (we can prove it easily from corollary 9 in Rockafellar & Uryasev, 2002).

The measure used in equation 3.1 almost (in the sense of equation 2.6) equals

\[
\mathbb{E}_q [ -y(x^T w + b) | \alpha_{1-v} (w, b) \leq -y(x^T w + b) \leq \alpha_{1-\nu O} (w, b) ] ,
\]

where \(\tilde{v} = (\nu I \tilde{m}_{\nu O} + |\nu O |I||)/|I|\). We call it truncated CVaR (trCVaR) and denote it by \(\text{tr-} \phi_{\nu I}^{1-\nu O}\) because equation 3.3 adds an upper bound, \(\alpha_{1-\nu O} (w, b)\), to equation 2.6 of the CVaR. The measure indicates the average of remaining \([\nu I \tilde{m}_{\nu O}]\) large losses after ignoring \(|\nu O |I||\) extremely large losses.
To ensure that equation 3.1 has an optimal solution, the range of $\nu_I$ is restricted depending on $J^2$, while $\nu_O$ can take $(0, 1)$. The upper bound of $\nu_I$ is determined by $\nu_{\text{max}}^I = 2 \min\{|I_+ \setminus J|, |I_- \setminus J|\} / |I \setminus J|$.

When $\nu_I > \nu_I^{\text{max}}$, equation 3.1 becomes an unbounded problem. Note that the optimal value of equation 3.1, $\text{tr-} \phi_1^{1-\nu_O}$, is decreasing with respect to $\nu_I$. We can compute $\nu_I^{\phi}(> 0)$ as in the case of Ev-SVM so that the optimal $\text{tr-} \phi_1^{1-\nu_O}$ is negative for $\nu_I \in (\nu_I^{\phi}, \nu_I^{\text{max}}]$, which we call the convex range for $\nu_I$.

Now we modify ER-SVM, equation 3.1, by replacing the nonconvex set $W$ by $\text{conv}(W)$ and call the resulting problem convex ER-SVM.

Lemma 1. As far as $\nu_I$ is in the convex range, ER-SVM, equation 3.1 is equivalent to convex ER-SVM.

Proof. To prove the lemma, it is enough to show that the set $W$ can be relaxed by $\text{conv}(W)$ without changing the optimality as far as the optimal value of ER-SVM, equation 3.1, is negative. Suppose on the contrary that convex ER-SVM has an optimal solution $(w^*, b^*, \alpha^*, J^*)$ satisfying $\|w^*\|^2 < 1$, although the optimal value of ER-SVM is negative. Then we have

$$\frac{1}{\|w^*\|} (\alpha^* + \frac{1}{\nu_I^{\text{max}} \nu_O} \sum_{i \in J^*} [-y_i (x_i^\top w^* + b^*) - \alpha^*]^+)$$

$$< \min_{w \in \text{conv}(W), b, \alpha, J \subseteq I, |J| \leq \nu_O |I|} \alpha + \frac{1}{\nu_I^{\text{max}} \nu_O} \sum_{i \in J} [-y_i (x_i^\top w + b) - \alpha]^+ < 0.$$  

Note that $(\frac{w^*}{\|w^*\|}, \frac{b^*}{\|w^*\|}, \frac{\alpha^*}{\|w^*\|}, J^*)$ is a feasible solution of the convex ER-SVM. The above strict inequalities show that the solution achieves a smaller objective value than $(w^*, b^*, \alpha^*, J^*)$, which contradicts the optimality of the convex ER-SVM. Therefore, $\|w^*\|^2 = 1$ holds, and this proves the lemma.

Lemma 1 implies that for $\nu_I \in (\nu_I^{\phi}, \nu_I^{\text{max}}]$, convex ER-SVM gives an optimal solution of ER-SVM, equation 3.1. But, when $\nu_I \leq \nu_I^{\phi}$, convex ER-SVM will lead to a trivial solution $(w = 0$ and $b = 0)$.

Choosing $\nu_I$ from the range depending on $J$ seems to be unrealistic to solve equation 3.1, because $J$ is a variable in that equation. However, our solution method, algorithm 3.3, repeatedly solves equation 3.1 by fixing the set $J$. $J$ is increased from $\{\emptyset\}$ to a set with size $[\nu_O |I|]$. Therefore, we can check whether $\nu_I$ is an appropriate parameter.

---

2Choosing $\nu_I$ from the range depending on $J$ seems to be unrealistic to solve equation 3.1, because $J$ is a variable in that equation. However, our solution method, algorithm 3.3, repeatedly solves equation 3.1 by fixing the set $J$. $J$ is increased from $\{\emptyset\}$ to a set with size $[\nu_O |I|]$. Therefore, we can check whether $\nu_I$ is an appropriate parameter.
The following theorem implies that ER-SVM, equation 3.1, is an extension of robust SVM, equation 2.4, because ER-SVM equals robust SVM only when \( \nu_I \) is in the convex range.

**Theorem 1.** Convex ER-SVM, robust SVM (see equation 2.4), and robust \( \nu \)-SVM

\[
\min_{w, b, \alpha, \eta \in \{0, 1\}^{|I|}} \frac{1}{2} \|w\|^2 + \alpha + \frac{1}{|I|} \sum_{i \in I} \eta_i \left[ -y_i(x_i^T w + b) - \alpha \right]^+ + \bar{s}(1 - \eta_i),
\]

(3.4)

are equivalent by setting \( C = -\alpha^* |I|, s = \frac{s}{\alpha^*} \) of equation 2.4 and \( \nu_O = \sum_{i \in I} (1 - \eta_i^*) \) and \( \nu_I = \frac{|I|}{\bar{s}} \) of equation 3.1 with the use of an optimal solution \((w^*, b^*, \alpha^*, \eta^*)\) of equation 3.4 for \( \bar{s} \geq 0 \) and \( \nu \) from the convex range.

