**RESEARCH REPORT # 2013-4**

Risk Management and Financial Engineering Lab
Department of Industrial and Systems Engineering
University of Florida, Gainesville, FL 32611

# Support Vector Classification with Positive Homogeneous Risk Functionals

**Peter Tsyurmasto** TSYURMASTO@UFL.EDU
*USA*

**Stan Uryasev** URYASEV@UFL.EDU
*USA*

**Jun-ya Gotoh** JGOTO@INDSYS.CHUO-U.AC.JP
*JAPAN*

**Editor:** Leslie Pack Kaelbling

## Abstract

Support Vector Machine (SVM) is often formulated as a structural risk minimization, which minimizes simultaneously an empirical risk and a regularization. At the same time, SVM can also be interpreted as a maximization of geometric margin. This paper revisits these two views on SVM by applying a risk management approach. A generalization of the maximum margin formulation is given, and shown to contain several classical versions of SVMs such as hard margin SVM, $\nu$-SVM, and extended $\nu$-SVM as its special cases, corresponding to the particular choices of risk functional. Under the assumption that the empirical risk function is positive homogeneous, we derived conditions under which the generalized formulation is equivalent to several structural risk minimization formulations. Sufficient conditions for the existence of an optimal solution and unbounded solution to the associated optimization problems are also given. Within the presented framework, we propose a new classification method based on the difference of two Conditional Value-at-Risk (CVaR) measures, which is robust to outliers. Computational experiments confirm that this formulation has a superior out-of-sample performance on datasets contaminated by outliers, compared to $\nu$-SVM.

## 1. Introduction

**Motivation.** The success of support vector machine (SVM) is based on a wide range of concepts from statistics, functional analysis, computer science, mathematical optimization, etc. This paper is focused on analysis of optimization formulations of SVM.

As presented in textbooks or tutorial papers (e.g., Burges (2000)), the most fundamental SVM can be traced back to the so-called maximum margin criterion for binary classification. It maximizes the so-called *geometric margin* when the given sample data set is (linearly) separable. Suppose we have a training dataset $(\xi_1, y_1), \ldots, (\xi_l, y_l)$ of features $\xi_i \in \mathbb{R}^m$ with binary class labels $y_i \in \{-1, +1\}$ for $i = 1, \ldots, l$, and a hyperplane specified by the equation $w^T x + b = 0$, $w \in \mathbb{R}^m \setminus \{0\}$, $b \in \mathbb{R}$. The following quantities gauge the degrees of misclassification on the basis of the geometric distances from data samples to the hyperplane:

$$d_i(w, b) = -\frac{y_i(w^T \xi_i + b)}{\|w\|}, \text{ for } i = 1, \ldots, l, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, i.e., $\|\xi\| := \sqrt{\xi_1^2 + \cdots + \xi_m^2}$. Inequality $d_i(w, b) \leq 0$ implies that the sample $i$ is correctly classified by the hyperplane, while $d_i(w, b) > 0$ implies wrong classification.

With the notation above, the maximum margin criterion is formulated as a minimization problem

$$\underset{w,b}{\text{minimize}} \quad \max\{d_1(w, b), ..., d_l(w, b)\}. \tag{2}$$

This criterion is appealing since 1) it is intuitively understandable; 2) it is a straightforward minimization of the so-called *generalization error bound* (e.g., Vapnik, 1999), which looks persuasive for people with deep statistical background. In addition, as long as the data set is separable, the margin maximization can be reduced to a convex quadratic programming problem (QP):

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\|w\|^2 \quad \text{subject to } - y_i(w^T \xi_i + b) \leq -1, \ i = 1, ..., l.$$

On the contrary, when the data set is not separable, the resulting convex QP becomes infeasible, and the soft margin SVM Cortes and Vapnik (1995) is developed by introducing slack variables. The soft margin SVM can then be considered as the simultaneous minimization of an empirical risk (associated with the slack variables introduced) and a regularization term (usually defined with the Euclidean norm of the normal vector of the hyperplane).

For example, with the aforementioned notation, the $C$-SVM, the most popular soft margin formulation, is represented by the bi-objective minimization:

$$\underset{w,b}{\text{minimize}} \quad \tfrac{1}{2}\|w\|^2 \quad + \quad C \cdot \tfrac{1}{l} \textstyle\sum_{i=1}^{l} \max\{1 - y_i(w^T \xi_i + b), 0\}, \tag{3}$$

with some constant $C > 0$. Note that the constant $C$ is a parameter balancing the two objectives and it needs to be determined. Based on this formulation, the $C$-SVM can be viewed as the structural risk minimization of the following form:

$$\text{minimize} \quad (\text{regularization}) \ + \ C \cdot (\text{empirical risk}). \tag{4}$$

However, readers may question the logic sketched above: "Why margin maximization (2) is not applied for the inseparable case?" or " What is the relation between the margin maximization (2) and $C$-SVM (3)?"

It is not hard to see that the margin maximization (2) results in a nonconvex optimization unless the data set is separable. But do we have any statistical (not tractability-based) motivation for avoiding the nonconvexity?

In addition, the principle of the structural risk minimization does not mention how it treats the trade-off between the empirical risk and the regularization. In other words, it can also be formulated in either

$$\text{minimize (empirical risk)} \;\; \text{subject to (regularization)} \leq E \text{ with a constant } E, \tag{5}$$

or

$$\text{minimize (regularization)} \;\; \text{subject to (empirical risk)} \leq -D \text{ with a constant } D. \tag{6}$$

However, with the empirical risk function employed in $C$-SVM (3), each of the above formulations (4), (5) and (6) results in a different classifier. In fact, $C$-SVM of the form (5) with $E > 0$ results in a meaningless solution with $w = 0$, in which no hyperplane is obtained. On the other hand, it is proved that the $\nu$-SVM (Schölkopf et al., 2000), another popular soft margin SVM, does not depend on the difference among the formulations (4), (5) and (6).

**Approach: A Generalized Risk Minimization.** The main focus of this paper is to find what factor makes such a difference. To explore this, we apply a risk management approach to the classification problem. The idea of the approach is to view the original criterion (2) as a minimization problem with objective function $\mathcal{R}(d_1, \ldots, d_l)$ referred to as *risk functional* and defined on geometric margins $d_1, \ldots, d_l$:

$$\underset{w,b}{\text{minimize}} \;\; \mathcal{R}(d_1(w,b), \ldots, d_l(w,b)), \tag{7}$$

where $\mathcal{R} : \mathbb{R}^l \to \mathbb{R}$ and $d_i$ is defined in (1). In other words, optimization problem (2) is a special case of (7) when risk functional $\mathcal{R}(d_1, \ldots, d_l) = \max\{d_1, ..., d_l\}$ is applied.

The area of risk management has been extensively grown in recent years. Specifically, *value-at-risk (VaR)* and *conditional value-at-risk (CVaR)* (Rockafellar and Uryasev, 2000) are currently widely used to monitor and control the market risk of financial instruments (Jorion, 1997; Duffie and Pan, 1997). A comprehensive recent study of risk measures can be found in Rockafellar and Uryasev (2013).

A merit of the generalized representation of the form (7) is to provide a unified scheme for various existing SVM formulations. Specifically, we show that hard-margin SVM (Boser et al., 1992), $\nu$-SVM (Schölkopf et al., 2000), extended $\nu$-SVM (Pérez-Cruz et al., 2003) and VaR-SVM (Tsyurmasto et al., 2013) in primal can be represented in compact forms with risk functionals (max-functional, conditional value-at-risk, value-at-risk) and derived from formulation (7).

It is not surprising that the form (7) has a profound relation to the structural risk minimization forms (4), (5) and (6), but the relation among them has been explored only in a limited number of articles by specifying its own structure. For example, Pérez-Cruz et al. (2003); Gotoh and Takeda (2005) study the relation of the $\nu$-SVM (Schölkopf et al., 2000) to the optimization problem of the form (7) with the CVaR. Recently, Gotoh et al. (2013a) develop a class of generalized formulations of SVM on the basis of the so-called *coherent risk measures* (Artzner et al., 1999), which includes the max-functional and the CVaR as special cases. The novelty of this paper is in more general assumptions on risk functional. In particular, our analysis can be applied to nonconvex risk measures, which are not covered by the existing papers. Further we show that the considered approach can derive a more general set of classifiers.

Specifically, we consider a class of risk functionals that are positive homogeneous. We first show that with this assumption, the structural risk minimizations of the different forms (4), (5) and (6), have the same set of optimal classifiers if the optimal value of (7) is nonpositive; On the other hand, neither (4), (5) or (6) provides any classifier if the optimal value of (7) is positive (Figure 2). The paper derives a sufficient condition for the exsitence of optimal solutions of the considered formulations. It is noteworthy that the obtained condition can be checked before the optimization, and is consistent with the fact pointed out by Burges (2000); Chang and Lin (2002); Gotoh and Takeda (2005); Gotoh et al. (2013b) where $\nu$-SVM is shown to be applicable not for all values of parameter $\nu$ in the range $[0, 1]$.

In addition, if the risk measure is positive homogeneous, the set of optimal classifiers obtained by either of the structural risk minimizations (4), (5) or (6) does not depend on the parameters $C, D$ or $E$. Namely, without bothering how to set those values, we can set $C = D = E = 1$ (Figure 2). This fact is pointed out in Schölkopf et al. (2000) for the $\nu$-SVM by using the duality. In contrast, we show without duality that such a property comes from the positive homogeneity of the corresponding risk functional. On the other hand, the aforementioned dependence on the parameters in the $C$-SVM can be explained by lack of positive-homogeneity.

**A New Formulation.** As a practical application of the developed methodology, we propose a new classification method referred to as *CVaR-*$(\alpha_L, \alpha_U)$*-SVM* with lower and upper parameters, $\alpha_L$ and $\alpha_U$ such that $0 \leq \alpha_L < \alpha_U < 1$. CVaR-$(\alpha_L, \alpha_U)$-SVM is a special case of the form (7) so that functional $\mathcal{R}$ is roughly an average between $100\alpha_L$ and $100\alpha_U$ percentiles of the set $\{d_1(w, b), \ldots, d_l(w, b)\}$ for each pair $w, b$. In fact, $\nu$-SVM is a special case of CVaR-$(\alpha_L, \alpha_U)$-SVM with parameters $\alpha_L = 1 - \nu$ and $\alpha_U = 1$. The proposed classifier has an additional parameter $\alpha_U$, which specifies that $(1 - \alpha_U) \cdot 100\%$ data samples with highest distances to the hyperplane are disregarded. When dataset does not contain outliers, the parameter $\alpha_U$ can be chosen (almost) equal to 1 and, thus, CVaR-$(\alpha_L, \alpha_U)$-SVM performs as good as $\nu$-SVM. However, when dataset is contaminated by outliers, CVaR-$(\alpha_L, \alpha_U)$-SVM has an advantage of stability to outliers, compared to $\nu$-SVM. The formulation of CVaR-$(\alpha_L, \alpha_U)$-SVM can be related to an existing formulation known as the ramp loss SVM Collobert et al. (2006). An advantage of the proposed formulation is parallel to that of $\nu$-SVM over the $C$-SVM. Namely, the CVaR-$(\alpha_L, \alpha_U)$-SVM has two interpretable parameters $\alpha_L, \alpha_U$, while the ramp loss SVM has a difficulty in the interpretation of the included parameter, and some exhaustive search is required for the parameter tuning. The computational experiments confirm that in presence of outliers, CVaR-$(\alpha_L, \alpha_U)$-SVM has a superior out-of-sample performance versus $\nu$-SVM.

The contribution of this paper can be summarized as follows:

- Based on the geometric margin, which is employed in the maximum margin SVM, we provide a unified scheme which contains several well-known SVMs;
- With the notion of risk functionals, we establish relations between existing SVM classifiers, and reveal that positive homogeneity plays a central role in the equivalence among the different formulations;

- We establish a sufficient condition for existence of optimal solution of (7).;
- We develop a nonlinear extension of (7);
- As a special case of (7), we propose a new classifier stable to outliers and empirically confirm its robust properties.

