Support Vector Machines Based on Convex Risk Functionals and General Norms

Jun-ya Gotoh, Stan Uryasev

RESEARCH REPORT # 2013-6

Department of Industrial and Systems Engineering
Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
E-mail: jgoto@indsys.chuo-u.ac.jp

Risk Management and Financial Engineering Lab
Department of Industrial and Systems Engineering
University of Florida,
303 Weil Hall, Gainesville, FL 32611, USA
E-mail: uryasev@ufl.edu

First draft: Nov 20, 2013
This draft: Mar 21, 2014
Correspondence should be addressed to: Jun-ya Gotoh

Abstract

This paper studies unified formulations of support vector machines (SVMs) for binary classification on the basis of convex analysis, especially, convex risk functionals theory, which is recently developed in the context of financial optimization. Using the notion of convex risk functionals, a pair of primal and dual formulations of the SVMs are described in a general manner, and duality results and optimality conditions are established in some standard setting. Interpretability of dual formulations is related to properties of the convex risk functionals. Besides, robust optimization modelings are easily incorporated. The formulation uses arbitrary norms for regularizers and incorporates new families of norms (in place of the ℓ_p-norms) in SVMs. We evaluate the proximity between SVMs with an arbitrary norm and the ℓ_2-norm, which also plays a role in kernelization. Numerical results demonstrate that with the new family of polyhedral norms, regularizer tuning can be efficiently incorporated via (possibly, parametric) linear programming, resulting in a better performance than the ℓ_2-regularizer.
Keywords: support vector machine, SVM, binary classification, convex risk functional, duality, norm

1 Introduction

Background. The primal formulation of *Support Vector Machines (SVMs)* has the form of a bi-objective minimization which is usually referred to as the *regularized empirical risk minimization (ERM)* or the *structural risk minimization* (see, e.g., Burges, 1998). In binary classification, a model, which is represented by a hyperplane, fits a given data set as precisely as possible, while the model would not overfit the data set. We consider the labeled data samples $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^n$ denotes attributes of sample i and $y_i \in \{\pm 1\}$ denotes its class label. Typically, in order to obtain a classification hyperplane $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{w}^\top \boldsymbol{x} = b\}$, the regularized ERM is written as an optimization problem:

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \underbrace{\mathcal{F}(\boldsymbol{L}(\boldsymbol{w},b))}_{\text{[Empirical Risk]}} + \underbrace{\gamma(\boldsymbol{w})}_{\text{[Regularizer]}},$$

where $(\boldsymbol{w}, b)$ are parameters determining a hyperplane, $\boldsymbol{L}(\boldsymbol{w}, b)$ is a vector of $\mathbb{R}^m$, representing an (empirical) loss dependent on $(\boldsymbol{w}, b)$,[1] $\mathcal{F}$ is a function of $\boldsymbol{L}$, gauging the averseness of the loss vector and referred

[1] If there is no need to mention the dependence on the parameters $(\boldsymbol{w}, b)$, we denote it by $\boldsymbol{L}$ (e.g., in Section 3).

to as *risk functional* in this paper, and γ is a function regularizing $\boldsymbol{w}$. This paper employs $\boldsymbol{L}(\boldsymbol{w}, b)$ with $L_i(\boldsymbol{w}, b) = -y_i(\boldsymbol{w}^\top \boldsymbol{x}_i - b), i = 1, ..., m$.[2]

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 an "Empirical Risk" of the form: $\mathcal{F}(\boldsymbol{L}) = \frac{C}{m}\sum_{i=1}^m \max\{L_i + 1, 0\}$ or $\mathcal{F}(\boldsymbol{L}) = \frac{C}{m}\sum_{i=1}^m (\max\{L_i + 1, 0\})^2$, which are called *hinge loss*.

As for "Regularizer," the use of the square of the ℓ_2-norm (or the Euclidean norm) of 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*, other norms can be alternatives. Indeed, the ℓ_1-norm is popular since its use leads to a sparse solution, or the associated minimization can be cast as a linear program (LP) when "Empirical Risk" can also be written by a system of linear inequalities. 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 point (i.e., $\boldsymbol{x}_i$) to a hyperplane (see, e.g., Mangasarian, 1999; Pedroso and Murata, 2001).

Proposed scheme. The primary purpose of this paper is to present unified formulations of SVMs for binary classification. One of the motivations for this comes out of the pursuit of a tractable SVM formulation in which parametrized families of polyhedral norms, which are recently studied by Gotoh and Uryasev (2013), 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).[3] However, the use of the ℓ_p-norm with $p \neq 1, 2$ or ∞ increases the computational complexity from the ℓ_1-, the ℓ_2- or the ℓ_∞-norm. In contrast, complexity of the new families do not vary significantly as a function of parameters.

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 functional can affect the tractability of the resultant optimization problem and/or the resulting classifier. To address this issue, we reexamine the basic formulation of SVMs.

More precisely, in order to make the resulting optimization problem well-defined and tractable, we reconsider the empirical risk functional as well as the regularizer. On the other hand, in order to focus on tractable formulations, we limit our attention to the case where both empirical risk functionals and regularizers are convex.[4]

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 $v(z)$ is a convex function on $\mathbb{R}$ (e.g., Christmann and Steinwart, 2004; Zhang, 2004; Bartlett, Jordan, and Mcauliffe, 2006; Rifkin and Lippert, 2007; Kloft et al., 2011). This formulation, however, does not include some important convex risk functionals. Indeed, the ν-SVM (Schölkopf et al., 2000) corresponds to a non-separable risk functional $\mathcal{F}(\boldsymbol{L}) = \min_\rho\{-\rho\nu + \frac{1}{m}\sum_{i=1}^m \max\{L_i + \rho, 0\}\}$, where $\nu \in (0, 1]$. Besides, if we expand the coverage beyond the separable functionals, we can treat, for instance, the log-sum-exp functional, 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 existence of regularizer.[5] This fact indicates the importance of the combination of risk functionals and regularizers.