**Proof.** We focus on the case where \( \nu \) is chosen from the convex range, that is the feasible solution \((w, b, \alpha, \eta) = (0, 0, 0, 1)\) is not optimal to robust \( \nu \)-SVM, equation 3.4 (here, \( \mathbf{1} \) indicates the all-ones vector). At first, we show that the optimal solution \((w^*, b^*, \alpha^*, \eta^*)\) of robust \( \nu \)-SVM, equation 3.4 is also optimal to the problem

\[
\min_{w, b, \alpha, \eta} \alpha + \frac{1}{|I|} \sum_{i \in I} \eta_i \left[ -y_i(x_i^T w + b) - \alpha \right]^+ \tag{3.5}
\]

s.t. \( \|w\| \leq \|w^*\|, \eta \in \{0, 1\}^{|I|}, \sum_{i \in I} (1 - \eta_i) \leq \sum_{i \in I} (1 - \eta_i^*) \).

If \((w^*, b^*, \alpha^*, \eta^*)\) is not optimal, equation 3.5 must have an optimal solution \((\tilde{w}, \tilde{b}, \tilde{\alpha}, \tilde{\eta})\) such as

\[
\tilde{\alpha} + \frac{1}{|I|} \sum_{i \in I} \tilde{\eta}_i \left[ -y_i(x_i^T \tilde{w} + \tilde{b}) - \tilde{\alpha} \right]^+ < \alpha^* + \frac{1}{|I|} \sum_{i \in I} \eta_i^* \left[ -y_i(x_i^T w^* + b^*) - \alpha^* \right]^+, \tag{3.6}
\]

\[
\|\tilde{w}\| \leq \|w^*\|, \sum_{i \in I} (1 - \tilde{\eta}_i) \leq \sum_{i \in I} (1 - \eta_i^*). \tag{3.7}
\]

However, inequalities in equations 3.6 and 3.7 imply that \((\tilde{w}, \tilde{b}, \tilde{\alpha}, \tilde{\eta})\) achieves a smaller objective value of equation 3.4 than \((w^*, b^*, \alpha^*, \eta^*)\), which contradicts the optimality of equation 3.4, and therefore, \((w^*, b^*, \alpha^*, \eta^*)\) is...
optimal to equation 3.5. Moreover, $(\|w^*\|, \frac{\nu}{\|w^*\|}, \frac{\alpha^*}{\|w^*\|}, \eta^*)$ is optimal to

$$\min_{w,b,\alpha,\eta} \alpha + \frac{1}{\nu} \sum_{i \in I^m} \eta_i \left[ -y_i (x_i^T w + b) - \alpha \right]^+$$

(3.8)

s.t. $w \in \text{conv}(W), \eta \in \{0, 1\}^{|I|}, \sum_{i \in I} (1 - \eta_i) \leq \nu |I|,$

which is the convex relaxation of ER-SVM, equation 3.2, that is, convex ER-SVM. We could show the equivalence between convex ER-SVM, equation 3.8, and robust $\nu$-SVM, equation 3.4.

Finally, we relate robust $\nu$-SVM, equation 3.4 to robust SVM equation 2.4. As far as $\nu$ is in the convex range, the optimal value, equation 3.5, is negative. Note that optimal $\alpha^*$ is also negative because it is less than or equal to the optimal value. If we fix $\alpha$ of equation 3.4 to $\alpha^*$ and minimize the objective function only over the remaining variables, nothing will change. By rescaling the set of variables of equation 3.4 by $-\alpha^*$ as $\hat{w} = \frac{w}{-\alpha^*}, \hat{b} = \frac{b}{-\alpha^*},$ we describe the objective function of equation 3.4 divided by $-\alpha^*$ as

$$-\frac{\alpha^*}{2} \|w\|_2^2 - 1 + \frac{1}{\nu |I|} \sum_{i \in I^m} \ell_{\eta_i,1} \left( \frac{x_i}{\alpha^*}, \frac{b}{-\alpha^*}; x_i, y_i \right)$$

$$= -\frac{\alpha^*}{2} \|\hat{w}\|_2^2 - 1 + \frac{1}{\nu |I|} \sum_{i \in I^m} \ell_{\eta_i,1} \left( \hat{w}, \hat{b}; x_i, y_i \right).$$

We can confirm that robust $\nu$-SVM, equation 3.4, reduces to robust SVM, equation 2.4, by setting $\nu, \tilde{s}$ of equation 3.4 and $C, s$ of equation 2.4, appropriately.

The theorem shows that ER-SVM is equivalent to robust SVM when $\nu^I \in (\nu^I_\phi, \nu^I_{\text{max}}].$ If $\nu^I$ of ER-SVM is in the range, the nonconvex constraint $w \in W$ can be relaxed as $w \in \text{conv}(W).$ when $0 < \nu^I < \nu^I_\phi,$ convex ER-SVM will achieve a trivial solution $(w = 0$ and $b = 0).$ In such a case, the non-convex constraint of ER-SVM is equivalent to $\|w\|_2^2 \geq 1,$ because for the inequality constraint, $\|w^*\|_2^2 = 1$ holds at the optimal solution, $w^*,$ as long as the optimal value of ER-SVM is positive, similar to the properties of $E_{\nu}$-SVM shown in section 2.1.2. By extending the permissible range of $\nu^I$ from $(\nu^I_\phi, \nu^I_{\text{max}}]$ to $[0, \nu^I_{\text{max}}],$ the margin is allowed to be negative as well as
Extended Robust Support Vector Machine 2553

Ev-SVM, equation 2.2, and as a result, ER-SVM can provide a nontrivial solution that robust SVM can not attain.

3.2 ER-SVM Heuristic Algorithm. When we use ER-SVM, equation 3.1, for classification, it seems troublesome to handle two parameters of ER-SVM, $\nu_O$ and $\nu_T$. If $\nu_O$ is set to the ratio of outliers, ER-SVM will achieve good prediction performance, but it is hard to predict the ratio in practical problem setting. In this section, we propose a heuristic algorithm for ER-SVM, which automatically tunes parameters of ER-SVM: $\nu_O$ and $\nu_T$ during execution. It makes parameter settings convenient if we cannot estimate the ratio $\nu_O$ of outliers in data sets.