The paper is structured as follows. Section 2 presents a risk management approach to classification problem by following a brief overview of the existing formulations and shows that SVMs can be expressed in compact form in terms of risk functionals. Section 3 proposes a classification model based on positive homogeneous risk functionals. Section 4 proposes CVaR-$(\alpha_L, \alpha_U)$-SVM. Section 5 extends to nonlinear CVaR-$(\alpha_L, \alpha_U)$-SVM. Section 6 contains results of computational experiments. Section 7 concludes.

## 2. Representation of SVMs with Risk Functionals in Primal

**A Brief Overview of Existing SVM Formulations for Binary Classification.** Let $(\xi_1, y_1), \ldots, (\xi_l, y_l)$ be a *training dataset* of features $\xi_i \in \mathbb{R}^m$ with binary class labels $y_i \in \{-1, +1\}$ for $i = 1, \ldots, l$. The original features $\xi_1, \ldots, \xi_l \in \mathbb{R}^m$ are transformed into features $\phi(\xi_1), \ldots, \phi(\xi_l) \in \mathbb{R}^n$, respectively, with the mapping $\phi : \mathbb{R}^m \to \mathbb{R}^n$. The goal of the SVM is then to construct a hyperplane specified by the equation $w^T x + b = 0$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$, which separates the transformed features $\phi(\xi_1), \ldots, \phi(\xi_l)$ with class labels $+1$ and $-1$ from each other in $\mathbb{R}^n$ space.

In order to find a hyperplane on the basis of the given training set and an (implicit) mapping $\phi$, minimizing violation of separation is a bottom line. We say that the dataset is *separable* if there exists $(w, b) \in \mathbb{R}^n \backslash \{0\} \times \mathbb{R}$ such that for each $i = 1, ..., l$, $y_i(w^T \phi(\xi_i) + b) > 0$. On the other hand, a data point $\xi_i$ is considered to be wrongly classified if $y_i(w^T \phi(\xi_i) + b) < 0$ for a given $(w, b)$. A reasonable criterion for determining a $(w, b)$ is to place the hyperplane so that the geometric distance from the worst classified data sample would be the nearest if dataset is inseparable, or that from the nearest data is maximized if data set is separable. This rule is formulated as a fractional programming (FP) problem:

$$\underset{w,b}{\text{maximize}} \quad \min_i \left\{ \frac{y_i(w^T \phi(\xi_i) + b)}{\|w\|} : i = 1, ..., l \right\}. \tag{8}$$

In this paper, the quantity $y_i(w^T \phi(\xi_i) + b)$ is referred to as the *margin* (of a sample $i$), while $y_i(w^T \phi(\xi_i) + b)/\|w\|$ is referred to as the *geometric margin* (of a sample $i$). The optimization problem (8) is known to be rewritten by a quadratic programming (QP) problem:

$$\left| \begin{array}{ll} \underset{w,b}{\text{maximize}} & \frac{1}{2}\|w\|^2 \\ \text{subject to} & y_i(w^T \phi(\xi_i) + b) \geq 1, i = 1, ..., l, \end{array} \right. \tag{9}$$

if the dataset is separable under the mapping $\phi$. This approach is known as *hard margin SVM* (see, e.g., Boser et al. (1992)). It is, however, noteworthy that the FP (8) remains well-defined even when dataset is inseparable, whereas QP (9) is then infeasible, i.e., it has no feasible solution.

In order to deal with non-separable data sets, formulation (9) was extended by introducing slack variables $z_1, ..., z_l$:

$$\left| \begin{array}{ll} \underset{w,b,z}{\text{minimize}} & \frac{1}{2}\|w\|^2 + \frac{C}{l}\sum_{i=1}^{l} z_i \\ \text{subject to} & y_i(w^T\phi(\xi_i) + b) \geq 1 - z_i, i = 1, ..., l, \\ & z_i \geq 0, i = 1, ..., l, \end{array} \right. \tag{10}$$

where $C > 0$ is a parameter to be tuned. (10) is referred to as *C-SVM*. It should be noted that (10) is rewritten by an unconstrained (convex) minimization of the form:

$$\underset{w,b}{\text{minimize}} \;\; \frac{1}{2}\|w\|^2 + \frac{C}{l}\sum_{i=1}^{l}[1 - y_i(w^T\phi(\xi_i) + b)]_+, \tag{11}$$

where $[x]_+ := \max\{x, 0\}$. The form (11) is referred to as a class of *structural risk minimization*, which is often considered as a central principle for machine learning methods. Namely, (11) is considered as the simultaneous minimization of the empirical classification error term $\sum_{i=1}^{l}[1 - y_i(w^T\phi(\xi_i) + b)]_+$ and the regularization term $\|w\|$, which is added to control overfitting.

On the other hand, the interpretation of the empirical error term has been left ambiguous in the sense that a data sample contributes to this term if the margin is less than 1 and no clear interpretation is given for the meaning of the value of 1. At the same time, the interpretation of the parameter $C$ is also not clear.

As a remedy for the ambiguous interpretation of the involved constants, i.e., 1 and $C$, Schölkopf et al. (2000) develop an alternative known as $\nu$-SVM:

$$\left| \begin{array}{ll} \underset{w,b,\rho,z}{\text{minimize}} & \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{l}\sum_{i=1}^{l} z_i \\ \text{subject to} & y_i(w^T\phi(\xi_i) + b) \geq \rho - z_i, i = 1, ..., l, \\ & z_i \geq 0, i = 1, ..., l, \end{array} \right. \tag{12}$$

where $\nu \in (0, 1]$ is a parameter to be tuned.

Corresponding to (12), Pérez-Cruz et al. (2003) pose an extended formulation:

$$\left| \begin{array}{ll} \underset{w,b,\rho,z}{\text{minimize}} & -\nu\rho + \frac{1}{l}\sum_{i=1}^{l} z_i \\ \text{subject to} & y_i(w^T\phi(\xi_i) + b) \geq \rho - z_i, i = 1, ..., l, \\ & z_i \geq 0, i = 1, ..., l, \\ & \|w\| = 1, \end{array} \right. \tag{13}$$

where $\nu \in (0, 1]$ is a parameter to be tuned. E$\nu$-SVM (13) is a nonconvex minimization formulation, but it extends the lower bound $\nu_{\min}$ of admissible range of $\nu$-SVM to $(0, \nu_{\max}]$.

We have so far overviewed several popular SVM formulations, and it is not hard to see that all are based on the margin, $y_i(w^T\phi(\xi_i)+b)$, or geometric margin, $y_i(w^T\phi(\xi_i)+b)/\|w\|$, and eventually formulated as a structural risk minimization form (e.g., (9) (11), (12) and (13)) or an FP (e.g., (8)). It is not hard to see that all these have a common feature in formulating a certain trade-off between the empirical risk and the regularization, but it is not clear how they are related to each other, especially in terms of the parameters therein, such as $C$ in (10) or $\nu$ in (12) and (13). One of the purposes of this paper is to reveal the connection among those formulations and to show that the positive homogeneity of the related function plays an important role.

**SVM Formulations with Various Risk Functionals.** This paper considers the following probabilistic setting. Let $\Omega = \{\omega_1, \ldots, \omega_l\}$ be a finite sample space of *outcomes* with equal probabilities $\Pr(\omega_i) = \frac{1}{l}$ for $i = 1, \ldots, l$ and $\xi : \Omega \to \mathbb{R}^n$, $y : \Omega \to \{-1, +1\}$ be a pair of discretely distributed random variables defined as $\xi(\omega_i) = \phi(\xi_i)$, $y(\omega_i) = y_i$ for $i = 1, \ldots, l$. Random *loss function* is defined for each outcome $\omega \in \Omega$ and *decision variables* $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ by

$$\mathcal{L}_\omega(w, b) = -y(\omega) \cdot [w^T \xi(\omega) + b]. \tag{14}$$

Loss function (14) represents a random error on the training set. Namely, negative loss $\mathcal{L}_{\omega_i}(w, b) < 0$ indicates that data sample $i$ is classified correctly, while positive loss $\mathcal{L}_{\omega_i}(w, b) > 0$ indicates that $\phi(\xi_i)$ is classified incorrectly, $i = 1, \ldots, l$.

A random *distance function* is defined for each outcome $\omega \in \Omega$ and decision variables $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ as a normalized loss function:

$$d_\omega(w, b) = \frac{\mathcal{L}_\omega(w, b)}{\|w\|}. \tag{15}$$

For each outcome $\omega \in \Omega$, the absolute value of (15) is a Euclidian distance between a vector $\xi(\omega)$ and a hyperplane $H = \{x \in \mathbb{R}^n | w^T x + b = 0\}$. Distance function (15) has a discrete distribution of realizations

$$-\frac{y_i \cdot [w^T \phi(\xi_i) + b]}{\|w\|}\Bigg|_{i=1}^{l}. \tag{16}$$

with equal probabilities $\frac{1}{l}$. Figure 1 shows a histogram for the realizations (16) of distances of data samples $(\phi(\xi_1), y_1), \ldots, (\phi(\xi_l), y_l)$ to the hyperplane for some fixed parameters $(w, b)$ of the hyperplane. Samples falling into light area (blue in color) are classified correctly, while samples falling into a dark area (red in color) are classified incorrectly. The goal is to select such parameters of hyperplane $(w, b)$ in (15) as to reduce the size of the right tail (marked with red color) of the histogram.

This problem can be formalized by introducing a risk functional $\mathcal{R}(\cdot)$, which converts a random distance $d_\omega(w, b)$ into a deterministic function $\mathcal{R}(d_\omega(w, b))$:

$$\underset{w,b}{\text{minimize}} \;\; \mathcal{R}\left(\frac{\mathcal{L}_\omega(w, b)}{\|w\|}\right). \tag{17}$$

An objective in (17) represents an aggregated loss in the right tail of (15) so that the lower the objective in (17), the lower the size of the right tail of (15) exceeding zero. Optimization problem (17) can be viewed as a margin-based classifier since it takes into account distances of data samples to the hyperplane and aggregates them by means of risk functional $\mathcal{R}(\cdot)$. Further we assume that $\mathcal{R}(C) = C$ for each constant $C \in \mathbb{R}$. Each choice of risk functional $\mathcal{R}(\cdot)$ defines a particular classifier.

In the literature of machine learning, the risk functional is usually minimized together with some regularization.

$$\underset{w,b}{\text{minimize}} \;\; \frac{\mathcal{R}\left(\mathcal{L}_\omega(w, b)\right)}{\|w\|}, \tag{18}$$

$$\underset{w,b}{\text{minimize}} \;\; \mathcal{R}(\mathcal{L}_\omega(w, b)) \quad \text{subject to} \;\; \|w\| = E, \quad E > 0, \tag{19}$$

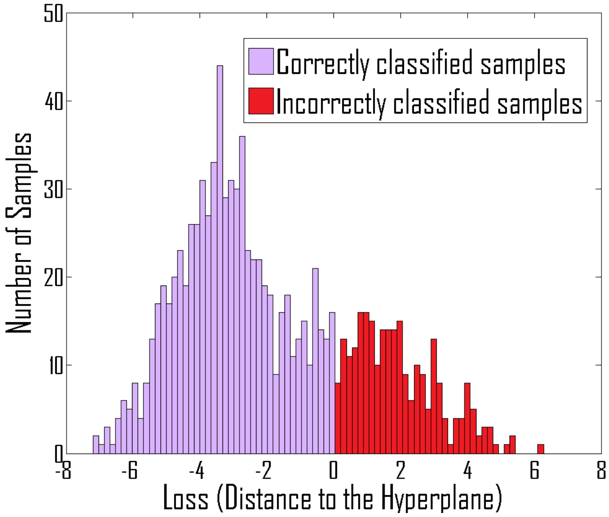

Figure 1: Histogram of the distances (1) of the data samples $(\phi(\xi_1), y_1), \ldots, (\phi(\xi_l), y_l)$ to the hyperplane for German Credit Data (The dataset was taken from UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/datasets.html`) when the decision variables are fixed.

$$\underset{w,b}{\text{minimize}} \;\; \mathcal{R}(\mathcal{L}_\omega(w,b)) \quad \text{subject to} \;\; \|w\| \leq E, \quad E > 0, \tag{20}$$

$$\underset{w,b}{\text{minimize}} \;\; \frac{1}{2}\|w\|^2 \quad \text{subject to} \;\; \mathcal{R}(\mathcal{L}_\omega(w,b)) \leq -D, \quad D > 0, \tag{21}$$

$$\underset{w,b}{\text{minimize}} \;\; \frac{1}{2}\|w\|^2 + C \cdot \mathcal{R}(\mathcal{L}_\omega(w,b)), \quad C > 0, \tag{22}$$

corresponding to particular risk functionals $\mathcal{R}(\cdot)$.

