Ann Oper Res (2017) 249:301–328
DOI 10.1007/s10479-016-2326-x

Support vector machines based on convex risk functions and general norms

Jun-ya Gotoh[1] · **Stan Uryasev**[2]

Published online: 30 September 2016

Abstract This paper studies unified formulations of support vector machines (SVMs) for binary classification on the basis of convex analysis, especially, convex risk functions theory, which is recently developed in the context of financial optimization. Using the notions of convex empirical risk and convex regularizer, a pair of primal and dual formulations of the SVMs are described in a general manner. With the generalized formulations, we discuss reasonable choices for the empirical risk and the regularizer on the basis of the risk function's properties, which are well-known in the financial context. In particular, we use the properties of the risk function's dual representations to derive multiple interpretations. We provide two perspectives on robust optimization modeling, enhancing the known facts: (1) the primal formulation can be viewed as a robust empirical risk minimization; (2) the dual formulation is compatible with the distributionally robust modeling.

Keywords Support vector machine · SVM · Binary classification · Convex risk function · Duality · Norm · Robust optimization

1 Introduction

Background. During the last two decades, *support vector machines (SVMs)* have become a popular methodology for binary classification and a number of modified formulations have been derived. To find a decision function that can predict the class labels of unseen data,

✉ Jun-ya Gotoh
jgoto@indsys.chuo-u.ac.jp

Stan Uryasev
uryasev@ufl.edu

[1] Department of Industrial and Systems Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

[2] Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall, Gainesville, FL 32611, USA

every SVM solves in general a bi-objective minimization which is defined with a set of given data samples, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$, and is usually referred to as the *regularized empirical risk minimization (ERM)* or the *structural risk minimization* (see e.g., Christopher 1998). Typically, the decision function is represented by a discriminant hyperplane, $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{w}^\top \boldsymbol{x} = b\}$, which separates $\mathbb{R}^n$ into two half-spaces, each corresponding to a class label. In order to obtain it, the regularized ERM minimizes the sum of an Empirical Risk and a Regularizer:

$$\underbrace{\mathcal{F}(\boldsymbol{L}(\boldsymbol{w}, b))}_{\text{[EmpiricalRisk]}} + \underbrace{\gamma(\boldsymbol{w}),}_{\text{[Regularizer]}}$$

over $(\boldsymbol{w}, b)$. Here, $\boldsymbol{L}(\boldsymbol{w}, b)$ is a vector in $\mathbb{R}^m$, representing a degree of misclassification over the given data samples with respect to the hyperplane, $\mathcal{F}$ is a function of $\boldsymbol{L}$, gauging the aversion to the vector and referred to as *risk function* in this paper, and γ is a function regularizing $\boldsymbol{w}$. Intuitively, minimization of Empirical Risk, the first term of the objective, seeks a hyperplane which would have smaller in-sample misclassification, while the Regularizer, the second term, prevents the hyperplane from overfitting to the samples.

This simple principle allows for a large freedom in the choice of Empirical Risk and Regularizer. Despite the generality, only several choices are popular in the literature. For example, the C-SVM (Cortes and Vapnik 1995), the most prevailing formulation, employs as Empirical Risk the *hinge loss*: $\mathcal{F}(\boldsymbol{L}) = \frac{C}{m}\sum_{i=1}^m \max\{L_i + 1, 0\}$ or $\mathcal{F}(\boldsymbol{L}) = \frac{C}{m}\sum_{i=1}^m (\max\{L_i + 1, 0\})^2$, where $C > 0$ is a user-defined constant.

As for the Regularizer, the use of the square of the ℓ_2-norm (or the Euclidean norm) of the normal vector, e.g., $\gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, is dominant. Although the use of the ℓ_2-norm naturally leads to the so-called *kernel trick* (see e.g., Christopher 1998), other norms are viable alternatives. For example, the ℓ_1-norm is popular since its use leads to a sparse solution. Besides, the use of any norm can be justified along the lines of geometric margin maximization by supposing that its dual norm is employed in gauging the distance of a data sample to the discriminant hyperplane (see e.g., Mangasarian 1999; Pedroso and Murata 2001).

Motivation and proposed scheme. The primary purpose of this paper is to seek reasonable forms of the regularized ERM on the basis of a generalized formulation. One of the motivations for this comes out of the pursuit of a tractable SVM formulation in which parametrized families of polyhedral norms, recently studied by Gotoh and Uryasev (2016), are employed as regularizers. A merit of the use of those parametrized families of norms is in tuning regularizers. The ℓ_p-norm family is used for the tuning of regularizers (e.g., Kloft et al. 2011). In consideration of current status of algorithm studies and solver software, the use of the ℓ_p-norm with $p \neq 1, 2,$ or ∞ is, however, not advantageous over the ℓ_1-, the ℓ_2-, or the ℓ_∞-norm in that its nonlinearity with respect to the parameter p may prevent an efficient parametric optimization. In contrast, the new families can be associated with linear program (LP) and employ an efficient parametric optimization (with respect to alternative parameters), and thus are more advantageous for the tuning of regularizers.

However, the introduction of such new norms requires a prudent approach. Indeed, the form of the regularizer and/or the choice of the empirical risk function can affect the validity of the optimization problem and/or the meaning of the resulting classifier. To address this issue, we reexamine the basic formulation of SVMs. Additionally, to retain the tractability of the popular SVMs, we limit our attention to the case where both empirical risk functions and regularizers are convex.

Method. The development of the formulations is based on *convex analysis*, especially the *Fenchel duality* (e.g., Rockafellar 1970). Although convex analysis is not typically used in machine learning literature, it is becoming popular (e.g., Rifkin and Lippert 2007; Kloft et al. 2011). As Rifkin and Lippert (2007) claim, the use of Fenchel duality is advantageous in developing duality theorems and establishing optimality conditions because we can derive most of the results in a standard setting just by applying established patterns of function operations. Thus, by treating both norms and empirical risk functions in a more general, we enjoy these advantages as well.

A novelty is the employment of the *convex risk function theory* established in mathematical finance (e.g., Föllmer and Schied 2002) and stochastic optimization (e.g., Ruszczyński and Shapiro 2005, 2006; Rockafellar and Uryasev 2013). The linkage to the risk function theory yields new perspectives on the SVM formulations. For example, the so-called ν-property of the ν-SVM (Schölkopf et al. 2000) can be analyzed via the connection to the conditional value-at-risk (CVaR), a popular risk measure in financial context (Rockafellar and Uryasev 2000, 2002) since the ν-SVM virtually employs a CVaR (Gotoh and Takeda 2005; Takeda and Sugiyama 2008).

More notable benefits of our approach are in relations to the three functions' properties: Monotonicity, Translation Invariance, and Positive Homogeneity, which are all often referred to in the context of financial risk management. We draw several insights on the regularized ERM, especially in relation to robust optimization modeling and the geometric and probabilistic interpretations.

Perspectives on robust optimization modeling. SVMs are often viewed as a data-driven optimization based on i.i.d. data samples, and are likely to be vulnerable to some perturbation of the data samples or deviation from the i.i.d. assumption. To cope with such a situation, the idea of optimizing the worst case, known as robust optimization, is a popular choice. In this paper we provide two insights on the relation to the robust modeling. First, considering a worst-case perturbation of the given data samples, we derive an interpretation of the regularized ERM as a robust optimization of the (unregularized) ERM, which is parallel to what Xu et al. (2009a) show by focusing on the use of the hinge loss. The perspective that the regularizer would make the ERM robust is enhanced in a broad way with the help of the risk function theory. Second, we demonstrate that with some risk functions $\mathcal{F}$, our framework can straightforwardly treat another type of robust optimization modeling, called distributionally robust optimization. This type of robust optimization assumes uncertainty in the probability measures whereas the first robust modeling approach assumes uncertainty in the support (or observed values of samples). We show that with a class of risk functions the distributionally robustified formulations can be established within convex optimization.

Novelty in the context of machine learning. The term "convex risk" itself is not new in the context of machine learning. Indeed, the empirical risk of the separable form $\mathcal{F}(\boldsymbol{L}) = \frac{1}{m}\sum_{i=1}^{m} v(L_i)$ has been discussed where v is a convex function on $\mathbb{R}$ (e.g., Christmann and Steinwart 2004; Zhang 2004; Bartlett et al. 2006; Rifkin and Lippert 2007; Kloft et al. 2011). This formulation, however, does not include some important convex risk functions that appear in machine learning methods. Indeed, the ν-SVM corresponds to an inseparable risk function $\mathcal{F}(\boldsymbol{L}) = \min_\rho\{-\rho\nu + \frac{1}{m}\sum_{i=1}^{m}\max\{L_i+\rho, 0\}\}$, where $\nu \in (0, 1]$. If we expand the coverage beyond the separable functions, we can treat for instance the log-sum-exp (or entropic) function, i.e., $\mathcal{F}(\boldsymbol{L}) = \ln\sum_{i=1}^{m}\exp(L_i)$, without removing the 'ln'-operator. Needless to say, the minimization of $\ln\sum_{i=1}^{m}\exp(L_i)$ is equivalent to the minimization of $\sum_{i=1}^{m}\exp(L_i)$. However, even a difference based on such a monotonic transformation may result in a different consequence because of the consideration of regularizer, which we will discuss in Sect. 3.1.

In addition, a class of inseparable risk functions include several existing formulations as special cases recently studied in Kanamori et al. (2013). Their paper shows the convergence of the obtained classifier to a classifier which attains the smallest expected misprediction. This indicates that the generalized formulation developed in the current paper is, at least, partly justified in a statistical way.

Moreover, with the help of the risk function theory, we show that dual formulations of such inseparable risk functions can be connected to a geometric or probabilistic interpretation. The geometric interpretation extends the existing papers (e.g., Crisp and Burges 2000; Bennett and Bredensteiner 2000; Takeda et al. 2013; Kanamori et al. 2013), while the probabilistic interpretation relates to the minimization of the φ-divergence (or, originally, f-divergence) (Csiszár 1967), as will be shown in Sect. 5. Interestingly, the probabilistic interpretation is further connected to the distributionally robust extension.

Further merits in practice. The presentation of general formulations, defined with only elementary operations, fits a recent trend of optimization software packages. Indeed, various convex functions are available as built-in functions in some software packages [e.g., PSG (American Optimal Decisions, Inc. 2009) and CVX (Grant and Boyd 2012)]. With these software packages, users can easily customize SVMs. Additionally, such a presentation may potentially fit recently developed algorithms for, e.g., nonsmooth and/or stochastic optimization. For example, the recent development of ℓ_1-minimization algorithms, such as the Fast Iterative Shrinkage Thresholding Algorithm (Beck and Teboulle 2009), suggests to directly handle the subgradient of the ℓ_1-norm.

Difference from relevant studies. Let us mention several related papers. Following Gotoh and Takeda (2005), Takeda and Sugiyama (2008) which derive ν-SVM (Schölkopf et al. 2000) and Eν-SVM (Perez-Cruz et al. 2003) as CVaR minimizations, Gotoh et al. (2014) extends CVaR to the coherent risk measures (Artzner et al. 1999), a subclass of convex risk functions, while preserving the nonconvexity in the formulation of the preceding studies. Also, Tsyurmasto et al. (2013) explore positively homogeneous risk functions, which are not necessarily convex. In contrast to the above papers, our paper disregards the nonconvexity, while generalizing the class of risk functions.

Xu et al. (2009b) mention the axioms of risk functions in relation to their robust optimization formulation. Takeda et al. (2013) propose a unified view using the presentation with uncertainty sets, although that paper does not relate to risk functions. A recent paper of Kanamori et al. (2013) studies a duality correspondence between empirical risk functions and uncertainty sets, sharing part of their formulation with ours. An advantage of the current paper over the above ones is a larger capability in a more systematic presentation on the basis of the theory of convex risk functions.

As for the use of general norms, numerous papers deal with non-ℓ_2-norms. Among such, Zhou et al. (2002) present a couple of formulations, which are shared with ours, and show some generalization bounds for them. However, to our best knowledge, all the existing papers focus on the ℓ_p-norms and only the ℓ_1- and the ℓ_∞-norms are employed for LP formulations of SVMs. In contrast, we employ other LP-representable norms such as the CVaR norm and the deltoidal norm (Gotoh and Uryasev 2016), which both include the ℓ_1- and the ℓ_∞-norms as special limiting cases.

Structure of the paper. The structure of this paper is as follows. Section 2 poses a general formulation of SVMs and explains how it includes existing ones. In particular, Sects. 2.2 and 2.3 introduce risk functions and regularizers, respectively, which are or can be used for SVMs. Section 3 examines the form of regularizer from two perspectives; Section 3.1 discusses the incompatibility of homogeneous empirical risk and regularizer, while Sect. 3.2 reveals a condition under which the regularized ERM can be viewed as a robust ERM. Section 4 derives duality theorem and optimality condition as well as dual formulation. Section 5 is devoted to interpretation of the dual formulation and some relations to distributionally robust optimization. Section 6 concludes the paper. Proofs of propositions are given in "Appendix".

To downsize the manuscript, this paper focuses on presentation of the general formulation. Readers who are interested in the proximity between the use of the non-ℓ_2-norm and that of the ℓ_2-norm or in remarks on the kernel trick and some numerical examples illustrating those theoretical results are referred to Gotoh and Uryasev (2013), the discussion version of this paper.

Notations. A vector in $\mathbb{R}^n$ is denoted in boldface and is written as a column vector in the inner products. In particular, $\mathbf{1}$ and $\mathbf{0}$ are the column vectors with all components equal to 1 and 0, respectively. Matrices are denoted also by boldface. In particular, we denote by $\operatorname{diag}(\boldsymbol{x})$ the square matrix whose diagonal elements are given by $\boldsymbol{x}$ and off-diagonal elements are all 0. The superscript '$\top$' denotes the transpose of vectors and matrices (e.g., $\boldsymbol{x}^\top = (x_1, \ldots, x_n)$). The inequality $\boldsymbol{x} \geq \boldsymbol{y}$ denotes $x_i \geq y_i, i = 1, \ldots, m$, and $\mathbb{R}^n_+ := [0, +\infty)^n = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \geq \mathbf{0}\}$; we denote by Π^m the unit simplex in $\mathbb{R}^m$, i.e., $\Pi^m := \{\boldsymbol{p} \in \mathbb{R}^m : \mathbf{1}^\top \boldsymbol{p} = 1, \boldsymbol{p} \geq \mathbf{0}\}$. For a set C, its relative interior is denoted by $\operatorname{ri}(C)$. We denote the ℓ_2-, the ℓ_1-, and the ℓ_∞-norms by $\|\cdot\|_2$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$, respectively. The notation $\|\cdot\|$ is reserved for any norm in $\mathbb{R}^n$. $(x)_+ := \max\{x, 0\}$. δ_C denotes the (0-∞) indicator function of a set $C \subset \mathbb{R}^n$, i.e., $\delta_C(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in C$; $+\infty$ otherwise. With a little abuse of the notation, we sometimes denote by $\delta_{c(\cdot)}$ the indicator function of a condition $c(\cdot)$, i.e., $\delta_{c(\cdot)}(\boldsymbol{x}) = 0$ if $c(\boldsymbol{x})$ is true; $+\infty$ otherwise. As a convention inspired by MATLAB, we extensively apply a function on $\mathbb{R}$ to a vector in $\mathbb{R}^m$. For example, with a function v on $\mathbb{R}$, we define $v(\boldsymbol{L}) := (v(L_1), \ldots, v(L_m))^\top$. Besides, we employ the notations '$./$' and '$.^k$' for component-wise division of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and power, respectively, i.e., $\boldsymbol{x}./\boldsymbol{y} = (x_1/y_1, \ldots, x_n/y_n)^\top$ and $\boldsymbol{x}.^k = (x_1^k, \ldots, x_n^k)^\top$.