The algorithm minimizes the objective value of ER-SVM with respect to $J$ and other variables alternately, gradually increasing the cardinality of $J$ from 0 to $\nu|I|$. When we use the idea of Larsen, Mausser, and Uryasev (2002), the algorithm observes the change of the solution with respect to increasing the cardinality of $J$ and stops the execution when the solution does not change much.

Larsen et al.'s (2002) efficient heuristic algorithm was proposed for the VaR minimization problem equation 2.5. We modify the algorithm for solving trCVaR minimization problem. Algorithm A1 in Larsen et al. (2002) approximately solves the $(1 - \nu)$-VaR minimization problem. The general line of thought behind the heuristic algorithm is simple. The algorithm systematically reduces the VaR by repeating two phases (starting with $k = 0$, $\nu_0 = \nu$ and $I_0 = I$): (1) solve a $(1 - \nu_k)$-CVaR problem using data set $I_k \subseteq I$, and (2) update $I_{k+1}$ so as to remove samples with large losses in the CVaR problem and reset $\nu_{k+1}$ so that $(1 - \nu_{k+1})|I_{k+1}| = (1 - \nu)|I|$ (the number of samples with zero penalty, shown in the white area in Figure 3, is the same). These phases are repeated until $|\nu||I|$ samples are removed.

We propose a heuristic algorithm, algorithm 1, for ER-SVM by modifying the algorithm of Larsen et al. (2002).\footnote{We constructed our algorithm by stopping (Larsen et al., 2002) algorithm in the middle when the solution of equation 3.9 does not change much from the previous solution and by slightly changing the CVaR problem. The original problem in Larsen et al. (2002) includes additional constraints to ensure that the resulting solution $\alpha$ will be the VaR. Since our purpose for using their method is not to find a minimized VaR, we removed those constraints. This modification makes the implementation of the algorithm easy. Indeed, we can use existing software and algorithms for solving $\nu$-SVM and Ev-SVM.} We use a parameter $\gamma \in (0, 1)$ to control the number of samples to be discarded in each iteration in addition to $\nu \in (0, \nu_{\text{max}})$. If we do not care about computation time, it is better to set a small value for $\gamma$. $C_k := [(1 - \nu + v(1 - \nu)^k)||I||] = |I_k|$ indicates the number of data samples used in the $k$th CVaR problem. We define $\nu_k := \frac{v(1-\nu)^k||I||}{|I_k|}$ for equation 3.9 in order to indicate the ratio of support vectors to $|I_k|$. As
Algorithm 1: ER-SVM Heuristic Algorithm.

**Step 0:** Assign a value for $\nu \in (0, \nu_{\text{max}}]$ and small values for $\gamma \in (0, 1)$ and $\epsilon > 0$. Let $k = 0$, $I_k = I$, $w_{-1}$ be arbitrary satisfying $\|w_{-1}\| = 1$ and $C_k := [(1 - \nu + \nu(1 - \gamma)^k)|I|]$. 

**Step 1:** Solve CVaR minimization with $\nu_k := \frac{\nu(1 - \gamma)^k|I|}{|I_k|}$:

$$\min_{w \in W, b, \alpha} \alpha + \frac{1}{\nu_k|I_k|} \sum_{i \in I_k} [-y_i(x_i^T w + b) - \alpha]^+. \quad (3.9)$$

If it is unbounded, stop the algorithm. Otherwise, let the optimal solution be $(w_k, b_k, \alpha_k)$.

**Step 2:** When $C_{k+1} = C_k$ or $|w_{k-1}^T w_k - 1| \leq \epsilon$, stop the algorithm.

**Step 3:** Select $C_{k+1}$ samples from $x_i$, $i \in I$, with small losses: $-y_i(x_i^T w_k + b_k) - \alpha_k$, and let the index set be $I_{k+1}$. Let $k = k + 1$ and go to step 1.

$k$ becomes larger, $\nu(1 - \gamma)^k$ becomes smaller, and therefore the algorithm terminates within finite iterations—more concretely, less than

$$\frac{1}{\log(1 - \gamma)} \log \frac{1 - \gamma}{\nu(2 - \gamma)|I|}.$$ 

If the iteration $k$ exceeds the above number, $C_{k+1} = C_k$ holds.

$|w_{k-1}^T w_k - 1|$ in the termination criterion shows how different the $k$th solution is from the previously obtained solution. When the solution does not change much even if samples with large losses are removed, this algorithm recognizes that all outliers were already removed and stops automatically. The solution in the first iteration is optimal to $(1 - \nu)$-CVaR minimization, equation 2.2. When the algorithm iterates sufficiently by setting $\epsilon = 0$, the resulting solution approximates an $(1 - \nu)$-VaR minimizer. Algorithm 1 gives an intermediate solution between the VaR minimizer and CVaR minimizer; it gets close to VaR minimizer from the CVaR minimizer as the iteration proceeds.

The algorithm automatically sets suitable values to $\nu_O$ and $\nu_I$ of ER-SVM. The following theorem implies which values are set to $\nu_O$ and $\nu_I$ for ER-SVM in every iteration of the algorithm.

**Theorem 2.** A feasible solution of equation 3.9 in each iteration is also feasible for equation 3.1 with $\nu_O = \nu(1 - (1 - \gamma)^k)$ and $\nu_I = \nu_k$.