**Example: Maximum Margin SVM and Hard-Margin SVM.** The maximum margin formulation (8) and the hard margin SVM (9) can be related to the *worst-case loss* (or *maximum loss*):

$$\sup(\mathcal{L}_\omega) := \max\{\mathcal{L}_{\omega_1}, ..., \mathcal{L}_{\omega_l}\}. \tag{23}$$

With worst-case loss (23), formulations (8) and (9) are special cases of (17) and (21) (with $D = 1$), respectively. □

**Example: $C$-SVM.** Based on (11), the $C$-SVM (10) corresponds to the *above-target-loss*:

$$\text{ATL}_t(\mathcal{L}_\omega) := \frac{1}{l}\sum_{i=1}^{l}[\mathcal{L}_{\omega_i} + t]_+. \tag{24}$$

with $t = 1$. Indeed, substitution of (24) with (14) into (22) results in (11), or equivalently, (10). □

**Example: $\nu$-SVM (Schölkopf et al. (2000)).** Using the fact that constraint $\rho \geq 0$ is redundant in the formulation as was shown by Burges (2000), $\nu$-SVM (12) can be expressed with the Conditional Value-at-Risk (CVaR) functional, i.e., $\mathcal{R}(\cdot) = \mathrm{CVaR}_\alpha(\cdot)$, which is given by

$$\mathrm{CVaR}_\alpha\left(\mathcal{L}_\omega\right) := \min_c \left\{ c + \frac{1}{(1-\alpha)l} \sum_{i=1}^{l} \left[\mathcal{L}_{\omega_i} - c\right]_+ \right\}, \tag{25}$$

where $\alpha \in [0,1)$, $[x]_+ := \max\{x, 0\}$. By using the (strong) duality theorem of linear programming (see, e.g., Boyd and Vandenberghe (2004)), we can obtain a useful formula computing the CVaR of a loss $\mathcal{L}_\omega$. Let $\mathcal{L}_{[i]}$ denote the $i$-th largest element among $\mathcal{L}_{\omega_1}, ..., \mathcal{L}_{\omega_l}$, i.e., $\mathcal{L}_{[1]} \geq \mathcal{L}_{[2]} \geq \cdots \mathcal{L}_{[l]}$. Then the CVaR of $\mathcal{L}$ is calculated by the formula:

$$\mathrm{CVaR}_\alpha(\mathcal{L}_\omega) = \frac{1}{pl} \left\{ \sum_{i=1}^{\lfloor p \rfloor} \mathcal{L}_{[i]} + (p - \lfloor p \rfloor) \mathcal{L}_{[\lfloor p \rfloor + 1]} \right\}, \tag{26}$$

where $p = (1-\alpha)l$. Consider the case where $(1-\alpha)l$ is an integer $k \in \{1, 2, ..., l\}$, then the formula (26) can be simplified as follows:

$$\mathrm{CVaR}_\alpha(\mathcal{L}) = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}_{[i]}.$$

In this sense, the CVaR can be considered as the conditional expectation of $\mathcal{L}$ of the largest $(1-\alpha)l$ elements. (See, e.g., Rockafellar and Uryasev (2013) for the detailed explanation and properties for CVaR).

The relation between $\nu$-SVM and the CVaR minimization was first noticed in Gotoh and Takeda (2005). □

**Example: Extended $\nu$-SVM (E$\nu$-SVM) (Pérez-Cruz et al. (2003)).** E$\nu$-SVM (13) can be expressed with the CVaR functional (25)

$$\underset{w,b}{\text{minimize}} \quad \mathrm{CVaR}_{1-\nu}\left(\mathcal{L}_\omega(w,b)\right) \quad \text{subject to } \|w\| = 1, \tag{27}$$

as was shown by Takeda and Sugiyama (2008). Indeed, substitution of the CVaR with (14) and $\alpha = 1 - \nu$ into (19) with $E = 1$ results in (27). □

**Example: VaR-SVM (Tsyurmasto et al. (2013)).** (VaR) is a quantile of the loss, i.e.,

$$\mathrm{VaR}_\alpha\left(\mathcal{L}_\omega\right) := \min_c \left\{ c : \Pr\{\mathcal{L}_\omega \leq c\} \geq \alpha \right\}, \tag{28}$$

where $\alpha \in (0,1)$. VaR-SVM can be expressed in the form (21) with VaR functional (28):

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 \quad \text{subject to } \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b)) \leq -1. \tag{29}$$

In contrast to the aforementioned examples, $\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b))$ is nonconvex with respect to w and b, and accordingly, (29) is a nonconvex minimization, which may have (nonglobal) local minima. □

## 3. Relation between SVM Formulations with Risk Functionals

In this section, we establish relations between optimization problems (17)-(22), summarized in Figure 2. First of all, we find out which of the aforementioned problems have the same set of classifiers for each risk functional $\mathcal{R}(\cdot)$.

**Proposition 1** *Classifiers defined in primal by optimization problems* (17) *and* (19) *provide the same separating hyperplane for each risk functional* $\mathcal{R}(\cdot)$ *and the constant* $E > 0$.

Indeed, if $(w^*, b^*)$ is an optimal solution of (17), then $(\frac{Ew^*}{\|w^*\|}, \frac{Eb^*}{\|w^*\|})$ is an optimal solution of (19). Conversely, if $(w^{**}, b^{**})$ is an optimal solution of (19), then $\lambda(w^{**}, b^{**})$ is an optimal solution of (17) for each $\lambda > 0$.

Notice that the constant $E$ in (19) and 20 can be set $E = 1$ due to the positive homogeneity of the norm function. Thus the optimization problems (19) and 20 can be equivalently recast with $E = 1$:

$$\underset{w,b}{\text{minimize}} \;\; \mathcal{R}(\mathcal{L}_\omega(w,b)) \quad \text{subject to} \;\; \|w\| = 1, \tag{30}$$

$$\underset{w,b}{\text{minimize}} \;\; \mathcal{R}(\mathcal{L}_\omega(w,b)) \quad \text{subject to} \;\; \|w\| \le 1. \tag{31}$$

Next we focus on a special case of positive homogeneous risk functionals, for which norm-constrained, risk-constrained and unconstrained formulations (20)-(22) under some additional conditions define the same classifiers.

### 3.1 Positive Homogeneous Risk Functionals

**Definition 2** *(Positive Homogeneity)* $\mathcal{R}(\lambda X) = \lambda\mathcal{R}(X)$ *for each random variable* $X$ *and constant* $\lambda > 0$, $\lambda \in \mathbb{R}$.

Note that among the risk functionals listed in the previous section, $\mathrm{ATL}_t$, which is given in (24), does not satisfy this property, and accordingly, we exclude the $C$-SVC from the analysis below. Several examples of positive homogeneous risk functionals are provided below.

- $\sup(\mathcal{L}_\omega) = \max\{\mathcal{L}_{\omega_1}, \ldots, \mathcal{L}_{\omega_l}\}$ Worst-Case Loss
- $\mathbb{E}(\mathcal{L}_\omega) = \frac{1}{l}\sum_{i=1}^{l} \mathcal{L}_{\omega_i}$ Expected Loss
- $ATL_0(\mathcal{L}_\omega) = \frac{1}{l}\sum_{i=1}^{l} [\mathcal{L}_{\omega_i}]_+$ Above-Zero Loss
- $MASD_t(\mathcal{L}_\omega) = \mathbb{E}(\mathcal{L}_\omega) + t\mathbb{E}([\mathcal{L}_\omega - \mathbb{E}(\mathcal{L}_\omega)]_+)$ with $t \ge 0$ Mean-Absolute Semi-Deviation
- $\mathrm{VaR}(\mathcal{L}_\omega) = \min_c \{c : \Pr\{\mathcal{L}_\omega \le c\} \ge \alpha\}$ Value-at-Risk
- $\mathrm{CVaR}(\mathcal{L}_\omega) = \min\limits_{c} \left\{ c + \frac{1}{(1-\alpha)l}\sum_{i=1}^{l} [\mathcal{L}_{\omega_i} - c]_+ \right\}$ Conditional Value-at-Risk
- $CRM(\mathcal{L}_\omega) = \max\limits_{q_1,\ldots,q_l} \{\sum\limits_{i=1}^{l} q_i\, \mathcal{L}_{\omega_i} : (q_1, ..., q_l) \in Q\}$ with $Q \subset \{(q_1, ..., q_l) : \sum\limits_{i=1}^{l} q_i = 1, q_i \ge 0, i = 1, ..., l\}$ Coherent Risk Measures

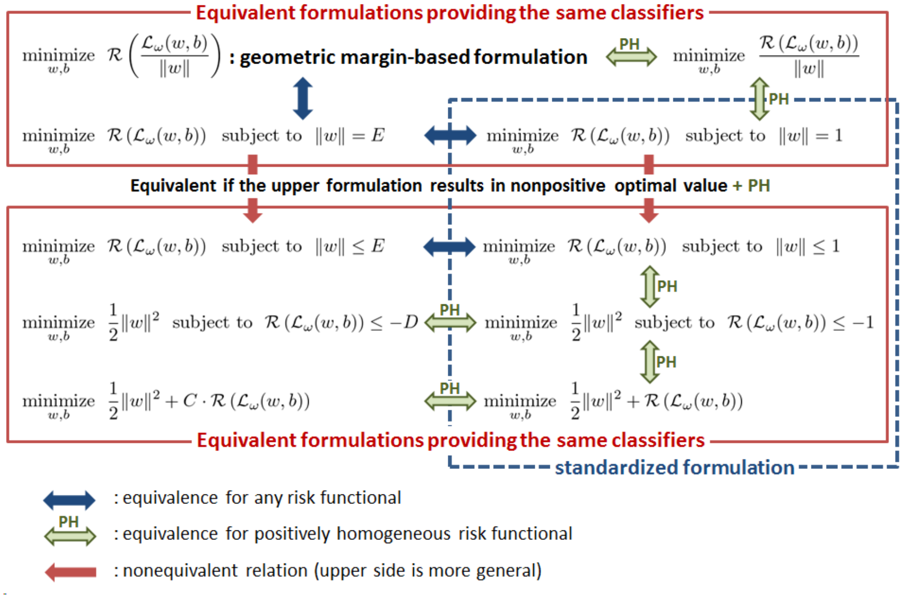
Equivalent formulations providing the same classifiers
$\underset{w,b}{\text{minimize}}\ \mathcal{R}\left(\frac{\mathcal{L}_\omega(w,b)}{\|w\|}\right)$ : geometric margin-based formulation
PH
$\underset{w,b}{\text{minimize}}\ \frac{\mathcal{R}(\mathcal{L}_\omega(w,b))}{\|w\|}$
PH
$\underset{w,b}{\text{minimize}}\ \mathcal{R}(\mathcal{L}_\omega(w,b))$ subject to $\|w\| = E$
$\underset{w,b}{\text{minimize}}\ \mathcal{R}(\mathcal{L}_\omega(w,b))$ subject to $\|w\| = 1$
Equivalent if the upper formulation results in nonpositive optimal value + PH
$\underset{w,b}{\text{minimize}}\ \mathcal{R}(\mathcal{L}_\omega(w,b))$ subject to $\|w\| \le E$
$\underset{w,b}{\text{minimize}}\ \mathcal{R}(\mathcal{L}_\omega(w,b))$ subject to $\|w\| \le 1$
PH
$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2$ subject to $\mathcal{R}(\mathcal{L}_\omega(w,b)) \le -D$
PH
$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2$ subject to $\mathcal{R}(\mathcal{L}_\omega(w,b)) \le -1$
PH
$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 + C \cdot \mathcal{R}(\mathcal{L}_\omega(w,b))$
PH
$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 + \mathcal{R}(\mathcal{L}_\omega(w,b))$
Equivalent formulations providing the same classifiers
standardized formulation
: equivalence for any risk functional
PH
: equivalence for positively homogeneous risk functional
: nonequivalent relation (upper side is more general)

Figure 2: Equivalence among several formulations with positive homogeneous risk functionals

If risk functional $\mathcal{R}(\cdot)$ is positive homogeneous (sometimes, denoted "PH"), all the structural risk minimization formulations, listed on the left-hand side of the above figure, have the same set of classifiers (independently of the values of the contained constants $C, D, E$), and accordingly, all the constants, i.e., $C, D, E$, in the structural risk minimization formulations can be set equal to 1, as listed on the right-hand side of the figure. On the other hand, the upper four formulations are equivalent to each other under the positive homogeneity and are more general than the below six ones in the sense that the lower formulations can achieve optimal solutions only if the upper ones have nonpositive optimal values.