[2]The usage of the words 'loss' and 'risk' is slightly different from the usual convention of machine learning literature. In this paper, 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 a correct classification. On the other hand, the 'risk' reflects the attitude towards the loss vector.

[3]Kloft et al. (2011) studies a kernel learning method based on the tuning of the ℓ_p-norms. Although kernel learning is beyond the scope of this paper, the replacement of the ℓ_p-norm with a parametrized family of polyhedral norms is similar to the standard SVM formulations.

[4]On the other hand, it is fair to mention that we exclude the nonconvexity which appears in SVMs. Among such are the ramp loss (e.g., Collobert et al., 2006; Tsyurmasto, Gotoh, and Uryasev, 2013), nonconvex regularizers (e.g., Huang et al., 2009), and geometric margins (Gotoh and Takeda, 2005; Gotoh, Takeda, and Yamamoto, 2013; Tsyurmasto, Gotoh, and Uryasev, 2013). However, we should emphasize that the convexity is a basic tool for those nonconvex extensions. First of all, the convexity provides a computationally tractable framework (e.g., Boyd and Vandenberghe, 2004). Second, nonconvex problems can often be well approximated by convex optimization (e.g., Candés and Plan, 2009) or approached by iterative applications of convex optimization (e.g., An and Tao, 2005).

[5]See Section 4.1 for a detailed discussion.

One may think that the treatment of such non-separable functionals is marginal. However, a class of non-separable functionals, which are further generalized in the current paper, includes several existing formulations as special cases recently studied in Kanamori, Takeda, and Suzuki (2013). Besides, their paper shows statistical consistency. This fact indicates that the generalized formulation developed in the current paper is, at least, partly justified in a statistical way. Furthermore, the ERM of such a non-separable risk functional is shown to be connected to the minimization of the φ-divergence (or, originally, f-divergence) (Csiszár, 1967).

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). In particular, 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 results in a standard setting just by applying established patterns of functional operations. Indeed, we can enjoy the merit more than other researches since we treat both norms and empirical risk functionals in a more general manner.

Another novelty is the employment of the *convex risk functional 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 functional theory enriches the understanding of the SVM formulations. In particular, based on the theory, we show that if the risk functional is monotonic and translation invariant, some geometric or probabilistic interpretations of the dual formulations are attained (Figure 1). This aspect can be viewed as an extension of the geometric interpretation discussed in the existing papers (e.g., Crisp and Burges, 2000; Bennett and Bredensteiner, 2000; Takeda, Mitsugi, and Kanamori, 2013; Kanamori, Takeda, and Suzuki, 2013). As mentioned, a merit of our approach is a mechanical derivation based on the Fenchel duality, even with non-ℓ_2-regularizers. In addition, gaining different interpretation enriches the theory itself. 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).

Furthermore, the theory tells which properties of the risk functionals are essential to robust optimization modelings. Xu, Caramanis, and Mannor (2009) show that a regularized hinge loss minimization can be viewed as a robust optimization of the (unregularized) ERM. While they employ the hinge loss as the empirical risk and a restricted uncertainty set, our framework deals with monotonic and translation invariant risk functionals and a larger uncertainty set. It is noteworthy that 1) both approaches induce the same so-called Tikhonov-type regularizer if the same norm is employed for defining uncertainty sets. In this sense, the use of general norm in the Tikhonov regularization is just viewed as a consequence of the choice of uncertainty set; 2) with our framework, the same regularizer is derived with various risk functionals in a simpler way.[6]

In addition, we demonstrate that with monotonic and translation invariant risk functionals, our framework can straightforward treat another type of robust optimization modeling, called the distributionally optimization. This robust modeling assumes uncertainty in the probability measures, while the aforementioned one assumes uncertainty in the support (or observed values of samples). It is usual to define the empirical risk with the uniform empirical probabilities. This fits the i.i.d.-sample assumption, which is, however, often violated in reality. To cope with this, we can consider to robustify the optimization problem against the deviation from the uniform assumption. We show that such robustified formulations can be established for a class of risk functionals within convex optimization.

On the other hand, the use of non-ℓ_2-norm regularizer may induce an incongruence of the representer theorem. With the generalized formulations, we evaluate the closeness between a non-ℓ_2-norm-based SVM and the ℓ_2-norm-based SVM. This provides a guidance for the choice of 'Regularizer.' In appendix, we add some remarks for the extension to kernelization.

Our formulation extends various existing formulations and allows for easy customizations. A straightforward merit of the use of the axiomatized class of functions is to systematically obtain interpretations

[6] A similar generalization is also considered by Livni, Crammer, and Globerson (2012) on the basis of a probabilistic interpretation. However, their formulation cannot deal with positively homogeneous risk functionals due to a reason which will be discussed in Section 4.3.

of the functions based on users' preference on empirical errors.[7] For example, if a user wants to make a credit scoring model, he/she can apply a favorable risk functional (e.g., mean-variance (Markowitz, 1952)) so as to increase the model's accountability. In general, enhancing interpretability deepens users' comprehension of SVMs and promotes further customization of SVMs.

Last, presentation based on convex risk functionals and norms which are general, but defined with elementary operations fits a recent trend of optimization software packages. Indeed, various convex functionals 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 those softwares, users can easily customize SVMs. Besides, such a presentation has a potential of application of recently developing algorithms such as indifferentiable optimization and stochastic optimization.