2 A general primal formulation of SVMs

This section introduces a general primal formulation of the binary classification and explains its motivations.

2.1 Loss, risk function, and regularized ERM

Suppose that a data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ is given, where $\boldsymbol{x}_i \in \mathbb{R}^n$ denotes the attributes of sample i and $y_i \in \{\pm 1\}$ denotes its binary label, which represents the class that sample i belongs to, $i = 1, \ldots, m$. Then the (binary) classification problem is formulated as the problem of finding a decision function $d : \mathbb{R}^n \to \{\pm 1\}$ defined as

$$d(\boldsymbol{x}) := \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x} - b) = \begin{cases} +1, & \text{if } \boldsymbol{w}^\top \boldsymbol{x} \geq b, \\ -1, & \text{if } \boldsymbol{w}^\top \boldsymbol{x} < b, \end{cases} \tag{1}$$

for predicting binary labels of unseen samples, $\boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_\ell$. Note that it is equivalent to find a vector $\boldsymbol{w} \neq \mathbf{0}$ and a scalar b.

To formulate the problem as an optimization, we quantify the misspecification of the sample labels by the decision function d by using some function $\mathcal{F}$ on some *loss* $\boldsymbol{L}$ associated with $(\boldsymbol{w}, b)$.

Let us first introduce the loss. Since a sample $\boldsymbol{x}_i$ can be considered misclassified if its *margin*, denoted by $y_i(\boldsymbol{x}_i^\top \boldsymbol{w} - b)$, is negative it is natural to define $\boldsymbol{L}$ as

$$\boldsymbol{L} = -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}b) \leftrightarrow L_i = -y_i(\boldsymbol{x}_i^\top \boldsymbol{w} - b),\ i = 1, \ldots, m, \tag{2}$$

where $\boldsymbol{Y} := \text{diag}(y_1, \ldots, y_m)$ and $\boldsymbol{X} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)^\top$. More precisely, the loss represents the degree of misclassification so that $L_i > 0$ means that the i-th sample is misclassified, while $L_i < 0$ implies correct classification.

Alternatively, with a kernel function k on $\mathbb{R} \times \mathbb{R}$ and $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^m$, the loss can be associated with a nonlinear discriminant surface defined by $\{\boldsymbol{x} \in \mathbb{R}^n : \sum_{h=1}^m y_h k(\boldsymbol{x}_h, \boldsymbol{x}_i)\alpha_h = b\}$. Denoting the kernel matrix by $\boldsymbol{K} = (y_i y_j k(x_i, x_j))_{i,j=1,\ldots,m} \in \mathbb{R}^{m \times m}$, the corresponding loss can be defined by

$$\boldsymbol{L} = -(\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}b) \leftrightarrow L_i = -y_i \left(\sum_{h=1}^m y_h k(\boldsymbol{x}_h, \boldsymbol{x})\alpha_h - b \right),\ i = 1, \ldots, m. \tag{3}$$

It is noteworthy that with a matrix $\boldsymbol{G}$ and a vector $\boldsymbol{w}$, both (2) and (3) have the form

$$\boldsymbol{L} = -(\boldsymbol{G}^\top \boldsymbol{w} - \boldsymbol{y}b), \tag{4}$$

which is linear with respect to $(\boldsymbol{w}, b)$. Accordingly, we hereafter suppose that the loss $\boldsymbol{L}$ has the linear form (4).

We are now in a position to formally introduce the empirical risk.

Let $\mathcal{F}$ be a function on $\mathbb{R}^m$ that can take '$+\infty$' and is proper and l.s.c. (lower semicontinuous):

- $\mathcal{F}$ is *proper* if $\mathcal{F}(\boldsymbol{L}) > -\infty$ for all $\boldsymbol{L} \in \mathbb{R}^m$ and $\text{dom}\,\mathcal{F} \neq \emptyset$, where '$\text{dom}\,\mathcal{F}$' denotes the effective domain of $\mathcal{F}$, i.e., $\text{dom}\,\mathcal{F} := \{\boldsymbol{L} \in \mathbb{R}^m : \mathcal{F}(\boldsymbol{L}) < +\infty\}$.
- $\mathcal{F}$ is *l.s.c.* if $\mathcal{F}(\boldsymbol{L}) \leq \liminf\limits_{i \to \infty} \mathcal{F}(\boldsymbol{L}_i)$ for any $\boldsymbol{L} \in \mathbb{R}^m$ and any sequence $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots \in \mathbb{R}^m$ converging to $\boldsymbol{L}$.

The empirical risk is then defined as $\mathcal{F}(\boldsymbol{L})$, and we call the function $\mathcal{F}$ a risk function. By construction, $\mathcal{F}(\boldsymbol{L})$ represents the undesirability of the loss defined above, i.e., less is better. Here we would like to note that the usage of the words 'loss' and 'risk' are slightly different from the usual convention of machine learning literature.

With a function $\gamma : \mathbb{R}^n \to [0, \infty]$, we consider a general SVM for binary classification of the following regularized ERM form:

$$p^\star := \inf_{\boldsymbol{w}, b}\ \mathcal{F}(-(\boldsymbol{G}^\top \boldsymbol{w} - \boldsymbol{y}b)) + \gamma(\boldsymbol{w}), \tag{5}$$

which is a minimization of a function consisting of two terms, as sketched in the introduction.

For example, the C-SVM (Cortes and Vapnik 1995) for the binary classification is formulated with the following (convex) quadratic programming (QP) problem:

$$\begin{aligned} \bar{p}^\star := \underset{\boldsymbol{w}, b, \boldsymbol{z}}{\text{minimize}}\ & \tfrac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \tfrac{C}{m}\boldsymbol{1}^\top \boldsymbol{z} \\ \text{subject to}\ & \boldsymbol{z} \geq -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}b) + \boldsymbol{1},\ \boldsymbol{z} \geq \boldsymbol{0}, \end{aligned} \tag{6}$$

where $C > 0$ is a user-defined parameter. This can be equivalently presented as

$$\underset{\boldsymbol{w}, b}{\text{minimize}}\ \underbrace{\frac{C}{m}\boldsymbol{1}^\top(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}b) + \boldsymbol{1})_+}_{\text{EmpiricalRisk}} + \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2}_{\text{Regularizer}},$$

which corresponds to (5) with $\mathcal{F}(\boldsymbol{L}) = (C/m)\mathbf{1}^\top(\boldsymbol{L}+\mathbf{1})_+$, $\boldsymbol{G} = \boldsymbol{X}^\top\boldsymbol{Y}$, and $\gamma(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. On the other hand, the ν-*SVM* (Schölkopf et al. 2000) solves another QP:

$$\begin{aligned} \tilde{p}^\star := \underset{\boldsymbol{w},b,\rho,\boldsymbol{z}}{\text{minimize}} \ & \tfrac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} - \rho + \tfrac{1}{m\nu}\mathbf{1}^\top\boldsymbol{z} \\ \text{subject to} \ & \boldsymbol{z} \geq -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}b) + \mathbf{1}\rho,\ \boldsymbol{z} \geq \mathbf{0}, \end{aligned} \tag{7}$$

where $\nu \in (0, 1]$ is a user-defined parameter. Similarly to the C-SVM, the QP (7) can be viewed as a regularized ERM:

$$\underset{\boldsymbol{w},b}{\text{minimize}} \ \underbrace{\min_{\rho}\Big\{-\rho + \frac{1}{\nu m}\mathbf{1}^\top((-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}b) + \rho\mathbf{1})_+)\Big\}}_{\text{EmpiricalRisk}} + \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2}_{\text{Regularizer}}.$$

Note that the risk function here is considered as the CVaR (see the list below for its definition), while, to our knowledge, no specific name was given in the SVM context, while the risk function of (6) is known as *hinge loss*.

Considering the tractability in optimization and duality, we assume convexity of $\mathcal{F}$ and γ throughout the paper:

- $\mathcal{F}$ is *convex* if $(1-\tau)\mathcal{F}(\boldsymbol{L}) + \tau\mathcal{F}(\boldsymbol{L}') \geq \mathcal{F}((1-\tau)\boldsymbol{L} + \tau\boldsymbol{L}')$ for all $\boldsymbol{L}, \boldsymbol{L}' \in \mathbb{R}^m$, $\tau \in (0, 1)$.

In the remainder of this section, let us see other existing alternatives of the convex risk function $\mathcal{F}$ (Sect. 2.2) and the convex regularizer γ (Sect. 2.3), which are covered by the generalized formulation (5).

2.2 Convex risk functions and their basic properties

Below, we give some examples of convex risk functions in binary classification.

- $\mathcal{F}(\boldsymbol{L}) = \text{Hinge1}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := t\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}+\mathbf{1})_+$: Hinge loss-based;
- $\mathcal{F}(\boldsymbol{L}) = \text{Hinge2}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{t}{2}\mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L}+\mathbf{1})_+{}^{\cdot 2})$: squared Hinge loss-based;
- $\mathcal{F}(\boldsymbol{L}) = \text{LSSVM}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{t}{2}\mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L}+\mathbf{1})^{\cdot 2})$: Least Square SVM-based;
- $\mathcal{F}(\boldsymbol{L}) = \text{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) := \min_c\{c + \frac{1}{1-\alpha}\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L} - c\mathbf{1})_+\}$: CVaR;
- $\mathcal{F}(\boldsymbol{L}) = \text{LR}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\ln(\mathbf{1} + \exp(\frac{1}{t}\boldsymbol{L})))$: Logistic Regression-based;
- $\mathcal{F}(\boldsymbol{L}) = \text{LSE}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{1}{t}\ln\mathbb{E}_{\boldsymbol{p}}(\exp(t\boldsymbol{L}))$: Log-Sum-Exp (or entropic);
- $\mathcal{F}(\boldsymbol{L}) = \text{SE}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{1}{t}\mathbb{E}_{\boldsymbol{p}}(\exp(t\boldsymbol{L}))$: Sum-Exp;
- $\mathcal{F}(\boldsymbol{L}) = \text{MV}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + \frac{t}{2}\mathbb{V}_{\boldsymbol{p}}(\boldsymbol{L})$: Mean-Variance,

where $t > 0$ and $\alpha \in [0, 1)$ are user-defined parameters, and $\mathbb{E}_{\boldsymbol{p}}(\cdot)$ denotes the mathematical expectation under a probability measure $\boldsymbol{p}$, i.e., $\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{x}) := \boldsymbol{p}^\top\boldsymbol{x}$. See Proof in "Appendix" of Gotoh and Uryasev (2013) for the additional explanation of CVaR (Proof of Proposition 2 section in "Appendix") and a list of the other risk functions which have potential to be useful (Proof of Theorem 2 section in "Appendix").

Here we emphasize that in addition to $\text{CVaR}_{(\alpha,\boldsymbol{p})}$, which appears in the ν-SVM, the Log-Sum-Exp risk function $\text{LSE}_{(1,\boldsymbol{p})}$, which appears in AdaBoost (Freund and Schapire 1997), is another notable inseparable-form risk function.

In addition to convexity, the following three properties are frequently considered in the context of financial risk management (e.g., Artzner et al. 1999).

- $\mathcal{F}$ is *monotonic* if $\mathcal{F}(\boldsymbol{L}) \geq \mathcal{F}(\boldsymbol{L}')$ for all $\boldsymbol{L}, \boldsymbol{L}' \in \mathbb{R}^m$ such that $\boldsymbol{L} \geq \boldsymbol{L}'$.
- $\mathcal{F}$ is *translation invariant* if $\mathcal{F}(\boldsymbol{L} + \tau\mathbf{1}) = \mathcal{F}(\boldsymbol{L}) + \tau$ for all $\tau \in \mathbb{R}$, $\boldsymbol{L} \in \mathbb{R}^m$.

- $\mathcal{F}$ is *positively homogeneous* if $\mathcal{F}(\tau \boldsymbol{L}) = \tau \mathcal{F}(\boldsymbol{L})$ for all $\tau > 0,\ \boldsymbol{L} \in \mathbb{R}^m$.

In particular, a proper l.s.c. convex risk function satisfying the above three properties is said to be *coherent*.

- $\mathcal{F}$ is *coherent* if it is a proper l.s.c. convex risk function satisfying monotonicity, translation invariance and positive homogeneity.

While the above properties make sense in financial risk management (see Artzner et al. 1999, for interpretations of these properties in the financial context), we need to examine the rationale of their role in the context of SVMs, i.e., with which properties $\mathcal{F}$ can be reasonable as a risk function for SVMs.

Among the three properties, the monotonicity seems to be less arguable since it requires that a larger misclassification, L_i, should be more penalized in the ERM. On the other hand, there seems to be no strong motivation for the other two properties at this point, unless they lead to tractable optimization problems. However, as will be shown in Sects. 3 to 5, these properties play crucial roles in the interpretation of the primal or dual formulation and in the validity of the combination of risk function and regularizer. Because of those facts, this paper considers the above properties. While the term "coherence" may be confusing, we simply use it to emphasize commonality with the financial risk function theory. We do not use this term to insist that "coherence" implies a more legitimate choice of function in the context of the classification task.

For later reference, we observe the following facts.

Proposition 1 *For a function $\mathcal{V}$ or $\mathbb{R}^m$, let us define another function on $\mathbb{R}^m$ of the form:*

$$\mathcal{F}(\boldsymbol{L}) = \inf_c \{c + \mathcal{V}(\boldsymbol{L} - c\mathbf{1})\}. \tag{8}$$

Then, $\mathcal{F}$ is convex if $\mathcal{V}$ is convex; $\mathcal{F}$ is monotonic if $\mathcal{V}$ is monotonic; $\mathcal{F}$ is translation invariant for any $\mathcal{V}$; $\mathcal{F}$ is positively homogeneous if $\mathcal{V}$ is positively homogeneous.

This proposition is a minor modification of Quadrangle Theorem of Rockafellar and Uryasev (2013) which restricts $\mathcal{F}$ to be convex (the above proposition does not make this assumption).

We should notice that we can view the formula (8) as an operation for making any function translation invariant.

In particular, we will refer to a special case of (8) having the form

$$\mathcal{F}(\boldsymbol{L}) = \mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L}) := \inf_c \{c + \mathbb{E}_{\boldsymbol{p}}(v(\boldsymbol{L} - c\mathbf{1}))\} \equiv \inf_c \left\{ c + \sum_{i=1}^{m} p_i v(L_i - c) \right\}, \tag{9}$$

where $v : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is a proper l.s.c. convex function. As an easy extension of Proposition 1, we can confirm the following properties.

Corollary 1 *Function* (9) *is translation invariant for any v.* (9) *is convex if v is convex;* (9) *is monotonic if v is monotonic;* (9) *is positively homogeneous if v is positively homogeneous; if v satisfies $v(z) \geq z + B$ for any z and a constant B,* (9) *is proper.*

This corollary is also a minor modification of Expectation Theorem of Rockafellar and Uryasev (2013), where v is assumed to be convex and satisfies the condition of the last statement above with $B = 0$.

We would emphasize that $\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}$ and $\mathrm{LSE}_{(1,\boldsymbol{p})}$ can be represented in the form of (9) with $v(z) = \frac{1}{1-\alpha}(z)_+$ and $\exp(z) - 1$, respectively, with each satisfying the condition.