Notice also that based on the set of positive homogeneous risk functionals $\mathcal{R}_1(\mathcal{L}_\omega), \ldots,$ $\mathcal{R}_k(\mathcal{L}_\omega)$ the new risk functionals can be obtained by applying operations that preserve positive homogeneity, e.g,

- $\mathcal{R}(\mathcal{L}_\omega) = \max\{\mathcal{R}_1(\mathcal{L}_\omega), \ldots, \mathcal{R}_k(\mathcal{L}_\omega)\}$
- $\mathcal{R}_k(\mathcal{L}_\omega) = \sum\limits_{i=1}^{k} \lambda_i \mathcal{R}_i(\mathcal{L}_\omega)$ for $\lambda_i \in \mathbb{R}$, $i = 1, \ldots k$.

### 3.2 Unboundedness of Optimal Solution

This section addresses an unboundedness of optimal solution of (17) for positive homogeneous risk functional $\mathcal{R}(\cdot)$. In addition, further we assume that $\mathcal{R}(\cdot)$ is also continuous:

**Definition 3** *(Continuity)* $\lim\limits_{k\to\infty} \mathcal{R}(\mathcal{L}_\omega^k) = \mathcal{R}(\mathcal{L}_\omega)$ *for each sequence* $\mathcal{L}_\omega^1, \mathcal{L}_\omega^2, \ldots$ *converging to* $\mathcal{L}_\omega$.

Notice that all mentioned in the previous section risk functionals except for value-at-risk are continuous. The following proposition states the conditions of unboundedness:

**Proposition 4** *Suppose that $\mathcal{R}(\cdot)$ is a positive homogeneous and continuous risk functional. Then optimization problem* (17) *is unbounded if*

$$\min\{\mathcal{R}(-y(\omega), \mathcal{R}(y(\omega)))\} < 0. \tag{32}$$

The proof of Proposition 4 is sketched in Appendix A.

### 3.3 Equivalence of Formulations with Positive Homogeneous Risk Functionals

Support Vector Machine is usually formulated using the structural risk minimization principle that can be expressed as a tradeoff between empirical risk and the regularization. Classification methods defined in primal as (20)-(22) provide several ways how this tradeoff can be expressed. With empirical function employed in C-SVM (3), all three formulations result in different classifiers. On the other hand, it is proved that all three formulation (20)-(22) with risk functionals $\mathrm{CVaR}_{1-\nu}(\cdot)$ define $\nu$-SVM (12). This fact can be explained by positive homogeneity of CVaR risk measure. This section aims at showing that for positive homogeneous risk functionals under a certain condition classifiers specified in primal by (20)-(22) provide the same separating hyperplane, which is also optimal for the geometric margin formulation (17).

**Proposition 5** *Suppose that $\mathcal{R}(\cdot)$ is a positive homogeneous and continuous risk functional. Then*

1. *If $(w^*, b^*)$ is an optimal solution of* (17) *with negative optimal objective value, then $\frac{E}{\|w^*\|}(w^*, b^*)$ is optimal for* (20).
2. *If $(w^*, b^*)$ is an optimal solution of* (20) *and $w^* \neq 0$, then $\lambda(w^*, b^*)$ is an optimal solution of* (17).

The proof of Proposition 5 is sketched in Appendix B.

**Proposition 6** *Suppose that $\mathcal{R}(\cdot)$ is a positive homogeneous and continuous risk functional. Then*

1. *If $(w^*, b^*)$ is optimal solution of* (17) *with negative optimal objective value, then $-\frac{D}{\zeta}(w^*, b^*)$ is an optimal solution of* (21)*, where $\zeta = |\mathcal{R}(\mathcal{L}_\omega(w^*, b^*))|$.*
2. *If $(w^*, b^*)$ is an optimal solution of* (21) *and $w^* \neq 0$, then $\lambda(w^*, b^*)$ is an optimal solution of* (17).

The proof of Proposition 6 is sketched in Appendix C.

**Proposition 7** *Suppose that $\mathcal{R}(\cdot)$ is a positive homogeneous and continuous risk functional. Then*

1. *If $(w^*, b^*)$ is optimal solution of* (17) *with negative optimal objective value, then $\frac{Cr^*}{\|w^*\|}(w^*, b^*)$ is an optimal solution of 22, where $r^* = -\frac{\mathcal{R}(\mathcal{L}_\omega(w^*,b^*))}{\|w^*\|}$.*
2. *If $(w^*, b^*)$ is an optimal solution of* (22) *and $w^* \neq 0$, then $\lambda(w^*, b^*)$ is an optimal solution of* (17).

The proof of Proposition 7 is sketched in Appendix D.

**Remark 8** *When* (22) *have non-zero optimal solutions,* (22) *determines the same hyperplane for different values of* $C > 0$*. Thus, the parameter* $C$ *can be set to* 1*.*

Propositions 5-7 imply the following corollary, which summarizes the relation between formulations (17), (20), (21), 22 .

**Corollary 9** *Suppose that* $\mathcal{R}(\cdot)$ *is a positive homogeneous and continuous risk functional. Then*

1. *When* (17) *has a negative optimal objective value, optimization problems* (17)*,* (20)*,* (21)*, 22 determine the same separating hyperplane.*
2. *When* (17) *has a positive or zero optimal objective value, optimization problems* (20)*,* (22) *have a trivial solution* ($w = 0$) *and* (21) *is infeasible.*

## 4. Existence of Optimal Solution

In this section, we explore under what condition there exists an optimal solution of optimization problem (17). We formulate the result for positive homogeneous and lower semi-continuous risk functionals $\mathcal{R}(\cdot)$. The lower semi-continuity assumption is more general than continuity and hold for all risk functionals listed in this dissertation.

**Proposition 10 (Existence of Optimal Solution)** *Suppose that* $\mathcal{R}(\cdot)$ *is positive homogeneous and lower semi-continuous. Then* (17) *has an optimal solution if*

$$\min\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} > 0 \tag{33}$$

The proof of Proposition 10 is sketched in Appendix E.

We should note that continuity is not necessary for the existence of an optimal solution or unboundedness of the minimization. In fact, although VaR, defined in (28), does not satisfy the (upper semi-)continuity, its minimization is unbounded if (32) holds (see Theorem 2 of Tsyurmasto et al. (2013) for details).

**Example: Expected loss.** Consider the case where the expected loss is used as the risk measure, i.e.,

$$\mathcal{R}(\mathcal{L}_\omega(w,b)) = \mathbb{E}[\mathcal{L}_\omega(w,b)] = -\sum_{\omega\in\Omega} \Pr(\omega) y(\omega)(w^T\phi(\xi(\omega)) + b).$$

Note that the expected loss is positive homogeneous and convex on $\mathbb{R}^m$ (i.e., continuous). In this case, we have

$$\begin{aligned}\min_{\omega\in\Omega}\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} &= \min\{\tfrac{1}{l}\sum_{i=1}^{l} y_i, -\tfrac{1}{l}\sum_{i=1}^{l} y_i\} \\ &= \min\{\tfrac{1}{l}(l_+ - l_-), \tfrac{1}{l}(l_- - l_+)\} = \begin{cases} < 0 & \text{if } l_+ \neq l_-, \\ = 0 & \text{if } l_+ = l_-. \end{cases}\end{aligned}$$

where $l_+ := |\{i : y_i = +1\}|$ and $l_- := |\{i : y_i = -1\}|$. Then Proposition 10 indicates that the expected loss-based SVM can have an optimal solution only if the number of samples of $y_i = +1$ is equal to that of $y_i = -1$. This is consistent with the result in Gotoh et al. (2013b), where a general probability setting $p_i = \Pr(\omega_i)$ is employed and the authors show that the condition $\sum_{i=1}^{l} p_i y_i = 0$ is the necessary and sufficient for the optimality. In addition, we should note that the expected loss based SVM admits any $b$ as an optimal solution even when the condition $l_+ = l_-$ holds, i.e., the number of the samples in each class is equal, and accordingly, it has the bounded optimal value. □

**Example: Worst-case loss (or maximum loss)** Given a set of data samples $\Omega = \{\omega_1, ..., \omega_l\}$ (or more specifically, $\{(\xi_1, y_1), \ldots, (\xi_l, y_l)\}$), consider the risk measure $\mathcal{R}(\cdot) = \sup(\cdot)$. The condition (33) is then given by

$$\min\{\sup\{-y_1, \ldots, -y_l\}, \sup\{y_1, \ldots, y_l\}\} = \min\{1, 1\} = 1 > 0.$$

Namely, when the worst-case loss is employed, the condition (33) is satisfied if and only if there is at least one sample of each class. □

**Example: CVaR.** Consider the case where the CVaR is used as the risk measure. By using the formula (26), we can easily check if the condition (33) is satisfied. Indeed, we can see that $\mathcal{R}(y(\omega)) > 0$ holds if and only if $(1-\alpha)l < 2l_+$ holds where $l_+ := |\{i : y_i = +1\}|$; $\mathcal{R}(-y(\omega)) > 0$ holds if and only if $(1-\alpha)l < 2l_-$ holds where $l_- := |\{i : y_i = -1\}|$. Accordingly, the condition (33) for the CVaR is given by

$$\alpha > \frac{1}{l}(1 - 2\min\{l_+, l_-\}). \tag{34}$$

It is noteworthy that this bound is consistent with the admissible range of the parameter $\nu$ for $\nu$-SVM Burges (2000). Also the condition (33) is consistent with the admissible range for the $\nu$-SVM shown in Chang and Lin (2002) and the condition in Lemma 2.2 of Gotoh and Takeda (2005), where the condition for the existence of an optimal solution of the geometric margin-based CVaR minimization formulation is given by $\alpha \geq 1 - 2\min\{\sum_{i:y_i=+1} p_i, \sum_{i:y_i=-1} p_i\}$ with the probability $p_i := \Pr(\omega_i)$. On the other hand, the condition (32) implies that for

$$\alpha < \frac{1}{l}(1 - 2\min\{l_+, l_-\}),$$

optimization problems (18) and (19)-(22) are unbounded. □

**Example: VaR.** Let us find an admissible range of parameter $\alpha$ for VaR-SVM (29). Since $y(\omega)$ is discretely distributed random variables with realizations $\{\underbrace{-1, \ldots, -1}_{l_-}, \underbrace{+1, \ldots, +1}_{l_+}\}$,

$$\mathrm{VaR}_\alpha(\pm y(\omega)) = \begin{cases} \geq 0 & \text{if } \alpha \geq \frac{l_\mp + 1}{l}, \\ < 0 & \text{if } \alpha < \frac{l_\mp + 1}{l}. \end{cases} \tag{35}$$

Thus, condition (33) holds when $\alpha \geq \min\{\alpha_+, \alpha_-\} + \frac{1}{l}$, while condition (32) holds when $\alpha < \min\{\alpha_+, \alpha_-\} + \frac{1}{l}$ (here we denote $\alpha_+ = \frac{l_+}{l}$ and $\alpha_- = \frac{l_-}{l}$ fractions of samples with positive and negative class labels, accordingly).