Relations and difference from an existing research Let us mention several related papers. Gotoh and Takeda (2005) apply CVaR to the geometric margin-based classification, and interpret ν-SVM (Schölkopf et al., 2000) as a CVaR minimization. Takeda and Sugiyama (2008) point out the equivalence between the nonconvex formulation of Gotoh and Takeda (2005) and Eν-SVM (Perez-Cruz et al., 2003), and construct a global optimization algorithm to solve the formulation. Gotoh, Takeda, and Yamamoto (2013) extends CVaR to the coherent risk measures (Artzner et al., 1999), a subclass of convex risk functionals, while preserving the nonconvexity in the formulation. Also, Tsyurmasto, Gotoh, and Uryasev (2013) explore positively homogeneous risk functionals, which is not necessarily convex. In contrast to the above papers, our paper disregards the nonconvexity, while generalizing the class of risk functionals.[8] Regarding the use of risk functionals in SVM, Xu et al. (2009) mention the axioms of risk functionals in relation to their robust optimization based formulation. Takeda, Mitsugi, and Kanamori (2013) propose a unified view using the presentation with uncertainty sets, although that paper does not relate to risk functionals. A recent paper of Kanamori, Takeda, and Suzuki (2013) studies a duality correspondence between empirical risk functions and uncertainty sets, and share parts of formulations 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 theory of convex risk functionals.

As for the use of general norms, numerous papers deal with non-ℓ_2-norms. Among such, Zhou, Zhang, and Jiao (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 another LP-representable norms, which include the ℓ_1- and the ℓ_∞-norms as special limiting cases.

The structure of this paper is as follows. Section 2 briefly overviews formulations of the SVM for binary classification. Section 3 summarizes several results of the theory of convex risk functionals. Section 4 explores the unified scheme for the convex risk functional-based SVM, which mainly aims to include a general polyhedral norms, while keeping the tractability. Section 6 explains the LP-representable SVM by employing two pairs of families of LP-representable norms. Some numerical examples also demonstrate how the theory works and regularizer tuning based on the LP-norms is performed. Section 7 concludes the paper. In appendix, we summarize some remarks on the use of general norms for kernelized SVMs, as well as proofs of theorems, a list of various risk functionals, and examples of LP formulations.

Notations. $\lfloor x \rfloor$ denotes the integer no greater than x, and $(x)_+ := \max\{x, 0\}$. $\mathrm{sign}(x) := +1$ if $x \geq 0$, and $\mathrm{sign}(x) := -1$ otherwise. A vector in $\mathbb{R}^n$ is denoted in boldface and is written as a column vector in the inner products. $\mathbf{1}_n$ is the column vector in $\mathbb{R}^n$ with all components equal to 1, and $\mathbf{0}$ is the vector whose elements are equal to 0. The superscript '$\top$' denotes the transpose of vectors and matrices (e.g., $\boldsymbol{x}^\top = (x_1, ..., x_n)$). $\boldsymbol{x} \geq \boldsymbol{y}$ denotes $x_i \geq y_i, i = 1, ..., m$. We denote by $\mathbb{\Pi}^m$ the unit simplex in $\mathbb{R}^m$ (or equivalently, the set of probability measures), i.e., $\mathbb{\Pi}^m := \{\boldsymbol{p} \in \mathbb{R}^m : \mathbf{1}_m^\top \boldsymbol{p} = 1, \boldsymbol{p} \geq \mathbf{0}\}$, and by $\mathbb{\Pi}_+^m$ the intersection of $\mathbb{\Pi}^m$ and the positive orthant, i.e., $\mathbb{\Pi}_+^m := \{\boldsymbol{p} \in \mathbb{R}^m : \mathbf{1}_m^\top \boldsymbol{p} = 1, p_i > 0, i = 1, ..., m\}$. On the other hand, $\mathbb{R}_+ := [0, +\infty)$. For a set C, its relative interior is denoted by $\mathrm{ri}(C)$. For example,

[7] Indeed, the (convex) risk functional theory has been developed in financial literature (e.g., Artzner et al., 1999) on the basis of axioms for the risk attitude of decision makers (e.g., investors, banks, etc.).

[8] Convex risk functionals treated in this paper is not limited to be monotonic. In that sense, our treatment is more related to Ruszczyński and Shapiro (2005), rather than Föllmer and Schied (2002).

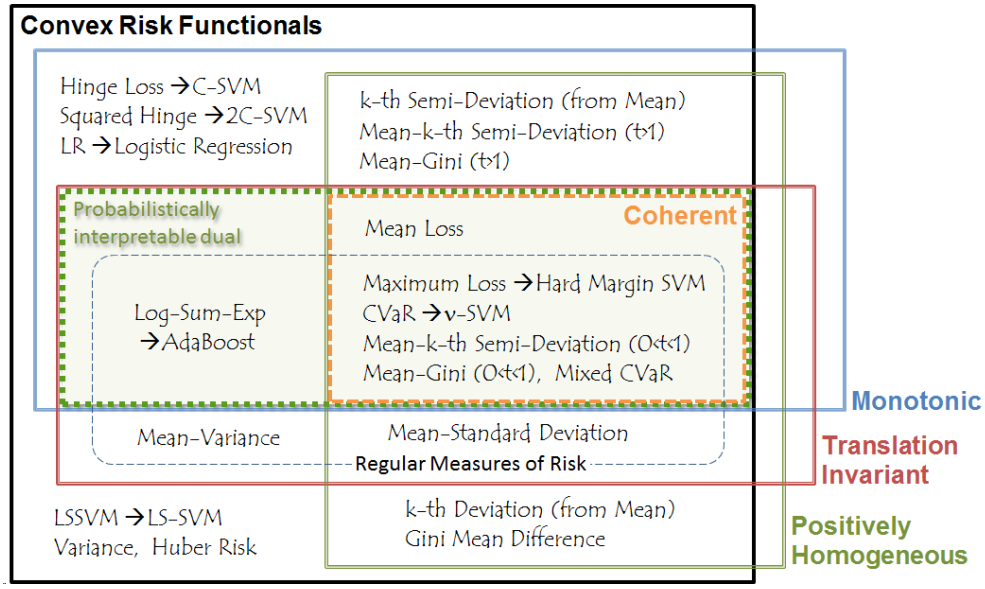


Figure 1: Classification of convex risk functionals and corresponding SVMs

Risk functionals are related to existing SVMs. We categorize the convex risk functionals on the basis of properties: monotonicity, translation invariance and positive homogeneity. Each property can be characterized in a dual representation. In particular, monotonic and translation invariant risk functionals are shown to be associated with geometric and probabilistic interpretations. Readers who are interested in a (partial) list of those risk functionals can jump to Section 3 and A.2.