In addition, via the formula (9), no-translation invariant functions such as Hinge1, Hinge2, and LSSVM can be transformed into translation invariant ones. For example, by employing

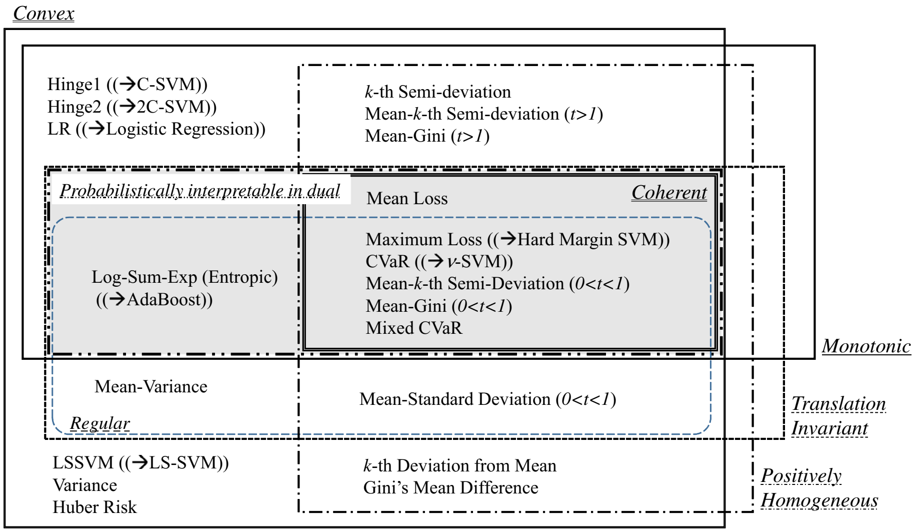

Fig. 1 Classification of convex risk functions and corresponding regularized ERMs

$v(z) = (1+z)_+/t$ in (9), Hinge1$_{(t,\boldsymbol{p})}$ is transformed to a translation invariant risk function Hinge1$^{\mathrm{OCE}}_{(t,\boldsymbol{p})} = \inf_c\{c+\frac{1}{t}\boldsymbol{p}^\top[((1-c)\mathbf{1}+\boldsymbol{L})_+]\}$. Note that this is equal to $1+\mathrm{CVaR}_{(1-t,\boldsymbol{p})}(\boldsymbol{L})$. Namely, CVaR can be considered as Hinge1 transformed by (9).

Such transformed functions are shown to be related to the uncertainty set-based representation of SVMs. Indeed, Kanamori et al. (2013) consider an SVM formulation which employs the risk function of the form (9) with $\boldsymbol{p} = (2/m)\mathbf{1}$. More precisely, one of their examples called Truncated Quadratic Loss corresponds to Hinge2 transformed by (9). An extension based on the above propositions will be discussed in Sect. 5.

Figure 1 illustrates relationships of some risk functions on the basis of their properties, indicating relations to several existing regularized ERMs. For the risk functions which are not mentioned in this paper, see Proof of Theorem 2 in "Appendix" of Gotoh and Uryasev (2013).

In addition to the properties described in the text, 'Regular' refers to l.s.c. convex risk functions $\mathcal{F}$ satisfying

- $\mathcal{F}(\tau\mathbf{1}) = \tau$ for all $\tau \in \mathbb{R}$, [consistency]
- $\mathcal{F}(\boldsymbol{L}) > \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L})$ for all $\boldsymbol{L}$ which does not satisfy $\boldsymbol{L} = \tau\mathbf{1}$ for some $\tau \in \mathbb{R}$. [aversity]

Rockafellar and Uryasev 2013 show that with a (proper) l.s.c. convex $v : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that $v(0) = 0$ and $v(x) > x$ when $x \neq 0$, the risk function defined in (9) is a regular measure of risk. In order to avoid unnecessary mix-up, we do not use 'risk measure' term in this paper, but use 'risk function' term instead. Besides, in Sect. 5, we show that the dual formulations of a set of monotonic translation-invariant risk functions which is shaded in the figure are interpreted as an optimization over probability distributions.

2.3 Regularizers with general norms and classification of SVMs

Following Rifkin and Lippert (2007), the regularizer γ is assumed to have the following properties.

$$\gamma : \mathbb{R}^n \to [0, +\infty] \text{ is an l.s.c. convex function such that } \gamma(\mathbf{0}) = 0. \tag{10}$$

In particular, we below consider the case where the regularizer $\gamma(\boldsymbol{w})$ is associated with an arbitrary norm as follows.

$$\gamma(\boldsymbol{w}) = \iota(\|\boldsymbol{w}\|),$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$, and $\iota : [0, +\infty) \to [0, +\infty]$ is non-decreasing and convex. In the following, we pay special attention to the following three regularizers.

$$\begin{aligned} &\text{(a)} \quad \gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ &\text{(b)} \quad \gamma(\boldsymbol{w}) = \|\boldsymbol{w}\| \\ &\text{(c)} \quad \gamma(\boldsymbol{w}) = \delta_{\|\cdot\|\le 1}(\boldsymbol{w}) \end{aligned}$$

where $\delta_{\|\cdot\|\le 1}$ denotes the indicator function defined as

$$\delta_{\|\cdot\|\le 1}(\boldsymbol{w}) = \begin{cases} 0, & \|\boldsymbol{w}\| \le 1, \\ \infty, & \|\boldsymbol{w}\| > 1. \end{cases}$$

Note that $\|\cdot\|$ denotes an arbitrary norm, while $\|\cdot\|_2$ denotes the ℓ_2-norm.

The cases (a) and (b) are categorized as the Tikhonov regularization, where the norms appear in the objective of the primal formulation, while the case (c) is categorized as the Ivanov regularization, where the norm appears in the constraint of the formulation, i.e., $\|\boldsymbol{w}\| \le 1$. These two styles often bring the same result (see e.g., Proposition 12 of Kloft et al. 2011). However, we have to pay attention to the difference because such equivalence depends on the risk function employed, which will be discussed in Sect. 3.1.

Despite several restrictions on the forms of the loss $\boldsymbol{L}$, the risk function $\mathcal{F}$, and the regularizer γ, the general formulation (5) covers a variety of optimization problem formulations for binary classification, as follows.

- 1-C-SVM (6): $\mathcal{F} = \text{Hinge1}_{(t, \frac{1}{m})}$, $\gamma(\cdot) = \frac{1}{2}\|\cdot\|_2^2$;
- 2-C-SVM: $\mathcal{F} = \text{Hinge2}_{(t, \frac{1}{m})}$, $\gamma(\cdot) = \frac{1}{2}\|\cdot\|_2^2$;
- ν-SVM (7): $\mathcal{F} = \text{CVaR}_{(1-\nu, \frac{1}{m})}$, $\gamma(\cdot) = \frac{1}{2}\|\cdot\|_2^2$;
- ℓ_1-regularized logistic regression (e.g., Koh et al. (2007)): $\mathcal{F} = \text{LR}_{(1, \frac{1}{m})}$, $\gamma(\cdot) = t\|\cdot\|_1$;
- AdaBoost (Freund and Schapire 1997), $\mathcal{F} = \text{LSE}_{(1, \frac{1}{m})}$, $\gamma = \delta_{\|\cdot\|_1\le 1} + \delta_{\mathbb{R}^n_+}$;
- LPBoost (Rätsch et al. 2000) $\mathcal{F} = \text{CVaR}_{(1-\nu, \frac{1}{m})}$, $\gamma = \delta_{\|\cdot\|_1\le 1} + \delta_{\mathbb{R}^n_+}$;
- LS-SVM (Suykens and Vandewalle 1999) $\mathcal{F} = \text{LSSVM}_{(t, \frac{1}{m})}$, $\gamma(\cdot) = \frac{1}{2}\|\cdot\|_2^2$,

where $\gamma = \delta_{\|\cdot\|_1\le 1} + \delta_{\mathbb{R}^n_+}$ corresponds to a regularizer, explicitly given by

$$(\delta_{\|\cdot\|_1\le 1} + \delta_{\mathbb{R}^n_+})(\boldsymbol{w}) = \begin{cases} 0, & \text{if } \mathbf{1}^\top \boldsymbol{w} \le 1,\ \boldsymbol{w} \ge \mathbf{0}, \\ +\infty, & \text{otherwise.} \end{cases}$$

With such a large possibility of the risk functions and the regularizers, the first research question we will consider is formulated as follows. What properties should $\mathcal{F}$ and γ have for the regularized ERM (5) to be reasonable or interpretable?

3 Insights on the general regularizer

To answer the question given at the end of the previous section, we first consider the regularized ERM (5) and later its dual formulation. In particular, this section draws two insights on the primal formulation (5): (1) an incompatible choice of the empirical risk and the regularizer; (2) a perspective as a robust optimization.

3.1 General formulations with non-ℓ_2-norm regularizers

Let us start with the following simple, but suggestive fact.

Proposition 2 *Suppose that both regularizer γ and risk function $\mathcal{F}$ are positively homogeneous. Then the primal* (5) *either attains the optimal objective value* 0*, where the solution* $(\boldsymbol{w}^\star, b^\star)$ *such that* $\boldsymbol{w}^\star = \boldsymbol{0}$ *is an optimal solution, or results in an unbounded solution such that* $p^\star = -\infty$.

See Proof of Proposition 2 in "Appendix" for the proof.

The above proposition shows a situation where (5) would be meaningless, having no optimal solution or resulting in a trivial solution satisfying $\boldsymbol{w} = \boldsymbol{0}$. Even if an optimization algorithm returns an optimal solution with $\boldsymbol{w} \neq \boldsymbol{0}$, such a solution is considered to be fragile since the all-zero solution is another optimal solution. Accordingly, the combination of a positively homogeneous function $\mathcal{F}$, such as CVaR, and a regularizer of the form $\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\|$ is not adequate for the classification problem. On the other hand, with a non-homogeneous ι, the regularizer given in the form $\gamma(\boldsymbol{w}) = \iota(\|\boldsymbol{w}\|)$ makes sense. For example, the case where $\gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, corresponding to $\iota(z) = \frac{1}{2}z^2$, works for the ν-SVM, which employs CVaR. (See also Remark 1 below.)

Thus, we may apply such a non-homogeneous ι to non-ℓ_2-norm, such as the ℓ_1- or the ℓ_∞-norm, so as not to make the Tikhonov regularizer positively homogeneous. However, such a strategy leads to a non-linear formulation and may reduce the advantage of using polyhedral norms. Therefore, below we consider the case of the Ivanov regularization, i.e., $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\|\leq 1}(\boldsymbol{w})$.

The Tikhonov and Ivanov regularizations are often considered identical. However, as the above observation indicates, a careful treatment is required. A notion of 'equivalence' only holds if a meaningful optimal solution is attained.

Remark 1 Tsyurmasto et al. (2013) consider the case where $\mathcal{F}$ is positively homogeneous (not necessarily convex) and the ℓ_2-norm is employed for the regularizer. They show that under some mild conditions, the following formulations are equivalent. By 'equivalent' we mean that the formulations provide the same (set of) classifiers.

$$\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{w}\|_2 \leq E, \end{array} \tag{11}$$

$$\underset{\boldsymbol{w},b}{\text{minimize}} \; C \cdot \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) + \frac{1}{2}\|\boldsymbol{w}\|_2^2, \tag{12}$$

$$\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} & \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) \leq -D. \end{array} \tag{13}$$

Here E, C, and D are positive constants, although the equivalence is independent of E, C, and D and therefore we can set $E = C = D = 1$. This is a virtue of the positive homogeneity of the risk function.

On the other hand, without positive homogeneity, the above independence does not hold. For example, with $\mathcal{F} = \text{Hinge1}_{(1,\boldsymbol{1}C/m)}$, which is not homogeneous, (12) is equal to the C-SVM (6). However, the equivalence to (11) or to (13) depends on E or D, respectively.

3.2 Interpretation of regularizers based on robust optimization modeling

Xu et al. (2009a) show that a regularized minimization of Hinge1 can be viewed as a robust optimization modeling. They suppose that the given data set $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\}$ suffers from some perturbation of the form $\{(\boldsymbol{x}_1 - \boldsymbol{\delta}_1, y_1), \dots, (\boldsymbol{x}_m - \boldsymbol{\delta}_m, y_m)\}$ with some $(\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_m)$ belonging to

$$\mathcal{T} := \left\{ (\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_m) : \sum_{i=1}^{m} \|\boldsymbol{\delta}_i\| \leq C \right\},$$

where $C > 0$ is a parameter deciding the size of the set and $\|\cdot\|$ is a norm. Under this uncertainty, they consider to minimize the worst-case ERM with Hinge1. Namely, they consider to minimize $\max_{\boldsymbol{\Delta}\in\mathcal{T}} \text{Hinge1}_{(m,\boldsymbol{1}/m)}(-\boldsymbol{Y}\{(\boldsymbol{X} - \boldsymbol{\Delta})\boldsymbol{w} - \boldsymbol{1}b\})$, where $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_m)^\top$.

Theorem 1 (Xu et al. 2009a) *Suppose that there is no decision function (1) correctly mapping all given samples, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$, to their labels, $y_1, \dots, y_m$, i.e., $\nexists(\boldsymbol{w}, b) \in (\mathbb{R}^n \setminus \{\boldsymbol{0}\}) \times \mathbb{R}$, $y_i = d(\boldsymbol{x}_i)$, $i = 1, \dots, m$. Then the following two optimization problems over $(\boldsymbol{w}, b)$ are equivalent.*

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \max_{(\boldsymbol{\delta}_1,\dots,\boldsymbol{\delta}_m)\in\mathcal{T}} \sum_{i=1}^{m} (1 - y_i\{\boldsymbol{w}^\top(\boldsymbol{x}_i - \boldsymbol{\delta}_i) - b\})_+,$$

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad C\|\boldsymbol{w}\|^\circ + \sum_{i=1}^{m} (1 - y_i(\boldsymbol{w}^\top\boldsymbol{x}_i - b))_+,$$

where $\|\cdot\|^\circ$ is the dual norm of $\|\cdot\|$, i.e., another norm defined by $\|\boldsymbol{w}\|^\circ := \max_{\boldsymbol{x}}\{\boldsymbol{x}^\top\boldsymbol{w} : \|\boldsymbol{x}\| \leq 1\}$.

In this subsection, we derive similar results for the case of monotonic and translation invariant risk functions. Note that Hinge1, which Xu et al. (2009a) employed, is not a case considered here since it is not translation invariant. In place of $\mathcal{T}$, we consider the following uncertainty.

$$\mathcal{S} := \{(\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_m) : \|\boldsymbol{\delta}_i\| \leq C, i = 1, \dots, m\}.$$

Note that $\mathcal{S}$ is called the box uncertainty in Xu et al. (2009a), and $\mathcal{S} \supset \mathcal{T}$ holds for the same C.

Theorem 2 *Let the function $\mathcal{F}$ be monotonic and translation invariant as well as proper, l.s.c. and convex. Then, for any $(\boldsymbol{w}, b)$, we have*

$$\max_{\boldsymbol{\Delta}\in\mathcal{S}} \mathcal{F}(-\boldsymbol{Y}\{(\boldsymbol{X} - \boldsymbol{\Delta})\boldsymbol{w} - \boldsymbol{1}b\}) = C\|\boldsymbol{w}\|^\circ + \mathcal{F}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}b)), \tag{14}$$

where $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_m)^\top \in \mathbb{R}^{m\times n}$.

See Proof of Theorem 2 in "Appendix" for the proof.

Theorem 2 shows that the Tikhonov regularization $\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\|$ can be interpreted as a consequence of the robustification of the (non-regularized) ERM not only for Hinge1, but also for a variety of other risk functions satisfying monotonicity and translation invariance. In particular, it is interesting to note that both approaches induce the same Tikhonov-type regularizer if the same norm is employed for defining the uncertainty sets, $\mathcal{S}$ and $\mathcal{T}$. In this sense, the use of a norm in the Tikhonov regularization is just viewed as a consequence of the choice of uncertainty set. Moreover, Theorem 2 derives the same regularizer in a simpler way under a class of risk functions and larger uncertainty set.