## 5. CVaR-$(\alpha_L, \alpha_U)$-SVM

In this section, we propose a new classifier formulated in the primal as

$$\min_{w,b} \, (1-\alpha_L)\cdot \mathrm{CVaR}_{\alpha_L}(d_\omega(w,b)) - (1-\alpha_U)\cdot \mathrm{CVaR}_{\alpha_U}(d_\omega(w,b)) \tag{36}$$

and further referred to as *CVaR-$(\alpha_L, \alpha_U)$-SVM.* It has two parameters: lower confidence level $\alpha_L \in [0,1]$ and upper confidence level $\alpha_U \in [0,1]$ such that $\alpha_L < \alpha_U$. In order to explain a meaning of the objective function of (36), we consider a profile representation of CVaR with confidence level $\alpha \in [0,1)$ applied to random distance function $d_\omega(w,b)$:

$$\mathrm{CVaR}_\alpha(d_\omega(w,b)) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\beta(d_\omega(w,b))d\beta, \tag{37}$$

where $\mathrm{VaR}_\beta(\cdot)$ is defined in (28), see Acerbi (2002) for the details. The definition (36) is illustrated with the Figure 3. Using (37), the objective function of (36) can be recast as

$$\int_{\alpha_L}^{\alpha_U} \mathrm{VaR}_\beta(d_\omega(w,b))d\beta.$$

Thus, objective (36) is roughly an average of distances between lower $\alpha_L$ and upper $\alpha_U$ percentiles of distribution $d_\omega(w,b)$, see Figure 4. In fact, $\nu$-SVM is a special case of CVaR-$(\alpha_L, \alpha_U)$-SVM with parameters $\alpha_L = 1-\nu$ and $\alpha_U = 1$. However, CVaR-$(\alpha_L, \alpha_U)$-SVM has an additional parameter $\alpha_U$, which roughly specifies that $(1-\alpha_U)\cdot 100\%$ data sample with highest distances to the hyperplane are disregarded. Outliers present in data are likely to have highest distances since they are misclassified the most and located far from the hyperplane. Thus, when dataset is free from outliers the parameter $\alpha_U$ can be chosen equal to 1 and, thus, CVaR-$(\alpha_L, \alpha_U)$-SVM performs as good as $\nu$-SVM. However, when dataset is contaminated by outliers, CVaR-$(\alpha_L, \alpha_U)$-SVM has an advantage of stability to outliers compared to $\nu$-SVM.

Note that CVaR-$(\alpha_L, \alpha_U)$-SVM is positive homogeneous. Clearly, problem (36) is a special case of (17) with $\mathcal{R}(\cdot) = (1-\alpha_L)\cdot \mathrm{CVaR}_{\alpha_L}(\cdot) - (1-\alpha_U)\cdot \mathrm{CVaR}_{\alpha_U}(\cdot)$. With Propositions 7, formulation (36) can be equivalently recast:

$$\underset{w,b}{\text{minimize}} \;\; \frac{1}{2}\|w\|^2 + (1-\alpha_L)\mathrm{CVaR}_{\alpha_L}(\mathcal{L}_\omega(w,b)) - (1-\alpha_U)\mathrm{CVaR}_{\alpha_U}(\mathcal{L}_\omega(w,b)) \tag{38}$$

as long as an optimal objective value of (17) is negative. Although (38) is unconstrained nonfractional optimization, it is still a nonconvex optimization. Accordingly, the deterministic global optimization methods (see, e.g., Horst and Tuy (2003); Horst et al. (2000); Horst and Thoai (1999)) are not promising except for very small instances. However, the objective of (38) is in the form of the so-called D.C. (difference of two convex functions) and efficient good heuristic algorithms such as DCA (see, e.g., Tao et al. (2005)) are available.

In addition, notice that the nonconvexity of the objective can be expected to be small when $\alpha_U$ is close to 1, i.e.,

$$\int_{\alpha_U}^1 \mathrm{VaR}_\beta(d_\omega(w,b))d\beta \approx 0, \text{ as } \alpha_U \approx 1,$$

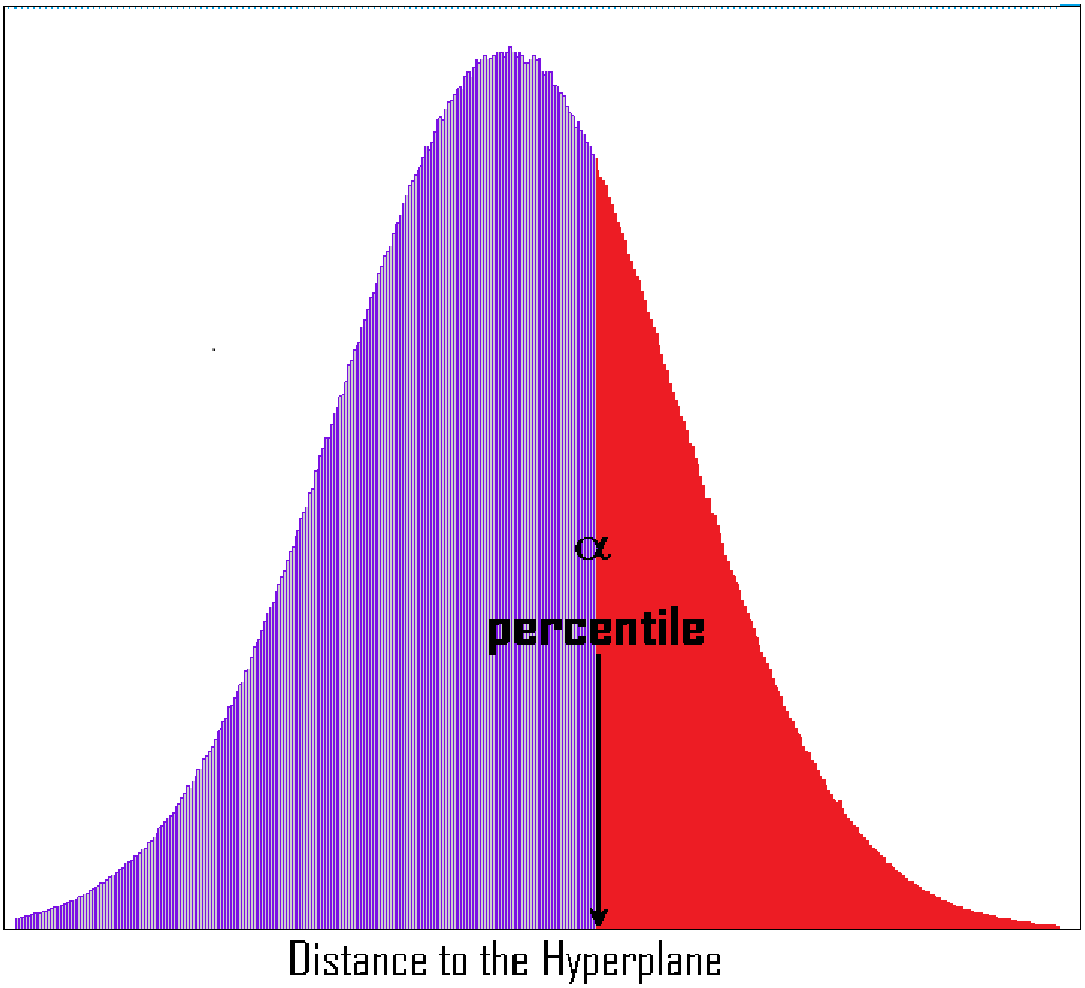

Figure 3: Distribution of distances to hyperplane. $\text{CVaR}_\alpha(\cdot)$ is calculated for distance function (15) as a normalized average of distances exceeding $\alpha$-percentile of distribution.

and, thus, objective (38) is expected to virtually remain convex. Since typically, data contains $< 3$ to 5% outliers, it is enough to set $\alpha_U \in [0.95, 1)$ for discarding those outliers.

In Section 6, we show that the heuristic approach to CVaR-$(\alpha_L, \alpha_U)$-SVM achieves a superior out-of-sample performance compared to $\nu$-SVM on the real-life data contaminated by outliers.

## 6. Nonlinear CVaR-$(\alpha_L, \alpha_U)$-SVM

This section provides a nonlinear extension of CVaR-$(\alpha_L, \alpha_U)$-SVM. Given a training set $\{(\xi_1, y_1), \ldots, (\xi_l, y_l)\}$ of features $\xi_i$ with binary class labels $y_i \in \{-1, 1\}$ for $i = 1, \ldots, l$, a

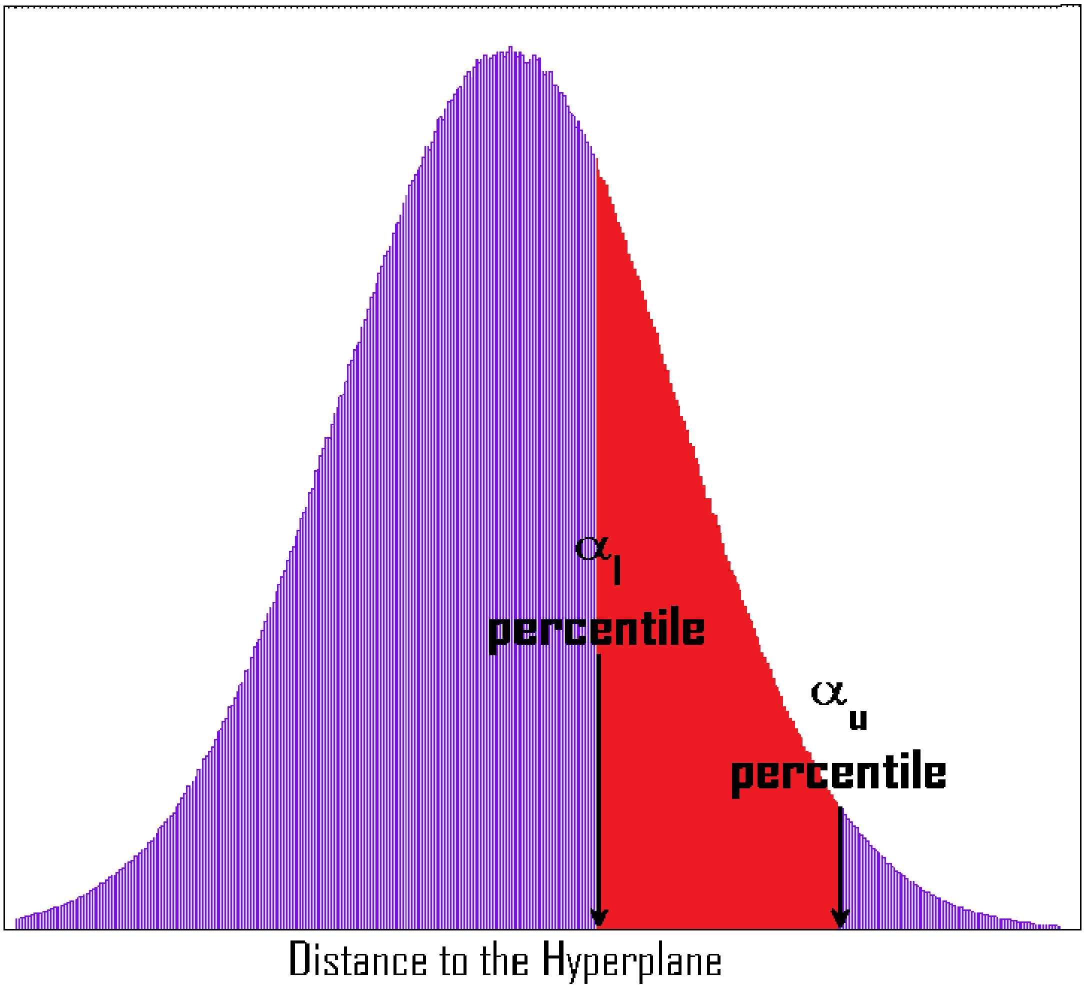

Figure 4: Distribution of distances to hyperplane. $(1-\alpha_L)\text{CVaR}_{\alpha_L}(\cdot)-(1-\alpha_U)\text{CVaR}_{\alpha_U}(\cdot)$ is calculated for distance function (15) as a normalized average between $\alpha_L$ and $\alpha_U$ percentiles of distribution.

non-linear SVM can be constructed by a transformation of the original features $\{\xi_1,\ldots,\xi_l\}$ into features $\{\phi(\xi_1),\ldots,\phi(\xi_l)\}$ with the mapping $\phi:\mathbb{R}^m\to\mathbb{R}^n$.