$\mathbb{\Pi}_+^m = \mathrm{ri}(\mathbb{\Pi}^m)$. 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. The 0-1 indicator for a condition $c(\cdot)$ on $\mathbb{R}$ is denoted by $[c(\cdot)]$, i.e., $[c(x)] = 1$ if $c(x)$ is true; 0 otherwise. As a convention inspired by MATLAB, we denote by $|\boldsymbol{x}|$ the vector whose i-th element is given by $|x_i|$, i.e., $|\boldsymbol{x}| := (|x_1|, ..., |x_n|)^\top$ for $\boldsymbol{x} \in \mathbb{R}^n$. Also, we apply plus, logarithmic, exponential and 0-1 indicator operations to a vector, respectively, as follows: $(\boldsymbol{x})_+ := ((x_1)_+, ..., (x_n)_+)^\top$, $\ln(\boldsymbol{x}) = (\ln x_1, ..., \ln x_n)^\top$, $\exp(\boldsymbol{x}) = (\exp x_1, ..., \exp x_n)^\top$ and $[\boldsymbol{x} \geq \boldsymbol{y}] := ([x_1 \geq y_1], ..., [x_n \geq y_n])^\top$. Besides, we employ the notation '$./$' and '$\cdot^k$' for component-wise division of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and power, respectively, i.e., $\boldsymbol{x}./\boldsymbol{y} = (x_1/y_1, ..., x_n/y_n)^\top$ and $\boldsymbol{x}^{\cdot k} = (x_1^k, ..., x_n^k)^\top$. Matrices are denoted also by boldface. In particular, we denote by $\boldsymbol{I}_n$ the $n \times n$ unit matrix, and by $\boldsymbol{E}_n$ the $n \times n$ matrix whose all elements are 1, i.e., $\boldsymbol{E}_n \equiv \mathbf{1}_n \mathbf{1}_n^\top$. We denote by $\mathrm{diag}\,(\boldsymbol{x})$ the square matrix whose diagonal elements are given by $\boldsymbol{x}$ and off-diagonal elements are all 0. Throughout the paper we denote the ℓ_2-norm by $\|\cdot\|_2$, while the ℓ_1- and ℓ_∞-norms by $\|\cdot\|_1$ and $\|\cdot\|_\infty$, respectively. The notation $\|\cdot\|$ is reserved for any norm in $\mathbb{R}^n$. Especially, $\langle\!\langle \cdot \rangle\!\rangle_\alpha$ and $(\!(\cdot)\!)_\tau$ are reserved, respectively, for the CVaR norm and the deltoidal norm, i.e., the convex combination of the ℓ_1- and ℓ_∞-norms. The asterisk '$*$' attached to a norm denotes the dual norm to a norm, i.e., $\|\boldsymbol{x}\|^\circ := \max_{\boldsymbol{z}}\{\boldsymbol{x}^\top \boldsymbol{z} : \|\boldsymbol{z}\| \leq 1\}$, whereas the asterisk '$*$' attached to a function (except any norm) indicates the conjugate of it, i.e., $f^*(\boldsymbol{x}) := \sup_{\boldsymbol{y}}\{\boldsymbol{x}^\top \boldsymbol{y} - f(\boldsymbol{y})\}$. Some operations are sometimes represented in a symbolical manner. For example, we denote by $\mathbb{E}_{\boldsymbol{p}}(\cdot)$ the mathematical expectation operator where $\boldsymbol{p}$ denotes the probability measure used, i.e., $\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{x}) := \boldsymbol{p}^\top \boldsymbol{x}$.

2 A brief overview of Support Vector Machine (SVM) formulations

This section introduces notations and briefly overviews popular SVM formulations for binary classification for later reference.

Suppose that a data set $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)\}$ is given, where $\boldsymbol{x}_i \in \mathbb{R}^n$ denotes the attributes of sample i and y_i describes its binary label, $i = 1, ..., m$. For convenience, the label is given by either $+1$ or -1.

SVM finds a hyperplane $\{\boldsymbol{x} : \boldsymbol{w}^\top \boldsymbol{x} = b\}$ for separating unseen samples with labels $+1$ and -1 (e.g., $\boldsymbol{x}_{m+1}, ..., \boldsymbol{x}_\ell$). The objective is to separate as much as possible points in a systematic, nonparametric manner.[9]

For example, let us consider the so-called C-SVM (Cortes and Vapnik, 1995). Letting $\boldsymbol{Y} := \mathrm{diag}(y_1, ..., y_m)$ and $\boldsymbol{X} := (\boldsymbol{x}_1, ..., \boldsymbol{x}_m)^\top$, C-SVM for the binary classification is formulated with the following (convex) quadratic programming (QP) problem:

$$\bar{p}^\star := \left| \begin{array}{ll} \underset{\boldsymbol{w}, b, \boldsymbol{z}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{C}{m}\mathbf{1}_m^\top \boldsymbol{z} \\ \text{subject to} & \boldsymbol{z} \geq -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b) + \mathbf{1}_m,\ \boldsymbol{z} \geq \mathbf{0}, \end{array} \right. \tag{1}$$

where $C > 0$ is a user-defined parameter. As sketched in Introduction, Problem (1) can be equivalently presented as the following bi-objective minimization problem:

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

With an optimal solution $(\boldsymbol{w}^\star, b^\star, \boldsymbol{z}^\star)$ to (1), the decision function is defined by

$$d(\boldsymbol{x}) := \mathrm{sign}(\boldsymbol{x}^\top \boldsymbol{w}^\star - b^\star). \tag{2}$$

Dual SVM formulations are frequently considered. For example, the dual problem to (1) is given by another QP:

$$\bar{d}^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\frac{1}{2}\|\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}\|_2^2 + \mathbf{1}_m^\top \boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0,\ \mathbf{0} \leq \boldsymbol{\lambda} \leq \frac{C}{m}\mathbf{1}_m. \end{array} \right. \tag{3}$$