On the other hand, there is a fact which is noteworthy in the light of Proposition 2. We see that if $\mathcal{F}$ is coherent (i.e., positively homogeneous in addition to the two properties supposed in Theorem 2), the unconstrained minimization of the worst-case empirical risk (14) is not adequate for the classification task. A similar generalization is also considered by Livni et al. (2012) on the basis of a probabilistic interpretation. However, their formulation cannot deal with positively homogeneous risk functions due to this reason.

To make sense of the employment of (14) as the empirical risk term when $\mathcal{F}$ is coherent, the addition of an Ivanov regularizer or non-homogeneous Tikhonov regularizer is required. With this view, a formulation

$$\frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\|\boldsymbol{w}\| + \mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}b)),$$

for example, makes sense in the light of Proposition 2 and can be viewed as a robust version of the ν-SVM. It is noteworthy that a composite regularizer of the form $\frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\|\boldsymbol{w}\|_1$ is known as the elastic net-type regularizer in the machine learning community and used also for SVMs.

4 Dual formulation and Fenchel duality

Dual SVM formulations are frequently considered. For example, the dual problem to the C-SVM (6) is given by another QP:

$$\begin{array}{rl} \bar{d}^\star := \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\frac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{\lambda} + \mathbf{1}^\top\boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \mathbf{0} \le \boldsymbol{\lambda} \le C\mathbf{1}/m. \end{array} \tag{15}$$

Strong duality between (6) and (15), i.e., $\bar{p}^\star = \bar{d}^\star$, holds under a mild condition. More importantly, with the optimality condition, which will be described later in Theorem 4, we have

$$\boldsymbol{w}^\star = \boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{\lambda}^\star, \tag{16}$$

where $\boldsymbol{w}^\star$ is an optimal solution to (6) and $\boldsymbol{\lambda}^\star$ is an optimal solution to (15). This equation leads to the so-called *representer theorem*, which provides a building block for the kernel-based nonlinear classification (e.g., Burges 1998). In fact, putting the condition (16) into (1), the decision function (1) can be rewritten with the optimal dual variables $\boldsymbol{\lambda}^\star$ as $d(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{x}^\top\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{\lambda}^\star - b^\star)$. See e.g., Chen et al. (2005) for the calculation of $b^\star$ on the basis of the dual solution $\boldsymbol{\lambda}^\star$.

On the other hand, the dual formulation of the ν-SVM (7) is given by another QP:

$$\begin{array}{rl} \tilde{d}^\star := \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\frac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \mathbf{1}^\top\boldsymbol{\lambda} = 1,\ \mathbf{0} \le \boldsymbol{\lambda} \le \mathbf{1}/(\nu m). \end{array} \tag{17}$$

Let $(\boldsymbol{w}^\star, b^\star, \rho^\star, \boldsymbol{z}^\star)$ and $\boldsymbol{\lambda}^\star$ be optimal solutions to (7) and (17), respectively. Similarly to the C-SVM, under a mild condition, we have the strong duality between (7) and (17), i.e., $\tilde{p}^\star = \tilde{d}^\star$, and the parallelism (16) between $\boldsymbol{w}^\star$ and $\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{\lambda}^\star$, again. Likewise, we can obtain a decision function on the basis of a dual solution to (17).

Note that the dual formulations and the optimality conditions for the ℓ_2-regularized SVMs, such as (15) and (17), are often derived via the Lagrangian duality theory (see e.g., Burges,

Table 1 Conjugates of convex risk functions examples

$\mathcal{F}^*(\boldsymbol{\lambda})$
$\text{Hinge1}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = -\mathbf{1}^\top\boldsymbol{\lambda} + \delta_{[\mathbf{0},t\boldsymbol{p}](\lambda_i)}(\boldsymbol{\lambda}) \equiv \sum_{i=1}^m(-\lambda_i + \delta_{[0,tp_i]})$
$\text{Hinge2}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \frac{1}{2t}\boldsymbol{\lambda}^{.2}./\boldsymbol{p} - \mathbf{1}^\top\boldsymbol{\lambda} + \delta_{\mathbb{R}^m_+}(\boldsymbol{\lambda}) \equiv \sum_{i=1}^m(\frac{\lambda_i^2}{2tp_i} - \lambda_i + \delta_{\mathbb{R}_+}(\lambda_i))$
$\text{LSSVM}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \frac{1}{2t}\boldsymbol{\lambda}^{.2}./\boldsymbol{p} - \mathbf{1}^\top\boldsymbol{\lambda} \equiv \sum_{i=1}^m(\frac{\lambda_i^2}{2tp_i} - \lambda_i)$
$\text{CVaR}^*_{(\alpha,\boldsymbol{p})}(\boldsymbol{\lambda}) = \delta_{\mathcal{Q}_{\text{CVaR}(\alpha,\boldsymbol{p})}}(\boldsymbol{\lambda})$ with $\mathcal{Q}_{\text{CVaR}(\alpha,\boldsymbol{p})}$ defined in (26)
$\text{LR}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \boldsymbol{p}^\top\{(t\boldsymbol{\lambda}./\boldsymbol{p})\ln(t\boldsymbol{\lambda}./\boldsymbol{p}) + (\mathbf{1} - t\boldsymbol{\lambda}./\boldsymbol{p})\ln(\mathbf{1} - t\boldsymbol{\lambda}./\boldsymbol{p})\} + \delta_{[\mathbf{0},\boldsymbol{p}/t]}(\boldsymbol{\lambda})$ $\equiv \sum_{i=1}^m\left[p_i\left\{(\frac{t\lambda_i}{p_i})\ln(\frac{t\lambda_i}{p_i}) + (1 - \frac{t\lambda_i}{p_i})\ln(1 - \frac{t\lambda_i}{p_i})\right\} + \delta_{[0,\frac{p_i}{t}]}(\lambda_i)\right] := \text{bitEnt}_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$
$\text{LSE}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \frac{1}{t}\boldsymbol{\lambda}^\top\ln(\boldsymbol{\lambda}./\boldsymbol{p}) + \delta_{\Pi^m}(\boldsymbol{\lambda}) \equiv \frac{1}{t}\sum_{i=1}^m\lambda_i\ln\frac{\lambda_i}{p_i} + \delta_{\Pi^m}(\boldsymbol{\lambda}) := \text{KL}_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$
$\text{SE}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \frac{1}{t}\boldsymbol{\lambda}^\top(\ln(\boldsymbol{\lambda}./\boldsymbol{p}) - \mathbf{1}) + \delta_{\mathbb{R}^m_+}(\boldsymbol{\lambda}) \equiv \sum_{i=1}^m\{\frac{1}{t}\lambda_i(\ln\frac{\lambda_i}{p_i} - 1) + \delta_{\mathbb{R}_+}(\lambda_i)\}$
$\text{MV}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \frac{1}{2t}\boldsymbol{p}^\top\{(\boldsymbol{\lambda}./\boldsymbol{p})^{.2} - \mathbf{1}\} + \delta_C(\boldsymbol{\lambda}) \equiv \frac{1}{2t}\sum_{i=1}^m p_i\{(\frac{\lambda_i}{p_i})^2 - 1\} + \delta_C(\boldsymbol{\lambda}) =: \chi^2_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$ with $C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \mathbf{1}^\top\boldsymbol{\lambda} = 1\}$

1998). In contrast, as shown below, the use of the Fenchel duality theory benefits to derive dual formulations and optimality condition under any combination of $\mathcal{F}$ and γ.

4.1 Formulations and duality of convex risk function-based SVMs

The dual problem to the general SVM (5) is derived as

$$d^\star := \sup_{\boldsymbol{\lambda}} -\gamma^*(\boldsymbol{G}\boldsymbol{\lambda}) - \mathcal{F}^*(\boldsymbol{\lambda}) - \delta_{\boldsymbol{y}^\top(\cdot)=0}(\boldsymbol{\lambda}), \tag{18}$$

where γ^* and $\mathcal{F}^*$ denote the conjugate functions of γ and $\mathcal{F}$, respectively, namely,

$$\gamma^*(\boldsymbol{w}) := \sup_{\boldsymbol{s}}\{\boldsymbol{w}^\top\boldsymbol{s} - \gamma(\boldsymbol{s})\}, \quad \mathcal{F}^*(\boldsymbol{\lambda}) := \sup_{\boldsymbol{L}}\{\boldsymbol{\lambda}^\top\boldsymbol{L} - \mathcal{F}(\boldsymbol{L})\}. \tag{19}$$

Since $\mathcal{F}$ and γ are proper l.s.c. convex functions, both γ^* and $\mathcal{F}^*$ are proper, l.s.c. and convex (e.g., Section 12 of Rockafellar 1970), and the dual (18) is a convex optimization problem. Table 1 lists conjugates of the aforementioned risk functions. As for the regularizers (a) to (c) introduced in Sect. 2.3, we have the following conjugate relations.

$$\begin{array}{lll}
\text{(a)} \;\; \gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 & \leftrightarrow & \gamma^*(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2, \\
\text{(b)} \;\; \gamma(\boldsymbol{w}) = \|\boldsymbol{w}\| & \leftrightarrow & \gamma^*(\boldsymbol{w}) = \delta_{\|\cdot\|^\circ \le 1}(\boldsymbol{w}), \\
\text{(c)} \;\; \gamma(\boldsymbol{w}) = \delta_{\|\cdot\| \le 1}(\boldsymbol{w}) & \leftrightarrow & \gamma^*(\boldsymbol{w}) = \|\boldsymbol{w}\|^\circ.
\end{array}$$

Note especially that the conjugate of γ also becomes a regularizer, i.e., satisfies the condition (10), if γ is a regularizer. In this sense, the squared ℓ_2-regularizer (a) is self-dual, while the Tikhonov regularizer (b) and the Ivanov regularizer (c) are dual to each other.

Obviously the known dual formulations such as (15) and (17) can be readily derived just by applying the above established patterns of conjugation. For example, with $\text{Hinge1}^*_{(t,\boldsymbol{p})}$ and $\gamma^*(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, we can reach the dual (15) of the C-SVM (6).

Given a pair of the primal and dual formulations, (5) and (18), respectively, we can describe the weak and strong duality theorems, as follows.

Proposition 3 *(Weak duality) The weak duality holds between* (5) *and* (18)*, i.e., we have* $p^\star \geq d^\star$.

Theorem 3 *(Strong duality) The strong duality holds between* (5) *and* (18)*, i.e., we have* $p^\star = d^\star$*, if either of the following conditions is satisfied:*

(a) There exists a $(\boldsymbol{w}, b)$ *such that* $\boldsymbol{w} \in \operatorname{ri}(\operatorname{dom}\gamma)$ *and* $-\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b \in \operatorname{ri}(\operatorname{dom}\mathcal{F})$.
(b) There exists a $\boldsymbol{\lambda} \in \operatorname{ri}(\operatorname{dom}\mathcal{F}^*)$ *such that* $\boldsymbol{y}^\top\boldsymbol{\lambda} = 0$.

Under (a), the supremum in (18) *is attained at some* $\boldsymbol{\lambda}$*, while under (b), the infimum in* (5) *is attained at some* $(\boldsymbol{w}, b)$*. In addition, if* $\mathcal{F}$ *(or equivalently,* $\mathcal{F}^*$*) is polyhedral, "*ri*" can be omitted.*

Proposition 3 is straightforward from the Fenchel's inequality (see Sections 12 and 31 of Rockafellar 1970). Theorem 3 can be obtained from the Fenchel-Rockafellar duality theorem. See Proofs of Theorem 3 and the modification of the condition (a) for the Ivanov regularization in "Appendix" for the details.

4.2 Duality correspondence for the case with the Ivanov regularizers

Regarding the incompatibility between the Tikhonov regularization of the form $\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\|$ and the positively homogeneous risk function $\mathcal{F}$ (Proposition 2), let us consider the SVM with the Ivanov regularization. The primal (5) and the dual (18) then become

$$\begin{array}{rl} p^\star := \underset{\boldsymbol{w},b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{w}\| \leq 1, \end{array} \tag{20}$$

and

$$\begin{array}{rl} d^\star := \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0, \end{array} \tag{21}$$

respectively. We denote by $(\mathcal{F}, \|\cdot\|)$ the pair of the primal and dual formulations (20) and (21) for an SVM.

For the Ivanov regularization case, the condition (a) of Theorem 3 can be a little more specific.

(a) There exists a $(\boldsymbol{w}, b)$ such that $\|\boldsymbol{w}\| < 1$ and $-\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b \in \operatorname{ri}(\operatorname{dom}\mathcal{F})$.

With the help of the Fenchel duality, the optimality conditions can be derived in a similar manner. When the Tikhonov-type ℓ_2-regularization, $\gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, is employed, the condition (16) is derived. As for the Ivanov regularization case, the condition is derived as follows.

Theorem 4 (Optimality condition) *Suppose* (20) *and* (21)*. In order that* $(\boldsymbol{w}^\star, b^\star)$ *and* $\boldsymbol{\lambda}^\star$ *be vectors such that*

$$\mathcal{F}(-(\boldsymbol{G}^\top\boldsymbol{w}^\star - \boldsymbol{y}b^\star)) + \delta_{\|\cdot\|\leq 1}(\boldsymbol{w}^\star) = -\|\boldsymbol{G}\boldsymbol{\lambda}^\star\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}^\star) - \delta_{\boldsymbol{y}^\top(\cdot)=0}(\boldsymbol{\lambda}^\star),$$

it is necessary and sufficient that $(\boldsymbol{w}^\star, b^\star)$ *and* $\boldsymbol{\lambda}^\star$ *satisfy the conditions:*

$$\boldsymbol{G}\boldsymbol{\lambda}^\star \in \mathcal{N}(\boldsymbol{w}^\star), \quad \|\boldsymbol{w}^\star\| \leq 1, \quad \boldsymbol{y}^\top\boldsymbol{\lambda}^\star = 0, \quad -\boldsymbol{G}^\top\boldsymbol{w}^\star + \boldsymbol{y}b^\star \in \partial\mathcal{F}^*(\boldsymbol{\lambda}^\star), \tag{22}$$

where $\mathcal{N}(\boldsymbol{w}^\star) := \{\boldsymbol{u} : \boldsymbol{u}^\top\boldsymbol{w}^\star = \|\boldsymbol{u}\|^\circ\}$*, and* $\partial\mathcal{F}^*(\boldsymbol{\lambda}^\star)$ *is the subdifferential of* $\mathcal{F}^*$ *at* $\boldsymbol{\lambda}^\star$*, i.e.,* $\partial\mathcal{F}^*(\boldsymbol{\lambda}^\star) := \{\boldsymbol{L} : \mathcal{F}^*(\boldsymbol{L}) \geq \mathcal{F}^*(\boldsymbol{\lambda}^\star) + \boldsymbol{L}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star),$ *for all* $\boldsymbol{\lambda}\}$.

This theorem is also straightforward from Theorem 31.3 of Rockafellar (1970). See the appendix for the detailed correspondence.

Note that the first and the second conditions in (22) can be rewritten by

$$\boldsymbol{w}^\star \in \arg\max_{\boldsymbol{w}}\{(\boldsymbol{\lambda}^\star)^\top\boldsymbol{G}^\top\boldsymbol{w} : \|\boldsymbol{w}\| \leq 1\}.$$

In particular, if we employ the ℓ_2-norm Ivanov regularization, this condition implies

$$\boldsymbol{w}^\star = \frac{\boldsymbol{G}\boldsymbol{\lambda}^\star}{\|\boldsymbol{G}\boldsymbol{\lambda}^\star\|_2}. \tag{23}$$

This condition, which claims a parallelism between $\boldsymbol{w}^\star$ and $\boldsymbol{\lambda}^\star$, corresponds to the one given in (16). Accordingly, as long as the ℓ_2-norm is employed, the two regularization results in the same decision function.