The transformation $\phi$ is usually implicitly specified by a kernel function $K(\xi,\xi')$ (Muller et al. (2001)). However, CVaR-$(\alpha_L,\alpha_U)$-SVM is not always convex and cannot be solved through its dual. This section shows how to construct non-linear CVaR-$(\alpha_L,\alpha_U)$-SVM with a kernel (e.g., Gaussian (RBF) kernel).

There exists a linear transformation $\psi$ of the original set of features $\{\xi_1,\ldots,\xi_l\}\subset\mathbb{R}^m$ such that the scalar products of $\psi(\xi_j)$ and $\psi(\xi_j)$ are equal to those produced by $K(\xi,\xi')$, i.e. $\langle\psi(\xi_i),\psi(\xi_j)\rangle = K(\xi_i,\xi_j)\equiv\langle\phi(\xi_i),\phi(\xi_j)\rangle$ for all $i$ and $j$ (see, e.g., Cristianini and Shawe-Taylor (2000)), so that the solution of the primal problem with the transformed features

$\{\psi(\xi_1), \dots, \psi(\xi_l)\} \subset \mathbb{R}^n$ coincides with that for the dual problem with the kernel $K(\xi, \xi')$ corresponding to the original transformation $\phi$ (Chapelle (2007)).

For features $\{\xi_1, \dots, \xi_l\}$, the kernel $K(\xi, \xi')$ yields a positive definite kernel matrix $K = \{K(\xi_i, \xi_j)\}_{i,j=1,\dots,l}$, which can be decomposed as

$$K = V\Lambda V^T \equiv (V\Lambda^{\frac{1}{2}})(V\Lambda^{\frac{1}{2}})^T, \tag{39}$$

where $\Lambda = diag(\lambda_1, \dots, \lambda_l)$ is a diagonal matrix with eigenvalues $\lambda_1 > 0, \dots, \lambda_l > 0$ and $V = (v_1, \dots, v_l)$ is an orthogonal matrix with corresponding eigenvectors $v_1, \dots, v_l$ of $K$. The representation (39) implies that $\psi : \xi_i \to (V\Lambda^{\frac{1}{2}})_i$, $i = 1, \dots, l$, is the sought linear transformation, where $(V\Lambda^{\frac{1}{2}})_i$ is row $i$ of the matrix $(V\Lambda^{\frac{1}{2}})$. Thus, the nonlinear version of (36) has the following explicit formulation

$$\underset{\lambda, \lambda_0}{\text{minimize}} \quad \frac{1}{2}\|\lambda\|^2 + (1-\alpha_L)\text{CVaR}_{\alpha_L}(\mathcal{L}_\omega(\lambda, \lambda_0)) - (1-\alpha_U)\text{CVaR}_{\alpha_U}(\mathcal{L}_\omega(\lambda, \lambda_0)) \tag{40}$$

with new decision variables $\lambda \in \mathbb{R}^l$, $\lambda_0 \in \mathbb{R}$ and discretely distributed loss function $\mathcal{L}_\omega(\lambda, \lambda_0)$ given by observations (scenarios)

$$\mathcal{L}_{\omega_i}(\lambda, \lambda_0) = -y_i[\lambda^T \phi(\xi_i) + \lambda_0],$$

and $\phi(\xi_i) = (V\Lambda^{\frac{1}{2}})_i$ for $i = 1, \dots, l$.

## 7. Computational Experiment

The computations are performed with MATLAB using Portfolio Safeguard (PSG)[1] solver, which applies advanced techniques for optimizing CVaR function. With PSG, solving problems involves three main stages:

1. *Mathematical formulation of optimization problem using precoded CVaR functions.* Typically, a problem formulation involves 5-10 operators of a meta-code. See, for instance, Appendix F with the example of problem formulation for optimization problem (38).

2. *Preparation of data for the PSG functions in an appropriate format.* In our experiment, CVaR functions are defined on the matrix of training samples $\{(\xi_1, y_1), \dots, (\xi_l, y_l)\}$.

3. *Solving the optimization problem with PSG using the predefined problem statement and data for PSG functions. The problem can be solved in several PSG environments, such as MATLAB environment and Run-File (Text) environment.*

The problems (36) and $\nu$-SVM are solved with datasets from UCI Machine Learning Repository[2]: Liver Disorders, Heart Disease, Indian Diabetes, German and Ionosphere.

The original features were normalized (zero mean, unit standard deviation). To calculate testing accuracy, we use 10-fold cross validation. We use 2/3 of the training set to solve (12) and (36) and the remaining 1/3 to fit the parameters $\nu$ in (12) and $\alpha_L, \alpha_r$ in (36). Parameter $\nu$ and $\alpha_L$ are selected from the grid 0 : 0.05 : 1, while parameter $\alpha_U$ is selected from the grid 0.9 : 0.01 : 1.

[1] `http://www.aorda.com/aod/welcome.action/psg.action`

[2] `http://archive.ics.uci.edu/ml/datasets.html`

Table 1: Experimental results for Liver Disorders Dataset, contaminated by outliers.

| Percent of Outliers (%) | $\nu$-SVM accuracy (%) | | | | CVaR-$(\alpha_L, \alpha_U)$-SVM accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | training | | testing | | training | | testing | |
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 0 | 70.43 | 1.18 | **66.13** | 2.09 | 72.22 | 1.47 | **71.17** | 1.73 |
| 1 | 69.07 | 1.81 | **65.43** | 2.55 | 72.72 | 2.74 | **71.65** | 2.65 |
| 5 | 60.52 | 2.41 | **60.65** | 1.72 | 72.26 | 1.35 | **68.91** | 1.14 |
| 10 | 61.63 | 1.92 | **59.57** | 2.46 | 74.15 | 2.01 | **70.78** | 1.86 |

Table 2: Experimental Results for Heart Disease Dataset, contaminated by outliers.

| Percent of Outliers (%) | $\nu$-SVM accuracy (%) | | | | CVaR-$(\alpha_L, \alpha_U)$-SVM accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | training | | testing | | training | | testing | |
| | Mean | STD | Mean | STD | Mean | STD | Mean | Std |
| 0 | 85.15 | 1.77 | **82.09** | 2.72 | 85.54 | 1.13 | **83.16** | 1.21 |
| 1 | 85.15 | 2.58 | **82.60** | 1.44 | 86.12 | 1.56 | **83.27** | 2.07 |
| 5 | 84.59 | 1.54 | **82.24** | 2.19 | 87.17 | 2.04 | **83.27** | 1.62 |
| 10 | 75.59 | 2.36 | **74.23** | 3.01 | 86.64 | 1.13 | **83.57** | 1.74 |

### 7.1 Linear SVM

Linear $\nu$-SVM and CVaR-$(\alpha_L, \alpha_U)$-SVM performed 66.13% and 71.17% on Liver Disorder Dataset, 82.09% and 83.16% on Heart Disease Dataset, 76.05% and 76.29% on Indian Diabetes Dataset, 69.43% and 71.27% on Ionosphere DataSet, accordingly. Outliers were generated by artificially multiplying a fraction of 0%, 1%, 5%, 10% of the original dataset by 1000. Tables 1, 2, 4, 3, 5 show performances of $\nu$-SVM and CVaR-$(\alpha_L, \alpha_U)$-SVM as the percentage of outliers increases, the performance of $\nu$-SVM drops, while CVaR-$(\alpha_L, \alpha_U)$-SVM has almost the same performance. Tables 7 and 6 show optimal parameters $\nu$ for $\nu$-SVM and $\alpha_L, \alpha_r$ for CVaR-$(\alpha_L, \alpha_U)$-SVM.

### 7.2 Non-Linear SVM

We compare out-of-sample performance of non-linear CVaR-$(\alpha_L, \alpha_U)$-SVM against $\nu$-SVM with Gaussian kernel. The original features are transformed as described in Section 6. Table 8 summarizes results of the experiment. CVaR-$(\alpha_L, \alpha_U)$-SVM outperformed $\nu$-SVM on all datasets, especially Indian Diabetes and Ionosphere.

## 8. Conclusion

The paper presented a unified scheme for classification based on geometric margin. The scheme encompasses several well-known SVMs. The relation between existing SVMs was established using positive homogeneity of corresponding risk functionals. A unified scheme with linear loss function was extended to non-linear case. As a special case of unified

Table 3: Experimental Results for German Credit Dataset, contaminated by outliers.

| Percent of Outliers (%) | $\nu$-SVM accuracy (%) | | | | CVaR-$(\alpha_L, \alpha_U)$-SVM accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | training | | testing | | training | | testing | |
| | Mean | STD | Mean | STD | Mean | STD | Mean | Std |
| 0 | 77.83 | 1.21 | **76.05** | 1.56 | 78.64 | 0.77 | **76.26** | 1.38 |
| 1 | 76.60 | 1.98 | **70.87** | 1.36 | 78.28 | 0.08 | **76.19** | 1.57 |
| 5 | 74.45 | 2.23 | **70.11** | 2.23 | 76.24 | 1.65 | **75.63** | 1.68 |
| 10 | 71.95 | 2.12 | **70.45** | 3.89 | 76.58 | 1.06 | **75.33** | 1.20 |

Table 4: Experimental Results for Indian Diabetes Dataset, contaminated by outliers.

| Percent of Outliers (%) | $\nu$-SVM accuracy (%) | | | | CVaR-$(\alpha_L, \alpha_U)$-SVM accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | training | | testing | | training | | testing | |
| | Mean | STD | Mean | STD | Mean | STD | Mean | Std |
| 0 | 77.68 | 2.13 | **77.99** | 1.64 | 77.68 | 2.60 | **77.99** | 1.71 |
| 1 | 77.17 | 1.34 | **74.15** | 1.82 | 77.91 | 2.02 | **76.84** | 1.78 |
| 5 | 66.00 | 3.69 | **62.86** | 4.13 | 78.88 | 1.46 | **77.79** | 2.28 |
| 10 | 61.18 | 8.19 | **59.66** | 7.63 | 78.23 | 2.92 | **77.24** | 2.17 |

scheme a new classifier based difference of two CVaR functions was proposed and its robust properties were confirmed empirically.

## Acknowledgments

This research has been supported by the AFOSR grant FA9550-11-1-0258, New Developments in Uncertainty: Linking Risk Management, Reliability, Statistics and Stochastic Optimization.