Strong duality between (1) and (3) holds under a mild condition, i.e., $\bar{p}^\star = \bar{d}^\star$. More importantly, with the optimality condition, we have

$$\boldsymbol{w}^\star = \boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}^\star, \tag{4}$$

where $\boldsymbol{\lambda}^\star$ is an optimal solution to (3). 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 (4) into (2), the decision function can be rewritten with the optimal dual variables $\boldsymbol{\lambda}^\star$:

$$d(\boldsymbol{x}) := \mathrm{sign}(\boldsymbol{x}^\top \boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}^\star - b^\star).^{10} \tag{5}$$

The *ν-SVM* (Schölkopf et al., 2000) solves another QP in place of (1):

$$\tilde{p}^\star := \left| \begin{array}{ll} \underset{\boldsymbol{w}, b, \rho, \boldsymbol{z}}{\text{minimize}} & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - \rho + \frac{1}{m\nu}\mathbf{1}_m^\top \boldsymbol{z} \\ \text{subject to} & \boldsymbol{z} \geq -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b) + \mathbf{1}_m \rho,\ \boldsymbol{z} \geq \mathbf{0}, \end{array} \right. \tag{6}$$

Similarly to C-SVM, (6) can be viewed as a regularized ERM:

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

[9] As in the usual convention of SVMs, we can extend the discussion below into a nonlinear case by considering a (possibly, implicit) mapping $\phi : \mathbb{R}^n \to \mathbb{R}^N$. For simplicity, however, we describe the method on the basis of the linear classification through Sections 2 to 6.

The dual formulation to (6) is given by another QP:

$$\tilde{d}^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\frac{1}{2}\|\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}\|_2^2 \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0,\ \mathbf{1}_m^\top \boldsymbol{\lambda} = 1,\ \mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}_m/m. \end{array} \right. \tag{7}$$

Let $(\boldsymbol{w}^\star, b^\star, \rho^\star, \boldsymbol{z}^\star)$ and $\boldsymbol{\lambda}^\star$ be optimal solutions to (6) and (7), respectively. Similarly to the C-SVM, under a mild condition, we have the strong duality between (6) and (7), i.e., $\tilde{p}^\star = \tilde{d}^\star$, and the parallelism (4) between $\boldsymbol{w}^\star$ and $\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}^\star$, again. Also, we can obtain the decision function (5) on the basis of a dual solution $\boldsymbol{\lambda}^\star$ to (7) (although there can be multiple ways for determining $b^\star$).[11]

In order to develop a general scheme, we next overview the theory of convex risk functionals, which will be employed (mainly) as empirical risk terms.

3 Theory of convex risk functionals

This section overviews the theory of convex functionals, especially focusing on their duality representation which will be exploited in the proposed SVM formulations.

3.1 Basic properties and examples of convex risk functionals

Let us consider uncertain outcomes which are represented as a random variable (e.g., denoted $\boldsymbol{L}$) on a sample space Ω (equipped with a sigma algebra). Without loss of generality, we assume that the random variable represents a quantity which is preferable as it is smaller, and we shall figuratively call it *(random) loss*. For simplicity, we limit our attention to a finite sample space, i.e., $\Omega := \{\omega_1, ..., \omega_m\}$ with $m < \infty$, and accordingly, we can identify each random variable as a vector in $\mathbb{R}^m$. Associated with this sample space, we consider a probability measure $\boldsymbol{p} \in \mathbb{\Pi}^m$, i.e., $p_i = \mathbb{P}(\omega_i)$, and referred to it as *reference probability*. Although the uniform probability measure, i.e., $\boldsymbol{p} = \mathbf{1}_m/m$, is often employed in machine learning context, it is not limited in this paper and user can put his/her own information into $\boldsymbol{p}$.

This paper employs the negative of the *margin* as the loss $\boldsymbol{L}$, since a wide range of binary classification methods are covered with it. In order to represent the averseness to a random loss (vector) by a single number, we define a function referred to as *risk functional*, $\mathcal{F}$, i.e., $\mathcal{F} : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$. Note that $\mathcal{F}$ can take '$+\infty$.' Let us denote by '$\mathrm{dom}\,\mathcal{F}$' the *effective domain* of a function $\mathcal{F}$, i.e., $\mathrm{dom}\,\mathcal{F} := \{\boldsymbol{L} \in \mathbb{R}^m : \mathcal{F}(\boldsymbol{L}) < +\infty\}$.

We assume that the risk functional $\mathcal{F}$ is proper and lower semi-continuous (l.s.c.) to make it appropriate for minimization.

- $\mathcal{F}$ is *proper* if $\mathrm{dom}\,\mathcal{F} \neq \emptyset$.
- $\mathcal{F}$ is *lower semi-continuous (l.s.c.)* if $\mathcal{F}(\boldsymbol{L}) \leq \lim_{i\to\infty} \mathcal{F}(\boldsymbol{L}_i)$ for any $\boldsymbol{L} \in \mathbb{R}^m$ and any sequence $\boldsymbol{L}_1, \boldsymbol{L}_2, ... \in \mathbb{R}^m$ converging to $\boldsymbol{L}$.

Considering the tractability of minimization which appears in many of SVMs, we limit our attention only to convex risk functionals.

- $\mathcal{F}$ 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 $\boldsymbol{L}, \boldsymbol{L}' \in \mathbb{R}^m,\ \tau \in (0,1)$.

Note that if $\mathcal{F}$ is proper convex, $\mathrm{dom}\,\mathcal{F}$ is a nonempty convex set. As below, most of risk functionals are defined with a reference probability $\boldsymbol{p}$. Therefore, we sometimes indicate the dependency of a risk functional $\mathcal{F}$ on $\boldsymbol{p}$ by denoting as $\mathcal{F}_{\boldsymbol{p}}$. Also, throughout the paper, we assume that $\boldsymbol{p}$ satisfies $p_i > 0$ for all $i = 1, ..., m$.

Let us give some examples of convex risk functionals in binary classification.