**Proof.** It is enough to check that equation 3.9 equals equation 3.1 with such $\nu_O$ and $\nu_I$ when $J$ is fixed to $I \setminus I_k$. Noticing that $|I \setminus I_k| = |I| - [(1 - \nu + \nu(1 - \gamma)^k)|I|] = [\nu_O|I|]$, we see that feasible solutions of equation 3.9 are also feasible for equation 3.1.
Especially if \( I_k \) of equation 3.9 equals to \( I \setminus J^* \) for the optimal solution \( J^* \) of equation 3.1, the optimal solution \((w^*, b^*)\) of equation 3.9 is also optimal for equation 3.1. Since we define \( I_k \) by the set of samples whose losses were small in equation 3.9, there is no guarantee that \( I_k = I \setminus J^* \) as well Larsen et al.’s (2012) algorithm. In general, it may not be easy to devise not only global optimization algorithms but also local optimization algorithms for nonconvex problems that have nonconvexity in both objective and constraint functions such as ER-SVM. However, algorithm 1 works well to find influential outliers for \( J \), as numerical experiments in Figure 7 show later.

To solve CVaR minimization equation 3.9 in the algorithm, we can use existing tools and software (Chang & Lin, 2001; Perez-Cruz et al., 2003; Takeda & Sugiyama, 2008) for solving \( \nu \)-SVM or \( E \nu \)-SVM. Therefore, the algorithm is easily implemented because the main part is solving equation 3.9 in order to solve ER-SVM. When \( \nu_k \in (0, \nu^k) \), we need to deal with equation 3.9 with the nonconvex set \( W \), to which we can use a local solution algorithm as in Perez-Cruz et al. (2003) and Takeda and Sugiyama (2008). Algorithm 1 becomes more efficient by using a warm-start strategy for equation 3.9 (i.e., by using \((w_{k-1}, b_{k-1}, \alpha_{k-1})\) as an initial solution). The problem, equation 3.9, does not change much in any iteration when setting a small value to \( \gamma \). Then small computation costs are needed to find the next solution, \((w_k, b_k, \alpha_k)\), by using a warm-start strategy.

4 Geometric Interpretation for ER-SVM

We give geometric interpretations for ER-SVM, equation 3.1, by showing the dual formulation of it. The dual ER-SVM is formulated as the minimum distance problem to a set that consists of samples \( x_i, i \in I \setminus J \). From theorem 1, we can give the same interpretation for robust SVM equation 2.4. The interpretation for robust SVM is not obvious from the formulation of equation 2.4 which uses truncated hinge losses. Furthermore, we also give a geometric interpretation for VaR minimization. Those interpretations help us to understand ER-SVM from the geometric viewpoint and recognize the difference of ER-SVM from VaR or CVaR-based methods.

4.1 Dual of ER-SVM. We define the Minkowski difference of two sets \( \mathcal{U}_+ \) and \( \mathcal{U}_- \) by

\[
\mathcal{U}_+ \ominus \mathcal{U}_- := \{x_+ - x_- | x_+ \in \mathcal{U}_+, x_- \in \mathcal{U}_-\}.
\]

When \( \mathcal{U}_+ \) and \( \mathcal{U}_- \) are compact convex sets and have interior points, their Minkowski difference \( \mathcal{U} \) is necessarily compact convex and has a nonempty interior. When \( \mathcal{U}_+ \) and \( \mathcal{U}_- \) do not intersect, \( \mathcal{U}_+ \cap \mathcal{U}_- = \emptyset \), we have \( 0 \notin \mathcal{U}_+ \ominus \mathcal{U}_- \). If they have an intersection, \( \mathcal{U}_+ \ominus \mathcal{U}_- \) contains the origin.
By taking the dual for ER-SVM, equation 3.1 in terms of $b$ and $\alpha$, it is transformed to

$$
\begin{align*}
\max_{w \in W, \|w\| \leq \nu |I|} & \min_{x_+ \in \mathcal{U}^{+}_{x}(f), x_- \in \mathcal{U}^{-}_{x}(f)} (x_+ - x_-)^T w, \quad \text{or} \quad (4.1) \\
\max_{w \in W, \|w\| \leq \nu |I|} & \min_{x \in \mathcal{U}^{-}_{x}(f)} x^T w,
\end{align*}
$$

where

$$
\mathcal{U}^{+}_{x}(f) := \mathcal{U}^{+}_{x}(f) \oplus \mathcal{U}^{-}_{x}(f),
$$

$$
\mathcal{U}^{+}_{x}(f) := \left\{ \sum_{i \in I_+ \setminus J} \lambda_i x_i : \sum_{i \in I_+ \setminus J} \lambda_i = 1, 0 \leq \lambda_i \leq \frac{2}{\nu |I \setminus J|}, i \in I_+ \setminus J \right\}. \quad (4.2)
$$

Let $w^*$ be its optimal solution. The bias $b^*$ of equation 3.1 can be computed by using margin support vectors, which are strictly on the margin ($x_i$ with $0 < \lambda_i^* < \frac{2}{\nu |I \setminus J|}$ in $\mathcal{U}^{+}_{x}(f^*)$). We can confirm the equivalence of ER-SVM, equation 3.1, to equation 4.1 by taking the dual of the inner-minimization problem in equation 4.1.

For equation 4.1, assume that $x_+$ and $x_-$ are representative points (or means) of each class. We can interpret the problem in the viewpoint of robust optimization (Ben-Tal, El-Ghaoui, & Nemirovski, 2009) by regarding $x_+$ and $x_-$ as uncertain inputs and preparing uncertainty sets $\mathcal{U}^{+}_{x}(f)$ where $x_+$ lie. Note that $\mathcal{U}^{+}_{x}(f)$ are convex sets constructed from training samples $x_i, i \in I_+ \setminus J$, respectively. ER-SVM, equation 4.1, finds a solution $w$ that is robust with respect to changes in the realization of $x_+$. 