Contrarily, if we employ a non-ℓ_2-norm, we have to pay attention to the deviation from the parallelism (23). See Section 6 of Gotoh and Uryasev (2013) for discussion of the proximity of the parallelism for a parameterized class of LP-representable norms.

Example 1 Employing $\mathcal{F} = \text{LSE}_{(t,\boldsymbol{p})}$, we have an SVM $(\text{LSE}_{(t,\boldsymbol{p})}, \|\cdot\|)$, where its dual is obtained as

$$\begin{array}{llll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \text{KL}_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) \equiv & \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \frac{1}{t}\sum_{i=1}^{m}\lambda_i \ln\frac{\lambda_i}{p_i} \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0, & \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{1}^\top\boldsymbol{\lambda} = 1,\ \boldsymbol{\lambda} \geq \boldsymbol{0}. \end{array} \tag{24}$$

Let us consider the optimality condition (22) for $(\text{LSE}_{(t,\boldsymbol{p})}, \|\cdot\|)$ in (24). Noting that at any $\boldsymbol{\lambda} \in \text{ri}(\Pi^m)$, the function $\mathcal{F}^*(\boldsymbol{\lambda}) = \text{KL}_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \frac{1}{t}\boldsymbol{\lambda}^\top\ln(\boldsymbol{\lambda}./\boldsymbol{p}) + \delta_{\text{ri}(\Pi^m)}(\boldsymbol{\lambda})$ has subdifferential $\partial\mathcal{F}^*(\boldsymbol{\lambda}) = \{\nabla\frac{1}{t}\boldsymbol{\lambda}^\top\ln(\boldsymbol{\lambda}./\boldsymbol{p})(\boldsymbol{\lambda}) + k\boldsymbol{1} : k \in \mathbb{R}\}$, the optimality condition is then explicitly given by

$$\begin{array}{l} (\boldsymbol{\lambda}^\star)^\top\boldsymbol{G}^\top\boldsymbol{w}^\star = \|\boldsymbol{G}\boldsymbol{\lambda}^\star\|^\circ, \quad \|\boldsymbol{w}^\star\| \leq 1, \quad \boldsymbol{y}^\top\boldsymbol{\lambda}^\star = 0, \\ -\boldsymbol{G}^\top\boldsymbol{w}^\star + \boldsymbol{y}b^\star = \frac{1}{t}(\ln\boldsymbol{\lambda}^\star./\boldsymbol{p} + \boldsymbol{1}) + \boldsymbol{1}k^\star, \quad \boldsymbol{1}^\top\boldsymbol{\lambda}^\star = 1 \quad \boldsymbol{\lambda}^\star > \boldsymbol{0}. \end{array}$$

Furthermore, consider the situation where the ℓ_2-norm is employed in $(\text{LSE}_{(t,\boldsymbol{p})}, \|\cdot\|_2)$, and there exists a solution $\boldsymbol{\lambda}^\star > \boldsymbol{0}$ such that $\|\boldsymbol{G}\boldsymbol{\lambda}^\star\|_2 > 0$, then we can find an optimal solution by solving a system of $n + m + 2$ equalities:

$$\boldsymbol{w}^\star = \frac{\boldsymbol{G}\boldsymbol{\lambda}^\star}{\|\boldsymbol{G}\boldsymbol{\lambda}^\star\|_2}, \quad \boldsymbol{y}^\top\boldsymbol{\lambda}^\star = 0, \quad -\boldsymbol{G}^\top\boldsymbol{w}^\star + \boldsymbol{y}b^\star = \frac{1}{t}(\ln\boldsymbol{\lambda}^\star./\boldsymbol{p} + \boldsymbol{1}) + \boldsymbol{1}k^\star, \quad \boldsymbol{1}^\top\boldsymbol{\lambda}^\star = 1,$$

and the optimal decision function is given by $d(\boldsymbol{x}) = \text{sign}(\frac{\boldsymbol{x}^\top\boldsymbol{G}\boldsymbol{\lambda}^\star}{\|\boldsymbol{G}\boldsymbol{\lambda}^\star\|_2} - b^\star)$. □

5 Perspectives on the dual formulation from various viewpoints

In this section, we demonstrate connections between the dual formulation (18) and the existing papers. We base our arguments on the correspondences between duality and the properties of the risk function $\mathcal{F}$.

5.1 Correspondence between risk function properties and dual formulations

By using the conjugate of $\mathcal{F}$, monotonicity, translation invariance, and positive homogeneity can be characterized in a dual manner as follows.

Theorem 5 (Ruszczyński and Shapiro 2006) *Suppose that $\mathcal{F}$ is l.s.c., proper, and convex, then we have*

1. *$\mathcal{F}$ is monotonic if and only if $\mathrm{dom}\,\mathcal{F}^*$ is in the nonnegative orthant;*
2. *$\mathcal{F}$ is translation invariant if and only if $\forall \boldsymbol{\lambda} \in \mathrm{dom}\,\mathcal{F}^*$, $\mathbf{1}^\top \boldsymbol{\lambda} = 1$;*
3. *$\mathcal{F}$ is positively homogeneous if and only if it can be represented in the form*
$$\mathcal{F}(\boldsymbol{L}) = \sup_{\boldsymbol{\lambda}} \{\boldsymbol{L}^\top \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathrm{dom}\,\mathcal{F}^*\}, \tag{25}$$
or equivalently, $\mathcal{F}^*(\boldsymbol{L}) = \delta_{\mathcal{Q}}(\boldsymbol{L})$ for a convex set $\mathcal{Q}$ in $\mathbb{R}^m$.

Note that the first and second statements of Theorem 5 imply the following expressions, respectively.

1. $\mathcal{F}$ is monotonic if and only if $\mathcal{F}(\boldsymbol{L}) = \sup_{\boldsymbol{\lambda}}\{\boldsymbol{L}^\top\boldsymbol{\lambda} - \mathcal{F}^*(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \geq \mathbf{0}\}$;
2. $\mathcal{F}$ is translation invariant if and only if $\mathcal{F}(\boldsymbol{L}) = \sup_{\boldsymbol{\lambda}}\{\boldsymbol{L}^\top\boldsymbol{\lambda} - \mathcal{F}^*(\boldsymbol{\lambda}) : \mathbf{1}^\top\boldsymbol{\lambda} = 1\}$;
3. $\mathcal{F}$ is monotonic and translation invariant if and only if $\mathcal{F}(\boldsymbol{L}) = \sup_{\boldsymbol{\lambda}}\{\boldsymbol{L}^\top\boldsymbol{\lambda} - \mathcal{F}^*(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \Pi^m\}$.

From Theorem 5, we see that $\mathrm{dom}\,\mathcal{F}^*$ plays an important role in characterizing risk functions. Let us denote this effective domain by $\mathcal{Q}_{\mathcal{F}}$ and call it *risk envelope*, i.e., $\mathcal{Q}_{\mathcal{F}} = \mathrm{dom}\,\mathcal{F}^*$.[1] In particular, by combining with Theorem 5, any coherent risk function can be characterized by a set of probability measures.

Corollary 2 (Artzner et al. 1999) *Any coherent risk function $\mathcal{F}$, we have $\mathcal{Q}_{\mathcal{F}} \subset \Pi^m$. On the other hand, for any set $\mathcal{Q} \subset \Pi^m$, the risk function defined as $\mathcal{F}(\boldsymbol{L}) := \sup\{\boldsymbol{L}^\top\boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathcal{Q}\} = \sup\{\boldsymbol{L}^\top\boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathrm{conv}(\mathcal{Q})\} = \max\{\boldsymbol{L}^\top\boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathrm{cl}(\mathrm{conv}(\mathcal{Q}))\}$ is coherent, where $\mathrm{conv}(\mathcal{Q})$ denotes the convex hull of a set $\mathcal{Q}$ and $\mathrm{cl}(\mathcal{Q})$ denotes the closure of a set $\mathcal{Q}$.*

For example, CVaR is coherent and can be represented with the risk envelope
$$\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})} := \left\{\boldsymbol{q} \in \Pi^m : \boldsymbol{q} \leq \boldsymbol{p}/(1-\alpha)\right\}, \tag{26}$$
i.e., $\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \max_{\boldsymbol{q}}\{\mathbb{E}_{\boldsymbol{q}}(\boldsymbol{L}) : \boldsymbol{q} \in \mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}\}$.

Based on Theorem 5, we can associate the constraints of dual formulations with the properties of the risk functions $\mathcal{F}$ employed in the primal formulation (5).

Proposition 4 1. *If $\mathcal{F}$ is monotonic, the dual problem (18) can be represented as*
$$\sup_{\boldsymbol{\lambda}} \; -\gamma^*(\boldsymbol{G}\boldsymbol{\lambda}) - \mathcal{F}^*(\boldsymbol{\lambda}) - \delta_C(\boldsymbol{\lambda}) \text{ with } C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0, \boldsymbol{\lambda} \geq \mathbf{0}\};$$

2. *If $\mathcal{F}$ is translation invariant, the dual problem (18) can be represented as*
$$\sup_{\boldsymbol{\lambda}} \; -\gamma^*(\boldsymbol{G}\boldsymbol{\lambda}) - \mathcal{F}^*(\boldsymbol{\lambda}) - \delta_C(\boldsymbol{\lambda}) \text{ with } C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0, \mathbf{1}^\top\boldsymbol{\lambda} = 1\};$$

[1] This terminology is a bit different from that in Rockafellar and Uryasev (2013). However, we use the same words for simplicity since there is a one-to-one correspondence between them.

3. *If $\mathcal{F}$ is positively homogeneous, the dual problem* (18) *can be represented as*

$$\sup_{\boldsymbol{\lambda}} -\gamma^*(\boldsymbol{G}\boldsymbol{\lambda}) - \delta_C(\boldsymbol{\lambda}) \text{ with } C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0\} \cap \mathcal{Q}_{\mathcal{F}}.$$

Example 2 If we limit our attention to positively homogeneous risk functions and the Ivanov regularization, the dual formulation (21) can be simplified with the help of its risk envelope $\mathcal{Q}_{\mathcal{F}}$:

$$\begin{aligned} p^\star := \underset{\boldsymbol{w},b}{\text{minimize}}\ & \sup_{\boldsymbol{q}}\{-\boldsymbol{q}^\top(\boldsymbol{G}^\top\boldsymbol{w} - \boldsymbol{y}b) : \boldsymbol{q} \in \mathcal{Q}_{\mathcal{F}}\} \\ \text{subject to}\ & \|\boldsymbol{w}\| \leq 1, \end{aligned} \tag{27}$$

and

$$\begin{aligned} d^\star := \underset{\boldsymbol{\lambda}}{\text{maximize}}\ & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ \\ \text{subject to}\ & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0, \quad \boldsymbol{\lambda} \in \mathcal{Q}_{\mathcal{F}}. \end{aligned} \tag{28}$$

There is a symmetric dual correspondence between the primal (27) and the dual (28). Precisely, the primal (27) has a norm constraint with $\|\cdot\|$ while its dual norm $\|\cdot\|^\circ$ appears in the objective of the dual (28); the positively homogeneous convex risk function $\mathcal{F}$ in the primal's objective corresponds to its risk envelope $\mathcal{Q}_{\mathcal{F}}$ in the dual's constraint.

Correspondingly, the condition (b) of Theorem 3 can be replaced with

(b) There exists a $\boldsymbol{\lambda}$ such that $\boldsymbol{y}^\top\boldsymbol{\lambda} = 0$ and $\boldsymbol{\lambda} \in \text{ri}\mathcal{Q}_{\mathcal{F}}$,

and the optimality condition (22) of Theorem 4 can be rewritten as follows:

$$\boldsymbol{G}\boldsymbol{\lambda}^\star \in \mathcal{N}(\boldsymbol{w}^\star),\ \|\boldsymbol{w}^\star\| \leq 1,\ \boldsymbol{y}^\top\boldsymbol{\lambda}^\star = 0,\ -\boldsymbol{G}^\top\boldsymbol{w}^\star + \boldsymbol{y}b^\star \in N_{\mathcal{Q}_{\mathcal{F}}}(\boldsymbol{\lambda}^\star),\ \boldsymbol{\lambda}^\star \in \mathcal{Q}_{\mathcal{F}}, \tag{29}$$

where $N_{\mathcal{Q}_{\mathcal{F}}}(\boldsymbol{\lambda}^\star)$ denotes the normal cone to the set $\mathcal{Q}$ at a point $\boldsymbol{\lambda}^\star \in \mathcal{Q}_{\mathcal{F}}$, i.e., $N_{\mathcal{Q}_{\mathcal{F}}}(\boldsymbol{\lambda}^\star) := \{\boldsymbol{L} \in \mathbb{R}^m : \boldsymbol{L}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star) \leq 0,\ \text{for all } \boldsymbol{\lambda} \in \mathcal{Q}_{\mathcal{F}}\}$.

The change in the final part of (29) comes from the fact that the subdifferential of the indicator function of a non-empty convex set is given by the normal cone to it (see p. 215 of Rockafellar 1970, for the details of the subdifferential of the indicator function). □

Remark 2 The primal formulation (27) with $\mathcal{Q}_{\mathcal{F}} \subset \Pi^m$ is a convex relaxation of the formulation developed by Gotoh et al. (2014), where the negative geometric margin and the coherent risk function are employed as the loss and the risk measure, respectively. On the other hand, the dual formulation is not mentioned in their paper since theirs includes some nonconvexity. □

If monotonicity and translation invariance are simultaneously supposed on $\mathcal{F}$, the dual variable, $\boldsymbol{\lambda}$, can be considered as a probability measure, i.e., $\boldsymbol{\lambda} \in \Pi^m$.

Corollary 3 *If $\mathcal{F}$ is monotonic and translation invariant, the dual problem* (18) *can be rewritten by*

$$\sup_{\boldsymbol{\lambda}} -\gamma^*(\boldsymbol{G}\boldsymbol{\lambda}) - \mathcal{F}^*(\boldsymbol{\lambda}) - \delta_C(\boldsymbol{\lambda}) \text{ with } C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0\} \cap \Pi^m.$$

Furthermore, if $\mathcal{F}$ is coherent, the third statement of Proposition 4 is valid with $\mathcal{Q}_{\mathcal{F}}$ such that $\mathcal{Q}_{\mathcal{F}} \subset \Pi^m$.