- $\mathcal{F}(\boldsymbol{L}) = \mathrm{Hinge1}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := t\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L} + \mathbf{1}_m)_+$: (1-norm) Hinge loss-based;

[11] Namely, for an optimal solution $\boldsymbol{\lambda}^\star$ and any samples $\boldsymbol{x}_i$ with $y_i = +1$ and $\boldsymbol{x}_{i'}$ with $y_{i'} = -1$ satisfying $\lambda_i^\star, \lambda_{i'}^\star \in (0, \frac{1}{m\nu})$, we can compute $b^\star$ by $b^\star = \frac{1}{2}(\boldsymbol{x}_i + \boldsymbol{x}_{i'})^\top \boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}^\star$.

- $\mathcal{F}(\boldsymbol{L}) = \mathrm{Hinge2}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{t}{2}\mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L}+\mathbf{1}_m)_+^{\cdot 2})$: squared (2-norm) Hinge loss-based;
- $\mathcal{F}(\boldsymbol{L}) = \mathrm{LSSVM}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{t}{2}\mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L}+\mathbf{1}_m)^{\cdot 2})$: Least Square SVM-based;
- $\mathcal{F}(\boldsymbol{L}) = \mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) := \min_c\{c + \frac{1}{1-\alpha}\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}-c\mathbf{1}_m)_+\}$: CVaR (conditional value-at-risk (Rockafellar and Uryasev, 2000));
- $\mathcal{F}(\boldsymbol{L}) = \mathrm{LR}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\ln(\mathbf{1}_m + \exp(\frac{1}{t}\boldsymbol{L})))$: Logistic Regression-based;
- $\mathcal{F}(\boldsymbol{L}) = \mathrm{LSE}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{1}{t}\ln \mathbb{E}_{\boldsymbol{p}}(\exp(t\boldsymbol{L}))$: Log-Sum-Exp;
- $\mathcal{F}(\boldsymbol{L}) = \mathrm{SE}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \frac{1}{t}\mathbb{E}_{\boldsymbol{p}}(\exp(t\boldsymbol{L}))$: Sum-Exp;
- $\mathcal{F}(\boldsymbol{L}) = \mathrm{MV}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + \frac{t}{2}\mathbb{V}_{\boldsymbol{p}}(\boldsymbol{L})$: Mean-Variance,

where $t > 0$ and $\alpha \in [0, 1)$. See Appendix A for the additional explanation of CVaR (Section A.1) and a list of the other risk functionals which have potential to be useful (Section A.2).

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 $\mathcal{F}(\boldsymbol{L}+\tau\mathbf{1}_m) = \mathcal{F}(\boldsymbol{L}) + \tau$ 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 functional satisfying the above three properties is said to be *coherent*.

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

See Artzner et al. (1999) for the interpretation of these properties in financial context.

In the later part of this paper, we show that these properties play a certain role also in SVM formulations. For later reference, we observe the following facts.

Proposition 1. *For a risk functional $\mathcal{V}$, let us define another risk functional of the form:*

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

Then, $\mathcal{F}$ is proper if so is $\mathcal{V}$, and for any $\boldsymbol{L}$, the infimum of (8) does not attain $-\infty$; $\mathcal{F}$ is l.s.c. if so is $\mathcal{V}$; $\mathcal{F}$ is convex if so is $\mathcal{V}$; $\mathcal{F}$ is monotonic if so is $\mathcal{V}$; $\mathcal{F}$ is translation invariant for any risk functional $\mathcal{V}$; $\mathcal{F}$ is positively homogeneous if so is $\mathcal{V}$.

Proof. Since $\mathcal{V}$ is proper, there is at least one $\boldsymbol{L}$ such that $\mathcal{F}(\boldsymbol{L}) < +\infty$. Therefore, if the infimum is above $-\infty$, $\mathcal{F}$ of the form (8) is proper. The other propositions are derived straightforward from the definition of each property. Thus, we omit the details. □

We should notice that the convexity of $\mathcal{V}$ is not necessary for the translation invariance. Indeed, we can view the formula (8) as an operation for making any risk functional translation invariant.[12]

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

$$\mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L}) = \inf_c\{c + \mathbb{E}_{\boldsymbol{p}}(v(\boldsymbol{L}-c\mathbf{1}_m))\}. \tag{9}$$

As an easy extension of Proposition 1, we can confirm the following properties.

[12] Rockafellar and Uryasev (2013) define the *regular measure of risk* by an l.s.c. convex risk functional $\mathcal{F}$ satisfying

- $\mathcal{F}(\tau\mathbf{1}_m) = \tau$ 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 $\boldsymbol{L} = \tau\mathbf{1}_m$ 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 functional 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 functional' term instead.

Corollary 1. *Functional (9) is translation invariant for any v. (9) is proper if so is v, and for any $\boldsymbol{L}$, the infimum in (9) does not attain $-\infty$; (9) is l.s.c. if so is v; (9) is convex if so is v; (9) is monotonic if so is v; (9) is positively homogeneous if so is v.*

For example, if v satisfies $v(z) \geq z$ for any z, (9) is proper. Some inseparable-form risk functionals such as CVaR and LSE, which appear in ν-SVM and AdaBoost (Freund and Schapire, 1997), respectively, can be represented with $v(z) = \frac{1}{1-\alpha}(z)_+$ and $\exp(z) - 1$, each satisfying the condition. In addition, via the formula (9), no-translation invariant functionals such as Hinge1, Hinge2 and LSSVM can be transformed into translation invariant ones. For example, by employing $v(z) = (1+z)_+/t$ in (9), $\mathrm{Hinge1}_{(t,\boldsymbol{p})}$ is transformed to a translation invariant risk functional $\mathrm{Hinge1}^{\mathrm{OCE}}_{(t,\boldsymbol{p})} = \inf_c\{c + \frac{1}{t}\boldsymbol{p}^\top[((1-c)\mathbf{1}_m + \boldsymbol{L})_+]\}$. Note that this is equal to $1 + \mathrm{CVaR}_{(1-t,\boldsymbol{p})}(\boldsymbol{L})$. Namely, CVaR can be considered as Hinge1 transformed by (8).