Bishop (2006) (see section 4.1.4 for Fisher’s linear discriminant) discusses the problem maximizing $(x_+ - x_-)^T w$ in terms of $w \in W$ for sample means $x_+, x_-$ of each classes. Although the problem selects a projection that maximizes the class separation, the resulting solution, $w^*$, can induce considerable class overlap in the projected space, between $\{x_i^T w^* | i \in I_+\}$ and $\{x_j^T w^* | j \in I_-\}$. The difficulty arises when the samples of each class are generated from class distributions with strongly nondiagonal covariances. Fisher’s linear discriminant finds a solution $w^*$ so as to give a small variance within each class in addition to maximizing the class separation, thereby minimizing the class overlap. We can say that ER-SVM, equation 4.1 tries to overcome the difficulty in another way by removing outliers with the use of $f$ and considering the worst case in uncertainty sets $\mathcal{U}^{+}_{x}(f)$ for $x_+$. 

4.2 Transformation of Dual ER-SVM Depending on $\mathcal{U}^{+}_{x}(f)$. In section 3.1, we referred to convex ER-SVM, equation 3.8, where the nonconvex constraint $w \in W$ is replaced by the convex one, $w \in \text{conv}(W)$, without
changing the optimality of ER-SVM as far as \( \nu \in (\nu_{\phi}, \nu_{\max}) \). In the range, the optimal value of ER-SVM is negative (see Figure 4). We can relate the sign of the optimal value of ER-SVM to the position of 0 relative to the interior of \( U_{\nu \cap} (J^*) \) for an optimal solution \( J^* \) of ER-SVM. Indeed, \( 0 \notin U_{\nu \cap} (J^*) \) holds if and only if the optimal value of ER-SVM is negative, and the interior of \( U_{\nu \cap} (J^*) \) includes 0 \((0 \in \text{int}(U_{\nu \cap} (J^*))) \) if and only if the optimal value of ER-SVM is positive. We can prove them by mimicking the proof of lemma 1 in Takeda et al. (2013).

Therefore, if \( 0 \notin U_{\nu \cap} (J^*) \), we find that ER-SVM is equivalent to convex ER-SVM, equation 3.8. Otherwise, \( w \in W \) is essentially equivalent to the nonconvex inequality \( \|w\| \geq 1 \). We cannot detect whether \( w \in W \) can be made convex or not a priori by using \( U_{\nu \cap} (J^*) \), because the optimal solution \( J^* \) is necessary. However, we can give geometric interpretations for ER-SVM in each case. Now we transform ER-SVM, equation 4.1, into two norm-minimization problems depending on \( U_{\nu \cap} (J^*) \).

**Theorem 3.** Suppose that \( J^* \) is the optimal solution of ER-SVM, equation 4.1. When \( 0 \notin U_{\nu \cap} (J^*) \), ER-SVM is equivalent to

\[
\max_{\|J\| \leq \nu_{\phi}} \min_{\|x\| \leq \nu_{\nu \cap} (J^*)} \|x\|, \text{ or equivalently,}
\]

\[
\max_{\|J\| \leq \nu_{\phi}} \min_{\|x\| \leq \nu_{\nu \cap} (J^*)} \|x_+ - x_-\|. \tag{4.3}
\]

When \( 0 \in \text{int}(U_{\nu \cap} (J^*)) \), it is equivalent to

\[
\min_{\|J\| \leq \nu_{\phi}} \|x\|, \tag{4.4}
\]

where \( \mathbb{R}^n \setminus U_{\nu \cap} (J) \) is the closure of the complement of the convex set \( U_{\nu \cap} (J) \).

**Proof.** For a bounded convex set \( U \), we define the support functional by

\[
g(w; U) := \max_{x \in U} \{x^T w\}.
\]

ER-SVM, equation 4.1 is rewritten as

\[
\min_{\|J\| \leq \nu_{\phi}} \min_{w \in W} g(-w; U_{\nu \cap} (J^*)). \tag{4.5}
\]

Theorem 3.1 of Briec (1997) implies

\[
\min_{x \in U_{\nu \cap} (J)} \|x\| = - \min_{\|w\| \leq 1} g(-w; U_{\nu \cap} (J)) \tag{4.6}
\]
Figure 5: The hyperplane given by ER-SVM, equation 4.1, and resulting $U^\nu_I(J^\ast)$ and $U^\nu_I(J^\ast)$ (shown in solid lines) for the solution $J^\ast$. The circle marks (or triangle marks) plot samples with label 1 (or label $-1$, resp). The removed samples, samples in $J^\ast$, which belong to class $-1$ are shown with the stars.

for any $J(\subseteq I)$ satisfying $0 \notin U^\nu_I(J)$ in addition to the relation $w^\ast = x^\ast/\|x^\ast\|$ between optimal solutions: $w^\ast$ of equation 4.1 and $x^\ast$ of equation 4.3. This leads to the equivalence of equation 4.5 to 4.3 because $w \in W$ in equation 4.5 can be convex by $w \in \text{conv}(W)$ due to the negative optimal value of ER-SVM (i.e., $0 \notin U^\nu_I(J^\ast)$).

On the other hand, proposition 3.1 of Briec (1997) shows

$$\min_{x \in \mathbb{R}^n \setminus U^\nu_I(J)} \|x\| = \min_{\|w\| \geq 1} g(-w; U^\nu_I(J))$$

for any $J(\subseteq I)$ satisfying $0 \in \text{int}(U^\nu_I(J))$ in addition to the relation $w^\ast = -x^\ast/\|x^\ast\|$ between optimal solutions: $w^\ast$ of equation 4.1 and $x^\ast$ of equation 4.4. This leads to the equivalence of equation 4.5 to 4.4, because $w \in W$ in equation 4.5 can be replaced by $\|w\| \geq 1$ due to the positive optimal value of ER-SVM.

Figure 5 depicts the hyperplane given by ER-SVM, equation 4.1, as well as the resulting $U^\nu_I(J^\ast)$ and $U^\nu_I(J^\ast)$ with the use of $J^\ast$ of equation 4.1.
ER-SVM could detect all outliers shown with the star marks as \( J^* \). In this case, \( U^+= (J^*) \cap U^-(J^*) = \emptyset \) (or equivalently, \( \emptyset \in U^-(J^*) \)) holds, and therefore, ER-SVM, equation 4.1, equals the dual problem, equation 4.3, of convex ER-SVM. As theorem 3 implies, the hyperplane of equation 4.1 was obtained by maximizing the minimum distance between \( U^+(J) \) and \( U^-(J) \) with respect to \( J \).