To deepen this probabilistic view, let us introduce the *φ-divergence* (Csiszár 1967; Ben-Tal and Teboulle 2007). Let $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be an l.s.c. convex function satisfying $\varphi(1) = 0$. With such φ, the φ-divergence of $\boldsymbol{q} \in \mathbb{R}^m$ relative to $\boldsymbol{p} \in \Pi_+^m$ is defined by

$$\mathcal{I}_\varphi(\boldsymbol{q}, \boldsymbol{p}) := \begin{cases} \mathbb{E}_{\boldsymbol{p}}(\varphi(\boldsymbol{q}./\boldsymbol{p})) \equiv \sum_{i=1}^m p_i\varphi\left(\frac{q_i}{p_i}\right), & \text{if } \boldsymbol{q} \text{ satisfies } \mathbf{1}^\top\boldsymbol{q} = 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Table 2 Examples of risk functions $\mathcal{F}_{\boldsymbol{p}}$ satisfying (30) and their $\mathcal{I}_{v^*}$, v^* and v

$\mathcal{F}_{\boldsymbol{p}}$	$\mathcal{I}_{v^*}(\boldsymbol{q},\boldsymbol{p}) \equiv \mathcal{F}^*_{\boldsymbol{p}}(\boldsymbol{q})$	$v^*(z)$	$v(z)$
$\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}$	$\delta_{\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}}(\boldsymbol{q})$	$\delta_{[0,1/(1-\alpha)]}$	$\max\{z,0\}/(1-\alpha)$
$\mathrm{LSE}_{(t,\boldsymbol{p})}$	$\mathrm{KL}_{(t,\boldsymbol{p})}(\boldsymbol{q})$	$\frac{z}{t}\log(\frac{z}{t})-\frac{z}{t}+1+\delta_{\mathbb{R}_+}(z)$	$\exp(tz)-1$
$\mathrm{MV}_{(t,\boldsymbol{p})}$	$((\boldsymbol{q}-\boldsymbol{p})^{.2})./(2t\boldsymbol{p})$	$(z-1)^2/(2t)$	$z+(t/2)z^2$

The φ-divergence generalizes the relative entropy. Indeed, with $\varphi(s) = s\log s - s + 1$ (and $0\ln 0 = 0$), $\mathcal{I}_\varphi(\boldsymbol{q},\boldsymbol{p})$ is the Kullback-Leibler divergence, i.e., $\mathrm{KL}_{(1,\boldsymbol{p})}(\boldsymbol{q})$, while with $\varphi(s) = (s-1)^2$, $\mathcal{I}_\varphi(\boldsymbol{q},\boldsymbol{p})$ is the modified χ^2-divergence, i.e., $\chi^2_{(1,\boldsymbol{p})}(\boldsymbol{q})$. See e.g., Table 2 of Reid and Williamson (2011), for the other examples.

Theorem 6 *Let v be a proper l.s.c. convex function on $\mathbb{R}$ such that $v(z) \geq z + B$ with some $B \in \mathbb{R}$. Then, the risk function (9) is proper l.s.c. convex, and it is valid $\mathcal{F}^*_{\boldsymbol{p}}(\boldsymbol{\lambda}) = \mathcal{I}_{v^*}(\boldsymbol{\lambda},\boldsymbol{p})$. Namely,*

$$\mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L}) = \inf_c\{c + \mathbb{E}_{\boldsymbol{p}}(v(\boldsymbol{L} - c\mathbf{1}))\} = \sup_{\boldsymbol{q}}\{\boldsymbol{q}^\top \boldsymbol{L} - \mathcal{I}_{v^*}(\boldsymbol{q},\boldsymbol{p})\}. \tag{30}$$

Furthermore, if there exists $z^\star$ such that $v(z^\star) = z^\star + B$, i.e., B is the minimum of $v(z) - z$ and $z^\star$ is the minimizer, the φ-divergence $\mathcal{I}_{v^*}(\boldsymbol{q},\boldsymbol{p})$ attains the minimum $-B$ at $\boldsymbol{q} = \boldsymbol{p}$. Furthermore, $\mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L})$ is monotonic if $\mathrm{dom}\, v^* \subset \mathbb{R}_+$; $\mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L})$ is positively homogeneous if $v^* = \delta_{[a,a']}$, where $a := \inf\{s : s \in \mathrm{dom}\, v^*\}$ and $a' := \sup\{s : s \in \mathrm{dom}\, v^*\}$.

See Proof of Theorem 6 section in "Appendix" for the proof.

Formula (30) indicates that the risk function of the form (9) is interpreted as a worst-case expected loss which is deducted with the φ-divergence where $\varphi = v^*$. In this view, we can associate each coherent risk function with a φ-divergence which is represented as the indicator function of a (closed convex) set. Table 2 demonstrates the correspondence of $\mathcal{F}_{\boldsymbol{p}}, \mathcal{I}_{v^*}(\boldsymbol{q},\boldsymbol{p})$, v^* and v for CVaR, LSE and MV risk functions. For any $\alpha \in [0,1)$ and $t > 0$, the functions $v(z) - z$ of both $\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}$ and $\mathrm{MV}_{(t,\boldsymbol{p})}$ attain the minimum and minimizer at $z^\star = 0$ and $B = 0$, respectively. On the other hand, $\mathrm{LSE}_{(t,\boldsymbol{p})}$ attains the minimum $B = (1 + \ln t)/t - 1$ at $z^\star = (1/t)\ln(1/t)$ for any $t > 0$. However, each $\mathcal{I}_{v^*}(\boldsymbol{q},\boldsymbol{p})$ attains its minimum at $\boldsymbol{q} = \boldsymbol{p}$. Accordingly, all of these functions in Table 2 can be related to the divergences relative to $\boldsymbol{p}$. See Ben-Tal and Teboulle (2007) for the details and an interpretation as the Optimized Certainty Equivalent based on a concave utility function $u(z) := -v(-z)$.

With the φ-divergence and Theorem 6, the second statement of Proposition 4 can be specified as follows.

Corollary 4 *If the risk function $\mathcal{F}$ is written by (30) with v^* such that v is monotonic, then the dual problem (18) is represented by*

$$\sup_{\boldsymbol{\lambda}}\ -\gamma^*(\boldsymbol{G\lambda}) - \mathcal{I}_{v^*}(\boldsymbol{\lambda},\boldsymbol{p}) - \delta_C(\boldsymbol{\lambda}) \text{ with } C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0\} \cap \Pi^m.$$

Note that this corollary is applicable to CVaR and LSE since they are both monotonic and translation invariant, which can be confirmed also by seeing that $\mathrm{dom}\, v^*$ of CVaR and LSE are in the nonnegative orthant (Table 2) and Theorem 6.

5.2 A connection to geometric interpretation

Crisp and Burges (2000), Bennett and Bredensteiner (2000), Takeda et al. (2013), Kanamori et al. (2013) show that dual problems of a couple of SVMs can be interpreted as the problem of finding two nearest points over two separate sets $I_+ := \{i \in \{1,\dots,m\} : y_i = +1\}$ and $I_- := \{i \in \{1,\dots,m\} : y_i = -1\}$, each corresponding to the data samples having the same label. With the dual formulation (18), we can easily derive the similar implication in a general manner.

To that purpose, let us consider the case where $\mathcal{F}$ is translation invariant and $\boldsymbol{G} = \boldsymbol{X}^\top \boldsymbol{Y}$. In this case, the constraint, $\boldsymbol{y}^\top \boldsymbol{\lambda} = 0$, of (18) can be represented by $\sum_{i\in I_-} \lambda_i = \sum_{i \in I_+} \lambda_i = \frac{1}{2}$, while the first term of the objective, i.e., $-\gamma^*(\boldsymbol{G}\boldsymbol{\lambda})$, can be $-\gamma^*(\sum_{i\in I_+} \boldsymbol{x}_i\lambda_i - \sum_{h\in I_-} \boldsymbol{x}_h\lambda_h)$. Consequently, with a change of variables $\mu_{+,i} := 2\lambda_i$ for $i \in I_+$ and $\mu_{-,i} := 2\lambda_i$ for $i \in I_-$, (18) (or (30)) can be represented as

$$\begin{aligned} -d^\star := \underset{\boldsymbol{\mu}_+,\boldsymbol{\mu}_-}{\text{minimize}}\ & \gamma^*\left(\frac{1}{2}\left(\sum_{i\in I_+} \boldsymbol{x}_i\mu_{+,i} - \sum_{h\in I_-} \boldsymbol{x}_h\mu_{-,h}\right)\right) + \mathcal{F}^*(\tfrac{1}{2}\boldsymbol{\mu}) \\ \text{subject to}\ & \sum_{i\in I_+} \mu_{+,i} = 1,\ \ \sum_{h\in I_-} \mu_{-,h} = 1, \end{aligned}$$

where $\boldsymbol{\mu} := 2\boldsymbol{\lambda}$.

As a further concrete interpretation, let us consider the case where $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\|\le 1}(\boldsymbol{w})$ and $\mathcal{F}$ is given in the form of (9) with monotonic v. Then (18) is represented as

$$\begin{aligned} -d^\star := \underset{\boldsymbol{\mu}_+,\boldsymbol{\mu}_-}{\text{minimize}}\ & \frac{1}{2}\left\|\sum_{i\in I_+} \boldsymbol{x}_i\mu_{+,i} - \sum_{i\in I_-} \boldsymbol{x}_i\mu_{-,i}\right\|^\circ + \mathcal{I}_{v^*}(\tfrac{1}{2}\boldsymbol{\mu}, \boldsymbol{p}) \\ \text{subject to}\ & \boldsymbol{\mu}_+ \in \Pi^{|I_+|},\ \ \boldsymbol{\mu}_- \in \Pi^{|I_-|}, \end{aligned} \tag{31}$$

where $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ are vectors consisting of $\mu_{+,i}$ and $\mu_{-,i}$, respectively. It is noteworthy that the formulation (31) is close to what Kanamori et al. (2013) demonstrate. Precisely, employing the ℓ_2-norm, they virtually present the regularized ERM of the form $\min_\rho\{-2\rho + \frac{1}{m}(\sum_{i=1}^m v(-y_i(\boldsymbol{x}_i^\top \boldsymbol{w} - b)) + \rho)_+\}$ subject to $\|\boldsymbol{w}\|_2^2 \le t$ with $t > 0$. It is easy to see that (31) contains their formulation as a special case.

Besides, the geometric interpretation of the ν-SVM demonstrated by Crisp and Burges (2000), Bennett and Bredensteiner (2000) can also be derived straightforwardly. For example, ν-SVM with $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\|\le 1}(\boldsymbol{w})$ can be explicitly represented as the geometric problem of the form

$$\min_{\boldsymbol{z}_+,\boldsymbol{z}_-} \|\boldsymbol{z}_+ - \boldsymbol{z}_-\|^\circ \ \text{ subject to } \boldsymbol{z}_+ \in \mathcal{Q}_+, \boldsymbol{z}_- \in \mathcal{Q}_-,$$

with

$$\begin{aligned} \mathcal{Q}_+ &:= \left\{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z} = \textstyle\sum_{i\in I_+} \boldsymbol{x}_i\mu_{+,i},\ \boldsymbol{\mu}_+ \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu, 2\boldsymbol{p}_+)}\right\}; \\ \mathcal{Q}_- &:= \left\{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z} = \textstyle\sum_{i\in I_-} \boldsymbol{x}_i\mu_{-,i},\ \boldsymbol{\mu}_- \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu, 2\boldsymbol{p}_-)}\right\}, \end{aligned}$$

where $\boldsymbol{p}_+ := (p_i)_{i\in I_+}$ and $\boldsymbol{p}_- := (p_i)_{i\in I_-}$ are supposed to have the elements in the same order as $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$, respectively, and $\mathcal{Q}_{\mathrm{CVaR}(\alpha,2\boldsymbol{p}_+)} \subset \Pi^{|I_+|}$ and $\mathcal{Q}_{\mathrm{CVaR}(\alpha,2\boldsymbol{p}_-)} \subset \Pi^{|I_-|}$. (Here we admit an abuse of the notation. Precisely, we put $2\boldsymbol{p}_+$ or $2\boldsymbol{p}_-$ in the place of a probability measure $\boldsymbol{p}$.) Note that $\mathcal{Q}_+$ and $\mathcal{Q}_-$ are exactly the reduced convex hulls in Crisp and Burges (2000), Bennett and Bredensteiner (2000). Note that along this line, we can derive the geometric interpretation of an SVM defined with a coherent risk function and a general norm, which is parallel to Kanamori et al. (2013).

In addition, the formulation (31) bridges the geometric interpretation and an information theoretic interpretation. Noting the relation

$$\mathcal{I}_{v^*}\left(\frac{1}{2}\boldsymbol{\mu}, \boldsymbol{p}\right) = \sum_{i\in I_+} p_i v^*\left(\frac{\mu_i}{2p_i}\right) + \sum_{h\in I_-} p_h v^*\left(\frac{\mu_h}{2p_h}\right),$$

the second term of the objective of (31) can be interpreted as a penalty on the deviations between the weight vectors $\boldsymbol{\mu}_+$ (or $\boldsymbol{\mu}_-$) and $2\boldsymbol{p}_+$ (or $2\boldsymbol{p}_-$, respectively). If we further suppose that

$$\sum_{i\in I_+} p_i = \sum_{h\in I_-} p_h \left(= \frac{1}{2}\right), \tag{32}$$

the penalty term is rewritten as $\mathcal{I}_\varphi(\boldsymbol{\mu}_+, \boldsymbol{r}_+) + \mathcal{I}_\varphi(\boldsymbol{\mu}_-, \boldsymbol{r}_-)$ where $\boldsymbol{r}_+ := (r_i)_{i\in I_+}$ with $r_i = p_i/\sum_{h\in I_+} p_h$ for $i \in I_+$ and $\boldsymbol{r}_- := (r_i)_{i\in I_-}$ with $r_i = p_i/\sum_{h\in I_-} p_h$ for $i \in I_-$. Namely, we can symbolically recast (31) as the minimization of

$$\frac{1}{2}\|\mathbb{E}_{\boldsymbol{\mu}}(\boldsymbol{x}|y = +1) - \mathbb{E}_{\boldsymbol{\mu}}(\boldsymbol{x}|y = -1)\|^\circ + \mathcal{I}_\varphi(\boldsymbol{\mu}, \boldsymbol{p}|y = +1) + \mathcal{I}_\varphi(\boldsymbol{\mu}, \boldsymbol{p}|y = -1),$$

over $\boldsymbol{\mu} \in \Pi^m$, where $\mathbb{E}_{\boldsymbol{\mu}}(\boldsymbol{x}|y = a)$ are conditional expectation, and $\mathcal{I}_\varphi(\boldsymbol{\mu}, \boldsymbol{p}|y = a)$ are divergence between conditional distributions, "$\boldsymbol{\mu}|y = a$" and "$\boldsymbol{p}|y = a$" with $a = +1$ or $a = -1$.

It is worth mentioning that the approach of Bennett and Mangasarian (1992) virtually employs the condition (32) and attains a nice performance in the breast cancer data set (see Wolberg et al. 2013).

5.3 Distributionally robust SVMs

Different from the robust optimization modeling described in Sect. 3.2, the so-called *distributionally robust optimization* is also popular in the literature. In this subsection, we show that a class of generalized SVM formulations described in this paper also fits into this robust optimization modeling approach.

In existing SVMs, the samples are usually assumed to be independently drawn from an unknown distribution, and the empirical probability $\boldsymbol{p} = \mathbf{1}/m$ is used as $\boldsymbol{p}$. However, such an i.i.d. assumption is often unfulfilled. For example, we can consider a situation where the samples are i.i.d. within each label samples while the (prior) distribution of labels, $\vartheta := \mathbb{P}\{y = +1\}(= 1 - \mathbb{P}\{y = -1\})$, is not known. Namely, we can assume $p_i = \vartheta/|I_+|$ for $y_i = +1$ and $p_i = (1 - \vartheta)/|I_-|$ for $y_i = -1$, but ϑ is under uncertainty. In such a case, the choice of the uniform distribution may not be the best.

In general, let us consider the case where $\boldsymbol{p}$ is under uncertainty of the form: $\boldsymbol{p} + \boldsymbol{\delta} \in P$ with some P satisfying $\boldsymbol{p} \in P \subset \Pi^m$. Similarly to Sect. 3.2, one reasonable strategy is to consider the worst case over the set P. Let us list examples of the uncertainty set P.