Such transformed functionals are shown to be related to the uncertainty set-based representation of SVMs. Indeed, Kanamori, Takeda, and Suzuki (2013) consider an SVM formulation which employs the risk functional of the form (9) with $\boldsymbol{p} = (2/m)\mathbf{1}_m$.[13] An extension based on the above propositions will be discussed in Section 5.1.

In addition to the aforementioned properties, we introduce another property related to the tractability of resulting optimization for classification task.

- $\mathcal{F}$ is *LP-representable* if $\mathcal{F}(\boldsymbol{L})$ is represented by a Linear Program, i.e., it is a minimum of a linear function over a set of linear constraints.

Table 1 summarizes the properties of the risk functionals listed above.

Table 1: Properties of convex risk functionals examples

$\mathcal{F}$	mono-tonic	trans. invar.	posi. homo.	coher-ent	reg.meas. risk	LP-rep.
$\mathrm{Hinge1}_{(t,\boldsymbol{p})}$	yes	no	no	no	no	yes
$\mathrm{Hinge2}_{(t,\boldsymbol{p})}$	yes	no	no	no	no	no
$\mathrm{LSSVM}_{(t,\boldsymbol{p})}$	no	no	no	no	no	no
$\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}$	yes	yes	yes	yes	yes	yes
$\mathrm{LR}_{(t,\boldsymbol{p})}$	yes	no	no	no	no	no
$\mathrm{LSE}_{(t,\boldsymbol{p})}$	yes	yes	no	no	yes	no
$\mathrm{SE}_{(t,\boldsymbol{p})}$	yes	no	no	no	no	no
$\mathrm{MV}_{(t,\boldsymbol{p})}$	no	yes	no	no	yes	no

3.2 Duality for convex risk functionals

As overviewed in Section 2, duality plays an important role in the optimization for SVMs. To deepen the theory presented below on the basis of duality, let us introduce the conjugate of a function. For a function $\mathcal{F}$ from $\mathbb{R}^n$ to $\mathbb{R} \cup \{\pm\infty\}$, its conjugate, $\mathcal{F}^*$, is defined by

$$\mathcal{F}^*(\boldsymbol{\lambda}) := \sup_{\boldsymbol{L}}\{\boldsymbol{\lambda}^\top \boldsymbol{L} - \mathcal{F}(\boldsymbol{L})\}. \tag{10}$$

The following fact is noteworthy.

Theorem 1 (Self duality of proper l.s.c. convex function (e.g., Rockafellar, 1970))**.** *For a function $\mathcal{F}$ from $\mathbb{R}^n$ to $\mathbb{R}\cup\{\pm\infty\}$, $\mathcal{F}^*$ is convex. Besides, if $\mathcal{F}$ is proper, l.s.c. and convex, so does $\mathcal{F}^*$ and $\mathcal{F}^{**} = \mathcal{F}$, and accordingly, the risk functional $\mathcal{F}$ is then also represented by*

$$\mathcal{F}(\boldsymbol{L}) = \sup_{\boldsymbol{\lambda}}\{\boldsymbol{L}^\top \boldsymbol{\lambda} - \mathcal{F}^*(\boldsymbol{\lambda})\}. \tag{11}$$

[13] One of their examples called Truncated Quadratic Loss corresponds to Hinge2 transformed by (9).

See Section 12 of Rockafellar (1970) for the proof.

Table 2 lists conjugates of the aforementioned risk functionals.

Table 2: Conjugates of convex risk functionals examples

$\mathcal{F}^*(\boldsymbol{\lambda})$	
$\mathrm{Hinge1}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= -\mathbf{1}_m^\top \boldsymbol{\lambda} + \delta_{[\mathbf{0},t\boldsymbol{p}](\lambda_i)}(\boldsymbol{\lambda}) \equiv \sum_{i=1}^m (-\lambda_i + \delta_{[0,tp_i]})$
$\mathrm{Hinge2}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= \frac{1}{2t}\boldsymbol{\lambda}^{\cdot 2}./\boldsymbol{p} - \mathbf{1}_m^\top \boldsymbol{\lambda} + \delta_{\mathbb{R}^m_+}(\boldsymbol{\lambda}) \equiv \sum_{i=1}^m (\frac{\lambda_i^2}{2tp_i} - \lambda_i + \delta_{\mathbb{R}_+}(\lambda_i))$
$\mathrm{LSSVM}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= \frac{1}{2t}\boldsymbol{\lambda}^{\cdot 2}./\boldsymbol{p} - \mathbf{1}_m^\top \boldsymbol{\lambda} \equiv \sum_{i=1}^m (\frac{\lambda_i^2}{2tp_i} - \lambda_i)$
$\mathrm{CVaR}^*_{(\alpha,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= \delta_{\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}}(\boldsymbol{\lambda})$ with $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}$ defined in (13)
$\mathrm{LR}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= \boldsymbol{p}^\top\{(t\boldsymbol{\lambda}./\boldsymbol{p})\ln(t\boldsymbol{\lambda}./\boldsymbol{p}) + (\mathbf{1}_m - t\boldsymbol{\lambda}./\boldsymbol{p})\ln(\mathbf{1}_m - t\boldsymbol{\lambda}./\boldsymbol{p})\} + \delta_{[\mathbf{0},\boldsymbol{p}/t]}(\boldsymbol{\lambda})$
	$\equiv \sum_{i=1}^m \left[p_i\left\{(\frac{t\lambda_i}{p_i})\ln(\frac{t\lambda_i}{p_i}) + (1-\frac{t\lambda_i}{p_i})\ln(1-\frac{t\lambda_i}{p_i})\right\} + \delta_{[0,\frac{p_i}{t}]}(\lambda_i)\right] := \mathrm{bitEnt}_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$
$\mathrm{LSE}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= \frac{1}{t}\boldsymbol{\lambda}^\top \ln(\boldsymbol{\lambda}./\boldsymbol{p}) + \delta_{\amalg^m}(\boldsymbol{\lambda}) \equiv \frac{1}{t}\sum_{i=1}^m \lambda_i \ln\frac{\lambda_i}{p_i} + \delta_{\amalg^m}(\boldsymbol{\lambda}) := \mathrm{KL}_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$
$\mathrm{SE}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= \frac{1}{t}\boldsymbol{\lambda}^\top(\ln(\boldsymbol{\lambda}./\boldsymbol{p}) - \mathbf{1}_m) + \delta_{\mathbb{R}^m_+}(\boldsymbol{\lambda}) \equiv \sum_{i=1}^m \{\frac{1}{t}\lambda_i(\ln\frac{\lambda_i}{p_i} - 1) + \delta_{\mathbb{R}_+}(\lambda_i)\}$
$\mathrm{MV}^*_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$	$= \frac{1}{2t}\boldsymbol{p}^\top\{(\boldsymbol{\lambda}./\boldsymbol{p})^{\cdot 2} - \mathbf{1}_m\} + \delta_C(\boldsymbol{\lambda}) \equiv \frac{1}{2t}\sum_{i=1}^m p_i\{(\frac{\lambda_i}{p_i})^2 - 1\} + \delta_C(\boldsymbol{\lambda}) =: \chi^2_{(t,\boldsymbol{p})}(\boldsymbol{\lambda})$
	with $C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \mathbf{1}_m^\top \boldsymbol{\lambda} = 1\}$