By fixing \( J \), equation 4.3 reduces to a convex problem, whereas equation 4.4 remains a nonconvex problem because of the constraint \( x \in IR^n \setminus U^-(J^*) \). The former problem corresponds to convex ER-SVM whose parameter \( \nu \) is in the convex range, \( \nu \in (\nu_{J^*\phi}, \nu_{J^*\max}) \). Indeed, the range was defined so that the optimal \( \min_{\nu \in (\nu_{J^*\phi}, \nu_{J^*\max})} \nu \) is negative, but it is equivalently defined so that \( \emptyset \in U^-(J^*) \) in the range. In other words, as long as \( \nu \in (\nu_{J^*\phi}, \nu_{J^*\max}) \), \( \emptyset \in U^-(J^*) \) holds and ER-SVM, equation 4.1, is equivalent to convex ER-SVM, robust SVM equation 2.4, and robust \( \nu \)-SVM (see equation 3.4). As \( \nu \) becomes smaller, the set \( U^-(J^*) \) becomes larger and the nonconvex case, \( \emptyset \in U^-(J^*) \), tends to happen.

### 4.3 Relation to CVaR and VaR Minimization Models

Takeda, Mitsugi, and Kanamori (2012) proposed a unified formulation, unified classification model (UCM), for binary classification.\footnote{UCM was called unified robust classification model and abbreviated by RCM in Takeda et al. (2012). The term robust is from robust optimization (Ben-Tal et al., 2009). However, since the term is misleading in the context of this letter, we denote RCM by UCM.} UCM is formulated as

\[
\min_{\nu \in \mathcal{W}} \max_{x_+ \in U^+, x_- \in U_-} -(x_+ - x_-) \top w \quad \text{or} \quad \min_{\nu \in \mathcal{W}} \max_{x \in U^+ \ominus U^-} -x \top w. \tag{4.7}
\]

Depending on the definition of \( U_\pm \), UCM embraces various classification methods such as the support vector machine (SVM) (Schölkopf & Smola, 2002), minimax probability machine (Lanckriet, Ghaoui, Bhattacharyya, & Jordan, 2002), and Fisher discriminant analysis (Fukunaga, 1990).

Indeed, Takeda et al. (2012) showed that UCM is equal to CVaR minimization (i.e., Ev-SVM or \( \nu \)-SVM) when \( U_\pm \) are reduced convex hulls with size \( \nu \), defined in Bennett and Bredensteiner (2000) and Crisp and Burges (2000) as

\[
U^\nu_\pm = \left\{ \sum_{i \in I_\pm} \lambda_i x_i : \sum_{i \in I_\pm} \lambda_i = 1, 0 \leq \lambda_i \leq \frac{2}{\nu |I|}, i \in I_\pm \right\}. \tag{4.8}
\]

The sets are polytopes that have been shrunk toward the centers as \( \nu \) becomes large. By taking the dual for the inner maximization in UCM with the
above $\mathcal{U}_\nu^+$ with respect to $x$, we have problem 2.7 minimizing $(1 - \nu)$-CVaR, $\phi_{1-\nu}(w, b)$, with respect to $(w \in W, b)$, which is equivalent to $\nu$-SVM or $\nu$-SVM.

Figure 6 shows the reduced convex hulls, $\mathcal{U}_\nu^+$ and $\mathcal{U}_\nu^-$, of equation 4.8. Because $\mathcal{U}_\nu^+ \cap \mathcal{U}_\nu^- \neq \emptyset$ in this case, UCM, equation 4.7, with these $\mathcal{U}_\nu^\pm$ equals to $\nu$-SVM. The hyperplane was obtained by solving $\nu$-SVM, equation 2.2. The CVaR-based model takes into account all losses induced from training samples; therefore, the reduced convex hulls shown in the figure are influenced by outliers. Therefore, the hyperplane is also influenced by outliers, compared with the hyperplane of $\hat{\nu}$-SVM (see Figure 5).

We can rewrite VaR minimization, equation 2.5, as

$$\min_{w \in W, ||J|| \leq \nu ||I||} \max_{x \in \mathcal{U}(J)} -x^Tw,$$

where

$$\mathcal{U}(J) := \mathcal{U}_+(J) \ominus \mathcal{U}_-(J),$$

$$\mathcal{U}_\pm(J) := \left\{ \sum_{i \in I_\pm} \lambda_i x_i : \sum_{i \in I_\pm} \lambda_i = 1, 0 \leq \lambda_i, i \in I_\pm \right\}. \quad (4.9)$$
Note that $U_{\pm}(J)$ of equation 4.9 are generated by deleting an upper bound $\frac{2}{v|I\setminus J|}$ of $\lambda_j$ from $U_{\pm}(J)$ of ER-SVM, equation 4.1.

We can again confirm that ER-SVM is a mixture model of CVaR and VaR minimization from the definition (see equation 4.2) of $U_{\pm}(J)$ of ER-SVM; it has an upper bound $\frac{2}{v|I\setminus J|}$ as in equation 4.8 of CVaR minimization and delete some portion of samples as $J$ from the set as in equation 4.9 of VaR minimization.