- $P = \mathcal{Q}_{\mathrm{Fi}(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K)} := \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K\}$, with $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K \in \Pi^m$;
- $P = \mathcal{Q}_{\mathrm{Dist}(\|\cdot\|', \boldsymbol{A}, \boldsymbol{p})} := \{\boldsymbol{\pi} \in \Pi^m : \boldsymbol{\pi} = \boldsymbol{p} + \boldsymbol{A}\boldsymbol{\zeta}, \|\boldsymbol{\zeta}\|' \leq 1\}$, with $\boldsymbol{A} \in \mathbb{S}^m_{++}$, $\|\cdot\|'$: a norm;
- $P = \mathcal{Q}_{\mathcal{I}_\varphi(t,\boldsymbol{p})} := \left\{\boldsymbol{\pi} \in \Pi^m : \mathcal{I}_\varphi(\boldsymbol{\pi}, \boldsymbol{p}) \leq t\right\}$, with $t > 0$,

where $\mathbb{S}^m_{++}$ denotes the $m \times m$ real symmetric positive definite matrices. The first example indicates the situation where K candidates $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K$ for $\boldsymbol{p}$ are possible. The second and third examples are the case where the possible deviations are given by convex sets defined with some norm $\|\cdot\|'$ and some φ-divergence, respectively. Specifically, $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|,\boldsymbol{A},\boldsymbol{p})}$ denotes a set of probability measures which are away from $\boldsymbol{p}$ with distance at most 1 under a norm $\|\cdot\|$ and a metric $(\boldsymbol{A}^{-1})^2$. Especially when $\|\cdot\| = \|\cdot\|_\infty$ and $\boldsymbol{A} = \mathrm{diag}(\overline{\boldsymbol{\zeta}})$ with some $\overline{\boldsymbol{\zeta}} \geq \boldsymbol{0}$, the set forms a box-type constraint, i.e., $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|_\infty,\mathrm{diag}(\overline{\boldsymbol{\zeta}}),\boldsymbol{p})} = \Pi^m \cap [\boldsymbol{p} - \overline{\boldsymbol{\zeta}}, \boldsymbol{p} + \overline{\boldsymbol{\zeta}}]$.

Let us consider that the risk function $\mathcal{F}_{\boldsymbol{p}}$ has the form (30). Namely, we consider a distributionally robust version of the primal formulation:

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \sup_{\boldsymbol{\pi}\in P} \mathcal{F}_{\boldsymbol{\pi}}(\boldsymbol{L}) + \gamma(\boldsymbol{w}). \tag{33}$$

Note that the worst-case risk function is given as

$$\text{Worst-}\mathcal{F}_P(\boldsymbol{L}) := \sup_{\boldsymbol{\pi}\in P} \mathcal{F}_{\boldsymbol{\pi}}(\boldsymbol{L}) = \sup_{\boldsymbol{\pi}\in P,\boldsymbol{q}} \{\boldsymbol{q}^\top \boldsymbol{L} - \mathcal{I}_\varphi(\boldsymbol{q},\boldsymbol{\pi})\} = \sup_{\boldsymbol{q}}\{\boldsymbol{q}^\top \boldsymbol{L} - \inf_{\boldsymbol{\pi}\in P} \mathcal{I}_\varphi(\boldsymbol{q},\boldsymbol{\pi})\}.$$

The last part indicates that $(\text{Worst-}\mathcal{F}_P)^*(\boldsymbol{q}) = \inf_{\boldsymbol{\pi}\in P} \mathcal{I}_\varphi(\boldsymbol{q},\boldsymbol{\pi})$, where we can independently show that this is convex in $\boldsymbol{q}$ as long as the φ-divergence is given with a convex φ. (See e.g., Section 3.2.6 of Boyd and Vandenberghe (2004), for the details.)

Proposition 5 *If P is a convex set, the dual formulation of the distributionally robust version* (33) *is given as the following convex minimization*

$$\begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{\pi}}{\textit{maximize}} & -\gamma^*(\boldsymbol{G\lambda}) - \mathcal{I}_\varphi(\boldsymbol{\lambda},\boldsymbol{\pi}) \\ \textit{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{1}^\top\boldsymbol{\lambda} = 1,\ \boldsymbol{\pi} \in P. \end{array}$$

If $P = \mathcal{Q}_{\mathrm{Fi}(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K)}$, the distributionally robust version of the generalized dual formulation is rewritten by

$$\begin{array}{ll} \underset{\boldsymbol{\lambda},\theta}{\textit{maximize}} & -\gamma^*(\boldsymbol{G\lambda}) - \theta \\ \textit{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{1}^\top\boldsymbol{\lambda} = 1,\ \theta \geq \mathcal{I}_\varphi(\boldsymbol{\lambda},\boldsymbol{\pi}_k),\ k = 1,\ldots,K. \end{array}$$

Although we describe the case of $P = \mathcal{Q}_{\mathrm{Fi}(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K)}$ separately from the case where P is a convex set, we can treat Worst-$\mathcal{F}_P$ in a unified manner when the original risk function $\mathcal{F}_{\boldsymbol{p}}$ is positively homogeneous. Indeed, we then have

$$\text{Worst-}\mathcal{F}_P(\boldsymbol{L}) = \sup_{\boldsymbol{p}\in P}\ \sup_{\boldsymbol{q}\in\mathcal{Q}_{\mathcal{F}}(\boldsymbol{p})} \boldsymbol{q}^\top\boldsymbol{L} = \sup_{\boldsymbol{q}}\{\boldsymbol{L}^\top\boldsymbol{q} : \boldsymbol{q} \in \bigcup_{\boldsymbol{p}\in P} \mathcal{Q}_{\mathcal{F}}(\boldsymbol{p})\}. \tag{34}$$

The union, $\cup_{\boldsymbol{p}\in P}\mathcal{Q}_{\mathcal{F}}(\boldsymbol{p})$, in (34) can be a nonconvex set. However, the convex hull of the union provides an equivalent coherent risk function. Namely, we have Worst-$\mathcal{F}_P(\boldsymbol{L}) = \sup_{\boldsymbol{q}}\{\boldsymbol{q}^\top\boldsymbol{L} : \boldsymbol{q} \in \mathrm{conv}(\cup_{\boldsymbol{p}\in P}\mathcal{Q}_{\mathcal{F}}(\boldsymbol{p}))\}$. Since the convex hull of the risk envelopes become another (possibly, larger) risk envelope, the distributionally robust coherent risk function-based SVM is also another coherent risk function-based SVM. Accordingly, with a positively homogeneous risk function $\mathcal{F}$, the dual of the distributionally robust version is given by

$$\begin{array}{rl} d^\star := \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G\lambda}\|^\circ \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{\lambda} \in \mathrm{conv}(\{\mathcal{Q}_{\mathcal{F}}(\boldsymbol{p}) : \boldsymbol{p} \in P\}). \end{array} \tag{35}$$

For example, if we employ the uncertainty sets P listed above, the distributionally robust version of ν-SVMs (Worst-CVaR$_{(1-\nu,P)}$, $\|\cdot\|$) are represented in the following dual forms, respectively:

[Finite-scenario uncertainty]

$$P = \mathcal{Q}_{\mathrm{Fi}(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K)} \leftrightarrow \begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{\pi},\boldsymbol{\tau}}{\text{maximize}} & -\|\boldsymbol{G\lambda}\|^{\circ} \\ \text{subject to} & \boldsymbol{y}^{\top}\boldsymbol{\lambda} = 0,\ \boldsymbol{1}^{\top}\boldsymbol{\lambda} = 1,\ \boldsymbol{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \boldsymbol{\pi} = \sum_{k=1}^{K} \tau_k \boldsymbol{p}_k,\ \boldsymbol{1}^{\top}\boldsymbol{\tau} = 1,\ \boldsymbol{\tau} \ge \boldsymbol{0}; \end{array}$$

[Distance-based uncertainty]

$$P = \mathcal{Q}_{\mathrm{Dist}(\|\cdot\|',\boldsymbol{A},\boldsymbol{1}/m)} \leftrightarrow \begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{\pi},\boldsymbol{\zeta}}{\text{maximize}} & -\|\boldsymbol{G\lambda}\|^{\circ} \\ \text{subject to} & \boldsymbol{y}^{\top}\boldsymbol{\lambda} = 0,\ \boldsymbol{1}^{\top}\boldsymbol{\lambda} = 1,\ \boldsymbol{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \boldsymbol{\pi} = \frac{1}{m}\boldsymbol{1} + \boldsymbol{A\zeta},\ \boldsymbol{1}^{\top}\boldsymbol{A\zeta} = 0,\ \|\boldsymbol{\zeta}\|' \le 1; \end{array}$$

[Entropy-based uncertainty]

$$P = \mathcal{Q}_{\mathrm{KL}(t,\boldsymbol{1}/m)} \leftrightarrow \begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{\pi}}{\text{maximize}} & -\|\boldsymbol{G\lambda}\|^{\circ} \\ \text{subject to} & \boldsymbol{y}^{\top}\boldsymbol{\lambda} = 0,\ \boldsymbol{1}^{\top}\boldsymbol{\lambda} = 1,\ \boldsymbol{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \boldsymbol{\pi}^{\top}\ln(m\boldsymbol{\pi}) \le t,\ \boldsymbol{1}^{\top}\boldsymbol{\pi} = 1, \end{array}$$

where the first one with $\|\cdot\|^{\circ} = \|\cdot\|_2$ is presented in Wang (2012), which extends it into a multi-class classification setting.

The distributionally robust SVMs presented above are different from existing ones (e.g., Wang et al. 2015) in that it is easy to obtain the dual formulation based on the dual representation of the inseparable risk function (30). As seen in preceding sections, (33) can be associated with φ-divergences and (35) is obtained straightforwardly with the help of the Fenchel duality.

It is noteworthy that the distributional robustification technique above incorporates prior knowledge on the distribution without significantly increasing the complexity of the optimization problem, specifically when such information is given by moments. For example,

- When the average of the j-th attribute of samples having the label $y_i = +1$ belongs in a certain interval $[l_j^+, u_j^+]$, we include this information into the dual problem as the constraint:

$$l_j^+ \le \sum_{i \in I_+} \pi_i x_{ij} \le u_j^+.$$

- When the prior probability of a sample being drawn from the group of label $y_i = +1$ is twice to thrice as large as that from $y_i = -1$, we include this information into the dual problem as the constraint:

$$2 \sum_{i \in I_-} \pi_i \le \sum_{i \in I_+} \pi_i \le 3 \sum_{i \in I_-} \pi_i.$$

Although it is known that simple robust optimization modeling often leads to excessively conservative results, adding experts' knowledge as constraints can be helpful to escape from those situations.

6 Concluding remarks

This paper studies formulations of SVMs for binary classification in a unified way, particularly considering the capability of inseparable risk functions and non-ℓ_2-norms, while also providing insights on the formulations from various perspectives.

- When using positively homogeneous functions, the choice of the form of regularizer requires careful attention. (Sect. 3.1).

- Corresponding to the dual characterizations of the three properties of the risk function (monotonicity, translation invariance, and positive homogeneity), we can express the dual formulation in interpretable ways (Sect. 5.1). More specifically, monotonic and translation invariant risk functions are shown to be associated with geometric and probabilistic interpretations (Sect. 5.2).
- In relation to robust optimization modeling, we draw two perspectives. With monotonic and translation invariant risk functions, the regularized ERM formulation can be viewed as a robust optimization (Sect. 3.2). Additionally, for these risk functions the distributionally robust modeling can be easily incorporated into the dual formulation (Sect. 5.3).

As stated in Introduction, a motivation of this study was the use of recently developed polyhedral norms for the regularizer. We see that the Ivanov regularization seems to be the unique solution for the combination with positive homogeneous risk measures, and that is the reason why we have focused on that case in the analysis (e.g., Sect. 4.2). Through an experiment, which is not reported in the current manuscript, we observed that within a comparable amount of time the use of a family of polyhedral norms could achieve a better out-of-sample performance than the standard ℓ_2-regularized SVM, whose difference is in the regularizer. See Gotoh and Uryasev (2013) for the details.

While we supposed that the argument of $\mathcal{F}$ was of the form $\boldsymbol{L} = -(\boldsymbol{G}\boldsymbol{w} - \boldsymbol{y}b)$ and only $\boldsymbol{w}$ was regularized (i.e., b was not regularized), we can treat a variety of existing formulations. On the other hand, excluded classes of risk functions or losses, such as

- $L_{i,j} = \boldsymbol{w}^\top \boldsymbol{x}_i - \boldsymbol{w}^\top \boldsymbol{x}_j$, where i are samples of $y_i = -1$ and j are samples of $y_i = +1$,

remain to be investigated.

This framework can be extended to other types of machine learning tasks, such as multiclass classifications and regression, in a similar manner. In particular, the application of the CVaR norms and the deltoidal norms to the multiple kernel learning (Kloft et al. 2011) can be a promising extension.

Acknowledgements The first version of this research was done during the visit of the first author to the University of Florida, USA. The first author appreciates the financial support by Chuo university and the environmental support by the Department of Industrial and Systems Engineering of the University of Florida. The research of the first author is supported in part by JSPS KAKENHI Grant Numbers 23710176 and 15K01204. Research of the second author of the paper was partially supported by the AFOSR Grant FA9550-11-1-0258, "New Developments in Uncertainty: Linking Risk Management, Reliability, Statistics and Stochastic Optimization." The authors acknowledge Prof. R. Terry Rockafellar for his comments and Mr. Matthew Norton for proofreading.

Appendix: Proofs

Proof of Proposition 2

Note that $(\boldsymbol{w}, b) = \boldsymbol{0}$ is feasible to (5) with the objective value 0. Accordingly, $p^\star \leq 0$. Suppose that there exists a solution $(\bar{\boldsymbol{w}}, \bar{b})$ whose objective value is negative. Then, for any $\tau > 1$, the solution $(\tau\bar{\boldsymbol{w}}, \tau\bar{b})$ attains a smaller objective value due to the positive homogeneity of the objective function $\gamma(\boldsymbol{w}) + \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b)$ with respect to $(\boldsymbol{w}, b)$. □

Proof of Theorem 2

For the sake of the simplicity, the following proof is based on the dual representation of the function $\mathcal{F}$.

From Corollary 4, the worst-case empirical risk is presented by

$$\begin{aligned}
&\max_{\boldsymbol{\Delta}\in\mathcal{S}} \mathcal{F}(-\boldsymbol{Y}\{(\boldsymbol{X}-\boldsymbol{\Delta})\boldsymbol{w}-\boldsymbol{1}b\})\\
&\quad= \max_{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m)\in\mathcal{S}} \max_{\boldsymbol{\lambda}\in\Pi^m}\left\{\sum_{i=1}^m \lambda_i\{-y_i(\boldsymbol{x}_i-\boldsymbol{\delta}_i)^\top\boldsymbol{w}+y_ib\}-\mathcal{F}^*(\boldsymbol{\lambda})\right\}\\
&\quad= \max_{\boldsymbol{\lambda}\in\Pi^m} \max_{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_m)\in\mathcal{S}}\left\{\sum_{i=1}^m \lambda_i(y_i\boldsymbol{w}^\top\boldsymbol{\delta}_i-y_i\boldsymbol{x}_i^\top\boldsymbol{w}+y_ib)-\mathcal{F}^*(\boldsymbol{\lambda})\right\}\\
&\quad= \max_{\boldsymbol{\lambda}\in\Pi^m}\left\{\sum_{i=1}^m \lambda_i(\max_{\|\boldsymbol{\delta}_i\|^\circ\le C} y_i\boldsymbol{w}^\top\boldsymbol{\delta}_i-y_i\boldsymbol{x}_i^\top\boldsymbol{w}+y_ib)-\mathcal{F}^*(\boldsymbol{\lambda})\right\} \qquad (\text{because } \boldsymbol{\lambda}\ge\boldsymbol{0})\\
&\quad= \max_{\boldsymbol{\lambda}\in\Pi^m}\left\{\sum_{i=1}^m \lambda_i(C\|\boldsymbol{w}\|-y_i\boldsymbol{x}_i^\top\boldsymbol{w}+y_ib)-\mathcal{F}^*(\boldsymbol{\lambda})\right\}\\
&\quad= C\|\boldsymbol{w}\|+\max_{\boldsymbol{\lambda}\in\Pi^m}\left\{\sum_{i=1}^m \lambda_i(-y_i\boldsymbol{x}_i^\top\boldsymbol{w}+y_ib)-\mathcal{F}^*(\boldsymbol{\lambda})\right\} \qquad (\text{because } \boldsymbol{1}^\top\boldsymbol{\lambda}=1)\\
&\quad= C\|\boldsymbol{w}\|+\mathcal{F}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{1}b)).
\end{aligned}$$

The fourth equality follows the fact that $\max_{\|\boldsymbol{\delta}_i\|^\circ\le C}\{y_i\boldsymbol{\delta}_i\boldsymbol{w}\}=C|y_i|\|\boldsymbol{w}\|=C\|\boldsymbol{w}\|$. □

Proofs of Theorem 3 and the modification of the condition (a) for the Ivanov regularization

To prove Theorem 3 we should confirm the correspondence between the functions and variables in Corollary 31.2.1 of Rockafellar (1970) and those in our setting.