## Appendix A. Proof of Proposition 4

Due to Proposition 1, it is enough to show that (19) is unbounded if the condition (32) holds. Suppose, for instance, that $\mathcal{R}(-y(\omega)) < 0$. Consider the behavior of the function $\mathcal{R}(\mathcal{L}_\omega(w, b))$ as $b \to +\infty$. Using the positive homogeneity of $\mathcal{R}(\cdot)$, we obtain

$$\mathcal{R}(\mathcal{L}_\omega(w, b)) = \mathcal{R}(-y(\omega)(w^T\xi(\omega) + b)) = b \cdot \mathcal{R}\left(-\frac{y(\omega)w^T\xi(\omega)}{b} - y(\omega)\right), \text{ for } b > 0.$$

Let us take an arbitrary sequence $b_k$ such that $\lim\limits_{k\to\infty} b_k = +\infty$. Then for each $\omega \in \Omega$, taking into account that $\|w\| = E$, we have

$$\lim_{k\to\infty} -\frac{y(\omega)w^T\xi(\omega)}{b_k} - y(\omega) = -y(\omega).$$

Table 5: Experimental Results for Ionosphere Dataset, contaminated by outliers.

| Percent of Outliers (%) | $\nu$-SVM accuracy (%) | | | | CVaR-$(\alpha_L, \alpha_U)$-SVM accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | training | | testing | | training | | testing | |
| | Mean | STD | Mean | STD | Mean | STD | Mean | Std |
| 0 | 78.39 | 1.29 | **69.43** | 1.88 | 84.14 | 1.36 | **71.27** | 1.73 |
| 1 | 77.33 | 2.13 | **68.51** | 1.23 | 80.09 | 1.03 | **70.96** | 2.68 |
| 5 | 76.29 | 2.56 | **66.04** | 3.02 | 79.96 | 1.73 | **71.01** | 2.92 |
| 10 | 77.22 | 1.97 | **66.94** | 1.05 | 71.31 | 2.18 | **70.70** | 1.92 |

Table 6: Optimal values of parameters of Linear CVaR-$(\alpha_L, \alpha u)$-SVM.

| Data Set | OPTIMAL PARAMETERS $(\alpha_L, \alpha_U)$<br>FOR DIFFERENT % OF OUTLIERS | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0% | | 1% | | 5% | | 10% | |
| | $\alpha_L$ | $\alpha_U$ | $\alpha_L$ | $\alpha_U$ | $\alpha_L$ | $\alpha_U$ | $\alpha_L$ | $\alpha_U$ |
| Liver Disorder | 0.3 | 0.96 | 0.45 | 0.9 | 0.55 | 0.88 | 0.5 | 0.85 |
| Heart Disease | 0.65 | 0.94 | 0.40 | 0.95 | 0.65 | 0.93 | 0.60 | 0.95 |
| German Credit | 0.45 | 0.95 | 0.50 | 0.98 | 0.60 | 0.94 | 0.60 | 0.90 |
| Indian Diabetes | 0.45 | 1 | 0.4 | 0.98 | 0.45 | 0.95 | 0.55 | 0.94 |
| Ionosphere | 0.65 | 0.92 | 0.40 | 0.90 | 0.35 | 0.97 | 0.50 | 0.91 |

Due to the continuity of $\mathcal{R}(\cdot)$, we obtain

$$\lim_{k\to\infty} \mathcal{R}\left(\frac{y(\omega)w^T\xi(\omega)}{b_k} - y(\omega)\right) = \mathcal{R}(-y(\omega)) < 0.$$

Therefore, $\lim\limits_{b\to+\infty} \mathcal{R}(\mathcal{L}_\omega(w,b)) = \lim\limits_{b\to+\infty} b\cdot\mathcal{R}\left(-\frac{y(\omega)w^T\xi(\omega)}{b} - y(\omega)\right) = -\infty$, which implies that (19) is unbounded.

Now, suppose that $\mathcal{R}(y(\omega)) < 0$. Applying the same reasoning as for the case $\mathcal{R}(-y(\omega)) < 0$, we obtain

$$\lim_{b\to-\infty} \mathcal{R}(\mathcal{L}_\omega(w,b)) = \lim_{b\to-\infty} \mathcal{R}(-y(\omega)(w^T\xi(\omega) + b)) =$$
$$\lim_{b\to-\infty} |b| \cdot \mathcal{R}\left(-\frac{y(\omega)w^T\xi(\omega)}{b} + y(\omega)\right) = \lim_{b\to-\infty} |b| \cdot \mathcal{R}(y(\omega)) = -\infty,$$

which implies that (19) is unbounded and condition $\min\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} \geq 0$ is necessary for the existence of an optimal solution of (19). □

## Appendix B. Proof of Proposition 5

1. Suppose that $(w^*, b^*)$ is an optimal solution of (17) with negative objective value, then $\left(\frac{Ew^*}{\|w^*\|}, \frac{Eb^*}{\|w^*\|}\right)$ is an optimal solution of (19) due to Proposition (1). Moreover, the optimal objective value of (19) is negative since $\mathcal{R}\left(\mathcal{L}_\omega\left(\frac{Ew^*}{\|w^*\|}, \frac{Eb^*}{\|w^*\|}\right)\right) = E\mathcal{R}\left(\frac{\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}\right) <$

Table 7: Optimal values of parameter $1-\nu$ for $\nu$-SVM.

| Data Set | OPTIMAL PARAMETERS $\nu$ FOR DIFFERENT % OF OUTLIERS | | | |
|---|---|---|---|---|
| | 0% | 1% | 5% | 10% |
| Liver Disorder | 0.80 | 0.75 | 0.95 | 0.95 |
| Heart Disease | 0.4 | 0.55 | 0.55 | 0.55 |
| German Credit | 0.55 | 0.60 | 0.95 | 0.95 |
| Indian Diabetes | 0.60 | 0.60 | 0.95 | 0.95 |
| Ionosphere | 0.70 | 0.75 | 0.90 | 0.95 |

Table 8: Experimental Results for kernel $\nu$-SVM and CVaR-$(\alpha_L, \alpha_r)$-SVM.

| Data Set | $\nu$-SVM accuracy (%) | | | | CVaR-$(\alpha_L, \alpha_U)$-SVM accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | training | | testing | | training | | testing | |
| | Mean | STD | Mean | STD | Mean | STD | Mean | Std |
| Liver Disorder | 77.68 | 2.13 | **77.99** | 1.64 | 77.68 | 2.60 | **77.99** | 1.71 |
| Heart Disease | 77.17 | 1.34 | **73.26** | 1.82 | 77.91 | 2.02 | **73.75** | 1.78 |
| German Credit | 66.00 | 3.69 | **69.68** | 4.13 | 78.88 | 1.46 | **69.96** | 2.28 |
| Indian Diabetes | 61.18 | 8.19 | **72.97** | 7.63 | 78.23 | 2.92 | **73.94** | 2.17 |
| Ionosphere | 87.30 | 4.24 | **87.30** | 7.63 | 78.23 | 2.92 | **88.67** | 2.17 |

0. Since the difference between the optimization problems (19) and (20) is only in a norm constraint on $w$, it is enough to show that $\|\hat{w}\| = E$ holds for each optimal solution $(\hat{w}, \hat{b})$ of (20). To obtain contradiction, suppose that $\|\hat{w}\| < E$. If $\hat{w} \neq 0$, then $\mathcal{R}\left(\mathcal{L}_\omega\left(\frac{E\hat{w}}{\|\hat{w}\|}, \frac{E\hat{b}}{\|\hat{w}\|}\right)\right) = \frac{E}{\|\hat{w}\|}\mathcal{R}\left(\mathcal{L}_\omega(\hat{w}, \hat{b})\right) < \mathcal{R}\left(\mathcal{L}_\omega(\hat{w}, \hat{b})\right)$ since $\frac{E}{\|\hat{w}\|} > 1$ and optimal objective value of (19) is negative. Thus, we obtain contradiction with optimality of $(\hat{w}, \hat{b})$. If $\hat{w} = 0$, then an objective value of (20) can be rewritten as $\mathcal{R}(-y(\omega)\cdot\hat{b})$. Since (17) has an optimal solution, then condition $\min\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} \geq 0$ holds due to Proposition 4. Thus, $\mathcal{R}\left(\mathcal{L}_\omega\left(\frac{Ew^*}{\|w^*\|}, \frac{Eb^*}{\|w^*\|}\right)\right) < 0 \leq \mathcal{R}\left(\mathcal{L}_\omega(\hat{w}, \hat{b})\right) = \mathcal{R}(-y(\omega)\cdot\hat{b})$, which contradicts an optimality of $(\hat{w}, \hat{b})$.

2. Suppose that $(w^*, b^*)$ is an optimal solution of (20) such that $w^* \neq 0$. Then it follows that $\|w^*\| = E$ as proved in the first part of this Proposition. Thus, $(w^*, b^*)$ is an optimal solution of (19). Applying Proposition (1), we obtain that $(\lambda w^*, \lambda b^*)$ is optimal solution of (17) for each $\lambda > 0$.

## Appendix C. Proof of Proposition 6

1. For short, denote $\zeta = |\mathcal{R}(\mathcal{L}_\omega(w^*, b^*))|$. Suppose that $(w^*, b^*)$ is optimal for (17) with negative optimal objective value, then $\frac{D}{\zeta}(w^*, b^*)$ is also optimal for (17) and $\frac{D}{\zeta}\mathcal{R}(\mathcal{L}_\omega(w^*, b^*)) = -D$. Then $\frac{D}{\zeta}(w^*, b^*)$ is optimal to:

$$\underset{w,b}{\text{minimize}} \;\; \mathcal{R}\left(\frac{\mathcal{L}_\omega(w,b)}{\|w\|}\right) \quad \text{subject to} \;\; \mathcal{R}(\mathcal{L}_\omega(w,b)) = -D, \qquad (41)$$

Optimization problem (41) can be equivalently recast:

$$\underset{w,b}{\text{minimize}} \;\; \frac{1}{2}\|w\|^2 \quad \text{subject to} \;\; \mathcal{R}(\mathcal{L}_\omega(w,b)) = -D, \tag{42}$$

Indeed $\mathcal{R}(-b \cdot y(\omega)) \geq 0$ due to Proposition 4, which implies that optimal solution of (42) satisfies condition $w \neq 0$. Thus, $(w^*, b^*)$ is optimal for (42). Let us show that $(w^*, b^*)$ is optimal for (21). To obtain contradiction, suppose that $(w^*, b^*)$ is not optimal for (21), i.e. $\mathcal{R}(\mathcal{L}_\omega(w^*, b^*)) < -D$, then $(\hat{w}, \hat{b}) = \frac{D}{|\mathcal{R}(\mathcal{L}_\omega(\hat{w},\hat{b}))|}(w^*, b^*)$ is feasible for (21) and $\|\hat{w}\| < \|w^*\|$, which contradicts the optimality of $(w^*, b^*)$. Thus, $(w^*, b^*)$ is optimal for (21).

2. Suppose that $(w^*, b^*)$ is an optimal solution of (21) and $w^* \neq 0$. Then $\mathcal{R}(\mathcal{L}_\omega(w^*, b^*)) = -D$ as proved above. With this fact and positive homogeneity of $\mathcal{R}(\cdot)$, the solution $(w^*, b^*)$ is optimal to (41) Then $\lambda(w^*, b^*)$ is optimal for (17) for each $\lambda > 0$.

## Appendix D. Proof of Proposition 7

1. Suppose that $(w^*, b^*)$ is an optimal solution of (17) with negative optimal objective value $\mathcal{R}\left(\frac{\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}\right) < 0$. For short, let us make a notation $r^* = -\mathcal{R}\left(\frac{\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}\right)$. Objective of (22) can be rewritten as follows:

$$\frac{1}{2}\|w\|^2 + C \cdot \mathcal{R}(\mathcal{L}_\omega(w,b)) = \begin{cases} \frac{1}{2}\|w\|^2 + C \cdot \frac{\mathcal{R}(\mathcal{L}_\omega(w,b))}{\|w\|}\|w\|, & \text{if } w \neq 0 \\ \mathcal{R}(-b \cdot y(\omega)), & \text{if } w = 0. \end{cases} \tag{43}$$

Since we have $\mathcal{R}(-b \cdot y(\omega)) \geq 0$ due to Proposition 4, it suffices to consider the case $w \neq 0$. Optimality of $(w^*, b^*)$ for (17) and positive homogeneity of $\mathcal{R}(\cdot)$ implies that $\frac{\mathcal{R}(\mathcal{L}_\omega(w,b))}{\|w\|} \geq \frac{\mathcal{R}(\mathcal{L}_\omega(w^*,b^*))}{\|w^*\|} = -r^*$ for each $(w,b) \in \mathbb{R}^{n+1}$, $w \neq 0$ and we have

$$\begin{aligned} \frac{1}{2}\|w\|^2 + C \cdot \frac{\mathcal{R}(\mathcal{L}_\omega(w,b))}{\|w\|}\|w\| &\geq \frac{1}{2}\|w\|^2 - Cr^*\|w\| \\ &= \frac{1}{2}(\|w\| - Cr^*)^2 - \frac{1}{2}C^2(r^*)^2 \geq -\frac{1}{2}(Cr^*)^2. \end{aligned} \tag{44}$$

and the lower bound $-\frac{1}{2}(Cr^*)^2$ in (44) is attained on the point $\frac{Cr^*}{\|w^*\|}(w^*, b^*)$. and consequently, the point $\frac{Cr^*}{\|w^*\|}(w^*, b^*)$ is optimal for (22).