5 Numerical Results

5.1 Properties of ER-SVM. The intermediate model, ER-SVM, has strong points of CVaR minimization and VaR minimization; a strong point of CVaR minimization is that the parameter choice of $v$ is not so sensitive for prediction, and a strong point of VaR minimization is that the optimal solution of VaR minimization is robust to outliers. We compared the performances of ER-SVM, equation 3.1, by algorithm 1, VaR minimization equation 2.5, by running the algorithm with $\epsilon = 0$ and CVaR minimization, equation 2.7. Recall that equation 2.7 equals Ev-SVM and especially for $v \in (\nu_{\phi}, \nu_{\max}]$, it reduces to $\nu$-SVM (Schölkopf et al., 2000). VaR minimization, equation 2.5, reduces to VaR-SVM (Tsyurmasto et al., 2014) for large $\nu$ (precisely, as long as $v > \bar{\nu}_\phi$; see Figure 4). ER-SVM equals robust SVM, equation 2.4, when $\nu \|I\| > \nu \|J\|$ (see lemma 1 and theorem 1).

We used synthetic data generated by following Xu et al. (2006). We generated two-dimensional samples with label $+1$ and $-1$ from two normal distributions with different mean vectors and the same covariance matrix. The optimal hyperplane for the noiseless data set is $h(x) = x_1 - x_2 = 0$ with $w = \frac{1}{\sqrt{2}}(1, -1)$ and $b = 0$. We added outliers only to the training set with label $-1$ by drawing samples uniformly from a half-ring with center 0, inner radius $R = 75$, and outer radius $R + 1$ in the space $x$ of $h(x) > 0$. The training set contained 50 samples from each class (100 in total), including outliers. The ratio of outliers in the training set was set to one of the values from 0 to 5% (10% only for Figure 8). The test set has 500 samples from each class (1000 in total). We repeated all the experiments 100 times, drawing training and test sets every repetition.

The parameter setting of $\gamma$ and $\epsilon$ in algorithm 1 for approximately solving ER-SVM is as follows: $\epsilon = 10^{-4}$ and $\gamma = 0.03/\nu$ (basic parameter setting throughout numerical experiments). This $\gamma$ makes the number of removed samples, $\gamma \|I\|$, unrelated to the choice of $\nu$.

Figure 7 shows the influence of outliers on performances of three methods for $\nu = 0.1, 0.2, \ldots, 0.9$. These figures show the average test errors with their estimation errors (the standard deviations divided by the square root of the number of trials, $\sqrt{100}$) over the 100 runs. $\nu_{\phi}$ indicates the maximum convexity threshold for CVaR minimization (Ev-SVM) among 100 trials. This indicates that ER-SVM with $\nu > \nu_{\phi}$ is equivalent to robust-SVM.
Figures 7a to 7d imply that ER-SVM (especially ER-SVM with $\nu \leq \nu_\phi$) achieved better performance as the ratio of outlier increases, while E$\nu$SVM’s performance became worse. We can confirm that the curve of the test error of ER-SVM was not so volatile with respect to $\nu$ compared to VaR minimization, and it achieved the lowest test errors among the three models in the presence of outliers.

We tested the precision and recall on the synthetic data including 10% outliers (i.e., 10 outliers are included in 100 samples) generated from the ring with inner radius $R = 20$. Precision is the number of outliers that algorithm 1 removed divided by the total number of removed samples, and recall is the number of outliers removed divided by the total number of existing outliers in the data set. Figure 8 (top) shows the precision recall curve with different values of $\epsilon$ that is used in the stopping criterion of algorithm 1. It varied from $\epsilon = 10^{-2}$ to $\epsilon = 10^{-8}$. The upper panel demonstrates that larger $\epsilon$ produces a high precision rate, implying that the outliers are detected at the early stage (i.e., small $k$) of algorithm 1. Figure 8 (bottom) shows the test errors of the classifiers obtained by algorithm 1 with corresponding $\epsilon$. The basic parameter setting $\epsilon = 10^{-4}$ used in the numerical experiments.
Figure 8: (Top) The precision-recall curve with different values of $\epsilon$ used in the stopping criterion of algorithm 1. (Bottom) The test errors with respect to different values of $\epsilon$. 

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/NECO_a_00647 by guest on 16 May 2021
Table 1: Synthetic Data ($\gamma = 0.03/\nu$, $\epsilon = 10^{-4}$, $R = 75$).

<table>
<thead>
<tr>
<th>Outlier Ratio</th>
<th>ER-SVM</th>
<th>robust SVM</th>
<th>Ev-SVM</th>
<th>C-SVM</th>
<th>VaR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1.86 ± 0.44</td>
<td>2.11 ± 0.61</td>
<td>1.80 ± 0.40</td>
<td>1.88 ± 0.51</td>
<td>2.07 ± 0.61</td>
</tr>
<tr>
<td>1%</td>
<td>1.91 ± 0.50</td>
<td>2.14 ± 0.86</td>
<td>1.99 ± 0.55</td>
<td>2.06 ± 0.64</td>
<td>2.22 ± 0.82</td>
</tr>
<tr>
<td>2%</td>
<td>1.96 ± 0.63</td>
<td>2.25 ± 0.90</td>
<td>2.38 ± 0.98</td>
<td>2.41 ± 1.01</td>
<td>2.38 ± 0.94</td>
</tr>
<tr>
<td>3%</td>
<td>2.10 ± 0.95</td>
<td>2.20 ± 1.01</td>
<td>2.83 ± 1.56</td>
<td>2.85 ± 1.59</td>
<td>2.53 ± 1.10</td>
</tr>
<tr>
<td>4%</td>
<td>2.14 ± 0.85</td>
<td>2.25 ± 1.17</td>
<td>4.00 ± 3.34</td>
<td>4.18 ± 3.44</td>
<td>2.56 ± 0.96</td>
</tr>
<tr>
<td>5%</td>
<td>2.42 ± 1.63</td>
<td>2.46 ± 1.82</td>
<td>5.64 ± 6.69</td>
<td>6.36 ± 6.94</td>
<td>2.57 ± 0.99</td>
</tr>
</tbody>
</table>

Notes: Average test error [%] ± standard deviations in 100 trials for synthetic data. The minimum average test error of five models for each outlier ratio is shown in bold.

stresses the precision rather than the recall, which leads to high-prediction performances (i.e., small test errors).