Corollary 31.2.1 of Rockafellar (1970) *Let f be a closed proper convex function on $\mathbb{R}^n$, let g be a closed proper convex function on $\mathbb{R}^m$, and let $\boldsymbol{A}$ be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then we have*

$$\inf_{\boldsymbol{z}}\{f(\boldsymbol{z})+g(\boldsymbol{A}\boldsymbol{z})\}=\sup_{\boldsymbol{\lambda}}\{-f^*(-\boldsymbol{A}^\top\boldsymbol{\lambda})-g^*(\boldsymbol{\lambda})\}$$

if either of the following conditions is satisfied:

(a) *There exists a $\boldsymbol{z}\in\mathrm{ri}(\mathrm{dom}\, f)$ such that $\boldsymbol{A}\boldsymbol{z}\in\mathrm{ri}(\mathrm{dom}\, g)$.*
(b) *There exists a $\boldsymbol{\lambda}\in\mathrm{ri}(\mathrm{dom}\, g^*)$ such that $-\boldsymbol{A}^\top\boldsymbol{\lambda}\in\mathrm{ri}(\mathrm{dom}\, f^*)$.*

Under the condition (a) the supremum is attained at some $\boldsymbol{\lambda}$, while under the condition (b) the infimum is attained at some $\boldsymbol{x}$. In addition, if $\mathcal{F}$ (or equivalently, $\mathcal{F}^*$) is polyhedral, "ri" *can be omitted.*

Indeed, letting $f(\boldsymbol{z},z_0)=\gamma(\boldsymbol{z})$, $\mathcal{F}=g$ and $\boldsymbol{A}=(-\boldsymbol{G}^\top,\boldsymbol{y})$, we see that $f^*(\boldsymbol{w},b)=\gamma^*(\boldsymbol{w})$ if $b=0$; $+\infty$ if $b\ne0$. This implies that $\boldsymbol{y}^\top\boldsymbol{\lambda}=0$ must holds in order to maximize $-f^*(-\boldsymbol{A}^\top\boldsymbol{\lambda})$. □

Modification of the condition (a) for the Ivanov regularization

We reach the conclusion just by observing that with $f(\boldsymbol{z},z_0)=\delta_{\{(\boldsymbol{u},v):\|\boldsymbol{u}\|\le1\}}(\boldsymbol{z},z_0)$, $f^*(\boldsymbol{w},b)=\|\boldsymbol{w}\|^\circ$ if $b=0$; $+\infty$ if $b\ne0$. (Note that the conjugate of an indicator function of a set is its support function.) □

Proof of Theorem 4

Similarly to Theorem 3, in order to prove Theorem 4, we just need to confirm the correspondence of the notation between it and Theorem 31.3 of Rockafellar (1970) and to simplify the results.

Theorem 31.3 (Rockafellar 1970) *Let f be a closed proper convex function on $\mathbb{R}^n$, let g be a closed proper convex function on $\mathbb{R}^m$, and let $\boldsymbol{A}$ be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then, in order that $\boldsymbol{z}^\star$ and $\boldsymbol{\lambda}^\star$ be vectors such that*

$$f(\boldsymbol{z}^\star) + g(\boldsymbol{A}\boldsymbol{z}^\star) = -f^*(-\boldsymbol{A}^\top\boldsymbol{\lambda}^\star) - g^*(\boldsymbol{\lambda}^\star),$$

it is necessary and sufficient that $\boldsymbol{z}^\star$ and $\boldsymbol{\lambda}^\star$ satisfy the Karush-Kuhn-Tucker (KKT) conditions:

$$-\boldsymbol{A}^\top\boldsymbol{\lambda}^\star \in \partial f(\boldsymbol{z}^\star), \quad \boldsymbol{A}\boldsymbol{z}^\star \in \partial g^*(\boldsymbol{\lambda}^\star). \tag{36}$$

For (20) and (21), the KKT condition (36) can be explicitly written by

$$\begin{pmatrix} \boldsymbol{G} \\ -\boldsymbol{y}^\top \end{pmatrix} \boldsymbol{\lambda}^\star \in \partial\delta_{\|\cdot\|\le 1}(\boldsymbol{w}^\star, b^\star), \quad -\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b \in \partial\mathcal{F}^*(\boldsymbol{\lambda}^\star).$$

Note that

$$\partial\delta_{\|\cdot\|\le 1}(\boldsymbol{w}, b) = \begin{cases} N_{\|\cdot\|\le 1}(\boldsymbol{w}, b), & \text{if } \|\boldsymbol{w}\| \le 1, \\ \emptyset, & \text{otherwise}, \end{cases}$$

where $N_{\|\cdot\|\le 1}(\boldsymbol{w}, b)$ denotes the normal cone to the set $\{(\boldsymbol{z}, z_0) \in \mathbb{R}^{n+1} : \|\boldsymbol{z}\| \le 1\}$ at a point $(\boldsymbol{w}, b)$ (see p. 215 of Rockafellar 1970, forthedetails), and

$$\begin{aligned} N_{\|\cdot\|\le 1}(\boldsymbol{w}, b) &= \{(\boldsymbol{z}, z_0) : (\boldsymbol{y} - \boldsymbol{w})^\top\boldsymbol{z} + (y_0 - b)z_0 \le 0, \text{ for all } (\boldsymbol{y}, y_0) \text{ such that } \|\boldsymbol{y}\| \le 1\} \\ &= \{(\boldsymbol{z}, 0) : \|\boldsymbol{z}\|^\circ \le \boldsymbol{z}^\top\boldsymbol{w}\}. \end{aligned}$$

Besides, by definition of the dual norm, we have $\boldsymbol{z}^\top\boldsymbol{w}^\star \le \|\boldsymbol{z}\|^\circ\|\boldsymbol{w}^\star\| \le \|\boldsymbol{z}\|^\circ$ for any $\boldsymbol{z} \in \mathbb{R}^n$ if $\|\boldsymbol{w}^\star\| \le 1$. Therefore, $(\boldsymbol{u}, 0) \in N_{\|\cdot\|\le 1}(\boldsymbol{w}^\star, b)$ for $\boldsymbol{w}^\star$ satisfying $\|\boldsymbol{w}^\star\| \le 1$ is equivalent to the condition that $\|\boldsymbol{u}\|^\circ = \boldsymbol{u}^\top\boldsymbol{w}^\star$. □

Proof of Theorem 6

Proof To show the first statement, it suffices to confirm the boundedness of $\mathcal{F}_{\boldsymbol{p}}$. Indeed, we see that $\mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L}) \ge \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + B > -\infty$. The l.s.c. convexity is obvious.

The expression (30) can be considered as a special case of Theorem 4.2 of Ben-Tal and Teboulle (2007) except for the conditions on v. Although the conditions are independent of (30), the proof is given for completeness. To see (30), observe that the right-hand side is equal to $\inf_c \sup_{\boldsymbol{\lambda}}\{\sum_{i=1}^m (L_i\lambda_i - p_i v^*(\frac{\lambda_i}{p_i})) - c(\sum_{i=1}^m \lambda_i - 1)\} = \inf_c\{c + \sum_{i=1}^m p_i \sup_{\lambda_i}\{\frac{\lambda_i}{p_i}(L_i - c) - v^*(\frac{\lambda_i}{p_i})\}\} = \inf_c\{c + \sum_{i=1}^m p_i v(L_i - c)\} = \mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L})$. To prove the third statement, first observe that $\min_{\boldsymbol{q}} \mathcal{I}_\varphi(\boldsymbol{q}, \boldsymbol{p}) = \sup_z \inf_{\boldsymbol{\zeta}}\{\sum_{i=1}^m v^*(\zeta_i) - z(\sum_{i=1}^m p_i\zeta_i - 1)\} = \sup_z\{z - \sum_{i=1}^m p_i \sup_{\zeta_i}\{z\zeta_i - v^*(\zeta_i)\}\} = \sup_z\{z - v(z)\} = z^\star - v(z^\star) = -B$. Also, due to the proper convexity of v, it is valid $\partial v^*(\zeta^\star) \ni z^\star$, which is equivalent to $\zeta^\star \in \partial v(z^\star)$ since v is l.s.c. (Theorem 23.5 of Rockafellar 1970). Obviously, $1 \in \partial v(z^\star)$, and accordingly, $\zeta^\star \equiv q_i^\star/p_i = 1$ is optimal, which proves the statement. Corresponding to Corollary 1, the sufficient conditions of the monotonicity and the positive homogeneity in terms of v^* are obtained by applying Theorem 5 to v and Quadrangle Theorem (c) of Rockafellar and Uryasev (2013). □

References

American Optimal Decisions, Inc. (2009). *Portfolio Safeguard* (PSG). www.aorda.com/aod/psg.action.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, *2*(1), 183–202.

Bennett, K. P., & Bredensteiner, E. J. (2000). Duality and geometry in SVM classifiers. In *Proceedings of the international conference on machine learning* (pp. 57–64).

Bennett, K. P., & Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, *1*, 23–34.

Ben-Tal, A., & Teboulle, M. (2007). An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, *17*, 449–476.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Chen, P.-H., Lin, C.-J., & Schölkopf, B. (2005). A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry*, *21*, 111–136.

Christmann, A., & Steinwart, I. (2004). On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, *5*, 1007–1034.

Christopher, J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Crisp, D. J., & Burges, C. J. C. (2000). A geometric interpretation of ν-SVM classifiers. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems 12* (pp. 244–250). Cambridge, Massachusetts: MIT Press.

Csiszár, I. (1967). Information-type measures of divergence of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 299–318.

Föllmer, H., & Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, *6*(4), 429–447.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Gotoh, J., & Takeda, A. (2005). A linear classification model based on conditional geometric score. *Pacific Journal of Optimization*, *1*(2), 277–296.

Gotoh, J., Takeda, A., & Yamamoto, R. (2014). Interaction between financial risk measures and machine learning methods. *Computational Management Science*, *11*(4), 365–402. doi:10.1007/s10287-013-0175-5.

Gotoh, J., & Uryasev, S. (2013). Support vector machines based on convex risk functionals and general norms. Research report #2013-5, Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida. Downloadable from www.ise.ufl.edu/uryasev/publications/.

Gotoh, J., & Uryasev, S. (2016). Two pairs of families of polyhedral norms versus ℓ_p-norms: Proximity and applications in optimization. *Mathematical Programming, Series A*, *156*(1), 391–431. doi:10.1007/s10107-015-0899-9.

Grant, M., & Boyd, S. (2012). CVX: MATLAB software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx.

Kanamori, T., Takeda, A., & Suzuki, T. (2013). Conjugate relation between loss functions and uncertainty sets in classification problems. *Journal of Machine Learning Research*, *14*, 1461–1504.

Kloft, M., Brefeld, U., Sonnenburg, S., & Zien, A. (2011). ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, *12*, 953–997.

Koh, K., Kim, S.-J., & Boyd, S. (2007). An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine Learning Research*, *8*, 1519–1555.

Livni, R., Crammer, K., & Globerson, A. (2012). A simple geometric interpretation of svm using stochastic adversaries. In *Proceedings of the 15th international conference on artificial intelligence and statistics*.

Mangasarian, O. L. (1999). Arbitrary-norm separating plane. *Operations Research Letters*, *24*, 15–23.

Pavlikov, K., & Uryasev, S. (2014). CVaR norm and applications in optimization. *Optimization Letters, 8*(7), 1999–2020.

Pedroso, J. P., & Murata, N. (2001). Support vector machines with different norms: Motivation, formulations and results. *Pattern Recognition Letters*, *22*, 1263–1272.

Perez-Cruz, F., Weston, J., Herrmann, D., & Schölkopf, B. (2003). Extension of the ν-SVM range for classification. In J. A. K. Suykens, G. Horvath, S. Basu, C. Micchelli, & J. Vandewalle (Eds.), *Advances in learning theory: Methods, models and applications 190* (pp. 179–196). Amsterdam: IOS Press.

Rätsch, G., Schölkopf, B., Smola, A. J., Mika, S., Onoda, T., & Müller, K.-R. (2000). Robust ensemble learning. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 207–219). Cambridge, MA: MIT Press.

Reid, M. D., & Williamson, R. C. (2011). Information, divergence and risk for binary experiments. *The Journal of Machine Learning Research*, *12*, 731–817.

Rifkin, R. M., & Lippert, R. A. (2007). Value regularization and Fenchel duality. *The Journal of Machine Learning Research*, *8*, 441–479.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton, New Jersey: Princeton University Press.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk*, *2*(3), 21–41.

Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, *26*, 1443–1471.

Rockafellar, R. T., & Uryasev, S. (2013). The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science, 16*(1–2), 33–53.

Ruszczyński, A., & Shapiro, A. (2005). Optimization of risk measures. In G. Calafiore & F. Dabbene (Eds.), *Probabilistic and randomized methods for design under uncertainty* (pp. 117–158). London: Springer.

Ruszczyński, A., & Shapiro, A. (2006). Optimization of convex risk functions. *Mathematics of Operations Research*, *31*(3), 433–452.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *12*(5), 1207–1245.

Suykens, J. A. K., & Vandewalle, J. P. L. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*(3), 293–300.

Takeda, A., Mitsugi, H., & Kanamori, T. (2013). A unified classification model based on robust optimization. *Neural Computation*, *25*(3), 759–804.

Takeda, A., & Sugiyama, M. (2008). ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25 th international conference on machine learning* (pp. 1056–1063).

Tsyurmasto, P., Gotoh, J., & Uryasev, S. (2013). Support vector classification with positive homogeneous risk functionals. Research report 2013-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida. Downloadable from www.ise.ufl.edu/uryasev/publications/.

Wang, X., Fan, N., & Pardalos, P. M. (2015). Robust chance-constrained support vector machines with second-order moment information. *Annals of Operations Research*,. doi:10.1007/s10479-015-2039-6.

Wang, Y. (2012). Robust ν-support vector machine based on worst-case conditional value-at-risk minimization. *Optimization Methods and Software*, *27*(6), 1025–1038.

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (2013). Wisconsin Diagnostic Breast Cancer (WDBC) Data Set. ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/WDBC/. Accessed July 24, 2013.

Xu, H., Caramanis, C., & Mannor, S. (2009a). Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, *10*, 1485–1510.

Xu, H., Caramanis, C., Mannor, S., & Yun, S. (2009b). Risk sensitive robust support vector machines. In *48th IEEE conference on decision and control (CDC09)*, Shanghai, China.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, *32*(1), 56–134.

Zhou, W., Zhang, L., & Jiao, L. (2002). Linear programming support vector machines. *Pattern Recognition*, *35*, 2927–2936.