2. Suppose that $(w^*, b^*)$ is an optimal solution of (22) and $w^* \neq 0$. To obtain contradiction, suppose that $\lambda(w^*, b^*)$ is not optimal solution of (17), i.e., there exists a $(\tilde{w}, \tilde{b})$ such that

$$\mathcal{R}\left(\frac{\mathcal{L}_\omega(\tilde{w},\tilde{b})}{\|\tilde{w}\|}\right) < \mathcal{R}\left(\frac{\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}\right). \tag{45}$$

Let $\tilde{r} = -\frac{\mathcal{R}(\mathcal{L}_\omega(\tilde{w},\tilde{b}))}{\|\tilde{w}\|}$. Then we have

$$\begin{aligned} &\frac{1}{2}\|w^*\|^2 + C \cdot \mathcal{R}(\mathcal{L}_\omega(w^*, b^*)) = \frac{1}{2}\|w^*\|^2 + C \cdot \mathcal{R}\left(\frac{\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}\right)\|w^*\| > \\ &\quad \frac{1}{2}\|w^*\|^2 + C \cdot \frac{\mathcal{R}(\mathcal{L}_\omega(\tilde{w},\tilde{b}))}{\|\tilde{w}\|}\|w^*\| = \frac{1}{2}(\|w^*\| - C\tilde{r})^2 - \frac{1}{2}(C\tilde{r})^2 \geq -\frac{1}{2}(C\tilde{r})^2. \end{aligned} \tag{46}$$

This implies that $\frac{C\tilde{r}}{\|\tilde{w}\|}(\tilde{w}, \tilde{b})$ attains a smaller objective value of (22) than $(w^*, b^*)$, contradicting the optimality of $(w^*, b^*)$. □

## Appendix E. Proof of Proposition 10

Due to Proposition 1, it is enough to show that (19) has an optimal solution if the condition (33) holds. Suppose that $\min\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} > 0$ and prove that (19) has an optimal solution. First, observe $\mathcal{R}(\mathcal{L}_\omega(w, b))$ is lower semi-continuous with respect to $(w, b)$. (Note that $\mathcal{R}(\cdot)$ is lower semi-continuous and $\mathcal{L}_\omega(\cdot, \cdot)$ is an affine function). Second, we prove that if $\min\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} > 0$, then $\lim\limits_{b \to \infty} \mathcal{R}(\mathcal{L}_\omega(w, b)) = \infty$. Let us find, for instance, $\lim\limits_{b \to +\infty} \mathcal{R}(\mathcal{L}_\omega(w, b))$. Again, applying the positive homogeneity of $\mathcal{R}(\cdot)$, we obtain

$$\mathcal{R}(\mathcal{L}_\omega(w, b)) = \mathcal{R}(-y(\omega)[w^T \xi(\omega) + b]) = b \cdot \mathcal{R}\left(-\frac{y(\omega) w^T \xi(\omega)}{b} - y(\omega)\right), \ \text{ for } b > 0.$$

Applying lower semi-continuity of $\mathcal{R}(\cdot)$, we obtain

$$\lim_{b \to +\infty} \mathcal{R}\left(-\frac{y(\omega) w^T \xi}{b} - y(\omega)\right) \geq \mathcal{R}(-y(\omega)) > 0.$$

Therefore, $\lim\limits_{b \to +\infty} \mathcal{R}(\mathcal{L}_\omega(w, b)) = \infty$. In a similar way, it can be shown that $\mathcal{R}(y(\omega)) > 0$ implies $\lim\limits_{b \to -\infty} \mathcal{R}(\mathcal{L}_\omega(w, b)) = \infty$.

Now, we show that an optimal solution of (19) exists, if $\min\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} > 0$. For each $\delta > 0$ the following optimization problem has an optimal solution

$$\underset{w,b}{\text{minimize}} \ \ \mathcal{R}(\mathcal{L}_\omega(w, b)) \quad \text{subject to} \quad \|w\| = E, \ |b| \leq \delta. \tag{47}$$

since the lower semi-continuous function $\mathcal{R}(\mathcal{L}_\omega(w, b))$ is minimized over a compact set. Denote by $(w_\delta^*, b_\delta^*)$ an optimal solution of problem (47). If there exists $\delta > 0$, such that for $(w_\delta^*, b_\delta^*)$ and for each $(w_\delta, b_\delta)$, $w_\delta \in \mathbb{R}^n$, $b_\delta \in \mathbb{R}$ satisfying $\|w_\delta\| = E$, $|b_\delta| \geq \delta$, we have $\mathcal{R}(\mathcal{L}_\omega(w_\delta^*, b_\delta^*)) \leq \mathcal{R}(\mathcal{L}_\omega(w_\delta, b_\delta))$, then $(w_\delta^*, b_\delta^*)$ is the optimal solution of (19). To obtain contradiction, suppose that for each $\delta > 0$ there exists $(\tilde{w}_\delta, \tilde{b}_\delta)$, $\tilde{w}_\delta \in \mathbb{R}^n$, $\tilde{b}_\delta \in \mathbb{R}$ satisfying $\|\tilde{w}_\delta\| = E$, $|\tilde{b}_\delta| \geq \delta$ such that $\mathcal{R}(\mathcal{L}_\omega(\tilde{w}_\delta, \tilde{b}_\delta)) < \mathcal{R}(\mathcal{L}_\omega(w_\delta^*, b_\delta^*))$. We have $\mathcal{R}(\mathcal{L}_\omega(w_\delta^*, b_\delta^*)) \leq \mathcal{R}(\mathcal{L}_\omega((w_1^*, b_1^*))$ for each $\delta \geq 1$ since minimization over a broader feasible region yields a lower optimal objective value. Therefore, $\mathcal{R}(\mathcal{L}_\omega(\tilde{w}_\delta, \tilde{b}_\delta)) < \mathcal{R}(\mathcal{L}_\omega(w_\delta^*, b_\delta^*)) \leq \mathcal{R}(\mathcal{L}_\omega(w_1^*, b_1^*))$ for $\delta \geq 1$. Since $\lim\limits_{\delta \to \infty} \tilde{b}_\delta = \infty$ and $\lim\limits_{b \to \infty} \mathcal{R}(\mathcal{L}_\omega(w, b)) = \infty$ (as we proved earlier), then $\lim_{\delta \to \infty} \mathcal{R}(\mathcal{L}_\omega(\tilde{w}_\delta, \tilde{b}_\delta)) = \infty$, which contradicts the bound $\mathcal{R}(\mathcal{L}_\omega(\tilde{w}_\delta, \tilde{w}_\delta)) \leq \mathcal{R}(\mathcal{L}_\omega(w_1^*, b_1^*))$ for $\delta \geq 1$. Thus, we proved that an optimal solution of (19) exists, if $\min\{\mathcal{R}(-y(\omega)), \mathcal{R}(y(\omega))\} > 0$.

## Appendix F. Example of PSG Meta-Code

This appendix presents the PSG meta-code for solving optimization problem (38). Meta-code, Data and Solutions can be downloaded from the link[3], see Problem 3c.

**Meta-Code for optimization problem** (38)

```
1. Problem:  problem_var_nu_svm, type = minimize

2. Objective:  objective_svm

3. 0.5*quadratic_matrix_quadratic(matrix_quadratic)

4. 0.7*cvar_risk_1(0.3,matrix_prior_scenarios)

5. -0.03*cvar_risk_2(0.97,matrix_prior_scenarios)

6. Box_of_Variables:  upperbounds =1,
   lowerbounds =-1

7. Solver:  VAN, precision = 6, stages = 6
```

Here we give a brief description of the presented meta-code. We boldfaced the important parts of the code. The keyword `minimize` tells a solver that (38) is a minimization problem. To define an objective function the keyword `Objective` is used. The objective function in (38) is defined in lines 3 through 5. The quadratic part of objective in the line 3 is defined by the keyword `quadratic` and the data matrix located in the file `matrix_quadratic_matrix.txt`. The $(1-\alpha_L)\mathrm{CVaR}_{\alpha_L}(\cdot) - (1-\alpha_U)\mathrm{CVaR}_{\alpha_U}(\cdot)$ part of objective in the lines 4,5 is defined by the keywords `cvar_risk_1`, `cvar_risk_2` and data matrix located in the file `matrix_prior_scenarios.txt`. The coefficients $\alpha_L$ and $\alpha_U$ are set to 0.3 and 0.97, accordingly. Also coefficients $C$ in the lines 3 and 4 is set to 1.

Notice that the meta code for (12) is the the keyword `var_risk_1` is replaced by the keyword `cvar_risk_1`.

## References

Carlo Acerbi. Spectral measures of risk: a coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] `http://www.ise.ufl.edu/uryasev/research/testproblems/advanced-statistics/case-study-nu-support-vector-machine-based-on-tail-risk-measures/`

David J Crisp Christopher JC Burges. A geometric interpretation of $\nu$-svm classifiers. *Advances in Neural Information Processing Systems 12*, 12:244–250, 2000.

Chih-Chung Chang and Chih-Jen Lin. Training $\nu$-support vector regression: theory and algorithms. *Neural Computation*, 14(8):1959–1977, 2002.

Olivier Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.

Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*, pages 201–208. ACM, 2006.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press, 2000.

Darrell Duffie and Jun Pan. An overview of value at risk. *The Journal of derivatives*, 4(3): 7–49, 1997.

Jun-ya Gotoh and Akiko Takeda. A linear classification model based on conditional geometric score. *Pacific Journal of Optimization*, 1:277–296, 2005.

Jun-ya Gotoh, Akiko Takeda, and Rei Yamamoto. Interaction between financial risk measures and machine learning methods. *Computational Management Science*, pages 1–38, 2013a.

Jun-ya Gotoh, Akiko Takeda, and Rei Yamamoto. Interaction between financial risk measures and machine learning methods. *Computational Management Science*, 2013b.

R Horst and Ng V Thoai. Dc programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

Reiner Horst and Hoang Tuy. *Global optimization: Deterministic approaches.* Springer, 2003.

Reiner Horst, Panos M Pardalos, and Nguyen Van Thoai. *Introduction to global optimization.* Kluwer Academic Pub, 2000.

Philippe Jorion. *Value at risk: the new benchmark for controlling market risk*, volume 2. McGraw-Hill New York, 1997.

K-R Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, 12(2):181–201, 2001.

Fernando Pérez-Cruz, Jason Weston, DJL Herrmann, and B Scholkopf. Extension of the nu-svm range for classification. *NATO SCIENCE SERIES SUB SERIES III COMPUTER AND SYSTEMS SCIENCES*, 190:179–196, 2003.

R Tyrrell Rockafellar and Stan Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1):33–53, 2013.

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

Akiko Takeda and Masashi Sugiyama. $\nu$-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on Machine learning*, pages 1056–1063. ACM, 2008.

Pham Dinh Tao et al. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1-4):23–46, 2005.

Peter Tsyurmasto, Michael Zabarankin, and Stan Uryasev. Value-at-risk support vector machine: Stability to outliers. 2013.

Vladimir Vapnik. *The nature of statistical learning theory*. springer, 1999.