5.2 Comparison to Existing Models. We compared the performance of ER-SVM to the following existing models: C-SVM; Ev-SVM; robust SVM (which was solved by the concave-convex procedure, CCCP (Collobert, Sinz, Weston, & Bottou, 2006)) and VaR minimization. C-SVM is a well-known classification method. Ev-SVM, equation 2.2 is an extension of $\nu$-SVM (equivalent to C-SVM) and VaR minimization, equation 2.5 is an extension of VaR-SVM (Tsyurmasto et al., 2014). We can say that ER-SVM is a robust variant of Ev-SVM whereas robust SVM is a robust variant of C-SVM. We used $s = 1$ for robust SVM, equation 2.4.

5.2.1 Synthetic Data. Table 1 shows the results (average error [%] ± standard deviations of test errors in 100 trials) of comparing four SVMs on the synthetic data set of the previous section. We found the best parameter setting from 9 candidates, $\nu = 0.1, 0.2, \ldots, 0.9$, for ER-SVM, Ev-SVM, and VaR minimization and from $C = 10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}$ for robust SVM, and C-SVM. Figure 9 depicts the results of Table 1: the average test error of each learning model by changing the ratio of outliers in data sets. ER-SVM, VaR minimization, and robust SVM were less influenced by outliers than standard SVM models such as C-SVM and Ev-SVM. Above all, the proposed model, ER-SVM, achieved a good prediction performance.

5.2.2 UCI Data Sets. We generated contaminated data sets from original data sets of UCI repository (Blake & Merz, 1998), that are shown in Table 2, by adding outliers as follows. We scaled all attributes of the original data set from $-1.0$ to $1.0$, generated outliers $\hat{x}$ uniformly from a ring with center 0 and radius $R$, and assigned the wrong label $\hat{y}$ to $\hat{x}$ by using optimal classifiers of Ev-SVM. The radius $R$ for generating outliers was properly set so that outliers have an impact on the test errors. In addition, the percentage of
outliers (outlier ratio) was increased until outliers had a large influence on the test errors.

We generated 10 contaminated data sets for each outlier ratio. Figure 10 shows the average of test errors on 10 contaminated data sets with the best parameter choice. The best parameter was chosen among 9 candidates at equal intervals from \( (0, \nu_{\text{max}}] \) for ER-SVM, \( E\nu\)-SVM, and VaR minimization and from 9 candidates (the same number of candidates for \( \nu \)) as \( C = 10^{-4}, 10^{-3}, \ldots, 10^3, 10^4 \) for robust SVM and C-SVM. The plots on each line of ER-SVM, \( E\nu\)-SVM, and VaR minimization indicate that the best
Figure 10: Comparison of prediction performances for contaminated UCI data sets. \(\left(\frac{\text{Outlier ratio}}{100} \times |I|\right)\) outliers were uniformly generated from a ring with center 0 and radius \(R\). The plots on each line of ER-SVM, Ev-SVM, and VaR minimization indicate that best parameter value \(\nu\) was attained in the nonconvex range.
parameter value $\nu$ was attained in the nonconvex range $\nu \leq \nu_\phi$; no plots indicate that it was attained in the convex range $\nu \in (\nu_\phi, \nu_{\text{max}}]$.

When the outlier ratio increases, ER-SVM tends to have good prediction performance over robust SVM in most cases. Especially, for the liver-disorders and diabetes data sets, nonconvex ER-SVM achieved good prediction performances compared to other methods. As for the other data sets, the best parameter $\nu$ of ER-SVM was chosen in the convex range for almost all outlier ratios; therefore, the differences in performance between convex ER-SVM and robust SVM were in the parameter setting $\nu$ and $C$. $\nu$ may be more easily adjusted than $C$. ER-SVM has a possibility of achieving better prediction performances with proper parameter setting.

6 Conclusion

We proposed extended robust SVM (ER-SVM), which minimizes an intermediate risk measure between the CVaR and VaR by expecting that the resulting model becomes less sensitive to outliers than $E\nu$-SVM. Our model, ER-SVM, is an extension of robust SVM (Xu et al., 2006; Wu & Liu, 2007). Indeed, if we set $\nu_T$ of ER-SVM to a value in the convex range, ER-SVM and robust SVM give the same classifier. Numerical experiments show the superior performance of our model over robust SVM, $\nu$-SVM, and $E\nu$-SVM in the presence of outliers. The effectiveness of the extended parameter range of ER-SVM contributes to the superior performance over robust SVM, whereas ignoring samples with large losses in ER-SVM contributes to the superior performance over $E\nu$-SVM.

ER-SVM includes two parameters, $\nu_O$ and $\nu_I$. We expect that if $\nu_O$ is set to the ratio of outliers, ER-SVM will achieve a good prediction performance, but it is hard to predict the ratio in practical problem setting. Therefore, in this letter, we proposed a heuristic algorithm for ER-SVM that automatically tunes these parameter values during execution.

Our algorithm has tuning parameter $\nu$, $\gamma$, and $\epsilon$. The prediction performance is not significantly affected by $\gamma$ and $\epsilon$. However, some guidelines for the choice of parameters would be helpful in practice. If we do not care about computation time, it is better to set a small value for $\gamma$. As for $\epsilon$, $10^{-2}$ to $10^{-5}$ may be appropriate for scaled data sets (see Figure 8). A challenging issue is to develop a reliable cross-validation method for $\nu$ in the existence of outliers.

In the future, we want to investigate a practical way to set parameters directly for $\nu_O$ and $\nu_I$ of ER-SVM. The parameter $\nu_O$ controls the number of ignored samples, and $\nu_I$ controls the thickness of the margin. The superior performance of ER-SVM over robust SVM is due to the extended permissible range for $\nu_I$. With small $\nu_I$ in the extended range, the margin of ER-SVM can be negative, different from robust SVM. We also need to check the problem settings and features of data sets where the negative margin works.
References


Received September 15, 2013; accepted April 30, 2014.