By using the conjugate, monotonicity, translation invariance and positive homogeneity can be characterized in a dual manner as follows.

Theorem 2. (**Dual characterization of risk functional properties (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 $\forall \boldsymbol{\lambda} \in \mathrm{dom}\,\mathcal{F}^*$, $\mathbf{1}_m^\top \boldsymbol{\lambda} = 1$;*
3. *$\mathcal{F}$ is positively homogeneous if and only if (11) can be represented in the form*

$$\mathcal{F}(\boldsymbol{L}) = \sup_{\boldsymbol{\lambda}}\{\boldsymbol{L}^\top \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathrm{dom}\,\mathcal{F}^*\}, \tag{12}$$

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

From Theorem 2, we see that $\mathrm{dom}\,\mathcal{F}^*$ plays an important role in characterizing risk functionals. Let us denote this effective domain by $\mathcal{Q}_{\mathcal{F}}$ and call it *risk envelope*, i.e., $\mathcal{Q}_{\mathcal{F}} = \mathrm{dom}\,\mathcal{F}^*$.[14] In particular, by combining with Theorem 2, any coherent risk functional can be characterized by a set of probability measures.

[14] 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.

Corollary 2 (Dual representation of coherent risk functionals (Artzner et al., 1999))**.** *Any coherent risk functional $\mathcal{F}$ is represented in the form (12) where $\mathcal{Q}$ is a set of probability measures. Furthermore, $\mathcal{Q}$ can be restricted to its convex hull, i.e., $\mathcal{F}(\boldsymbol{L}) = \max\{\boldsymbol{L}^\top \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathrm{conv}(\mathcal{Q})\}$ where $\mathrm{conv}(\mathcal{Q})$ denotes the convex hull of a set $\mathcal{Q}$.*

For example, CVaR is coherent and can be represented with the risk envelope

$$\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})} := \{\boldsymbol{q} \in \mathbb{\Pi}^m : \boldsymbol{q} \le \boldsymbol{p}/(1-\alpha)\}, \tag{13}$$

i.e., $\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \max_{\boldsymbol{q}}\{\mathbb{E}_{\boldsymbol{q}}(\boldsymbol{L}) : \boldsymbol{q} \in \mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}\}$.

To deepen this theoretic 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 $\varphi(1) = 0$.[15] With such φ, the φ-divergence of $\boldsymbol{q} \in \mathbb{R}^m$ relative to $\boldsymbol{p} \in \mathbb{\Pi}^m_+$ is defined by

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

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 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 φ-divergences examples.

Theorem 3. *Let v be a proper l.s.c. convex function on $\mathbb{R}$ such that $v(z) \ge z + B$ with some $B \in \mathbb{R}$. Then, the risk functional (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,*

$$\mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L}) = \inf_c\{c + \mathbb{E}_{\boldsymbol{p}}(v(\boldsymbol{L} - c\mathbf{1}_m))\} = \sup_{\boldsymbol{q}}\{\mathbb{E}_{\boldsymbol{q}}(\boldsymbol{L}) - \mathcal{I}_{v^*}(\boldsymbol{q},\boldsymbol{p})\}. \tag{14}$$

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 Section B.1 for the proof.

Formula (14) indicates that the risk functional of the form (9) is interpreted as a worst-case expected loss which is deducted with the φ-divergence where $\varphi = v^*$.[16] In this view, we can associate each coherent risk functional with a φ-divergence which is represented as the indicator function of a (closed convex) set (see, e.g., Ben-Tal and Teboulle, 2007).

Table 3 demonstrates the correspondence of $\mathcal{F}_{\boldsymbol{p}}$, $\mathcal{I}_{v^*}(\boldsymbol{q},\boldsymbol{p})$, v^* and v for CVaR, LSE and MV risk functionals. 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 functionals in Table 3 can be related to the divergences relative to $\boldsymbol{p}$.

4 Convex risk functional-based SVMs for linear binary classification

This section provides primal and dual formulations for convex risk functional-based SVMs in a unified manner on the basis of convex analysis.

[15] We can limit $\mathrm{dom}\varphi$ to be nonnegative for measuring the distance of (probability) measure. However, for a generality, we here do not. This treatment is common with Ben-Tal and Teboulle (2007). Indeed, the nonnegativity is related to the monotonicity of φ^*, as will be stated in Theorem 3.

[16] See Ben-Tal and Teboulle (2007) for an interpretation as the Optimized Certainty Equivalent based on a concave utility function $u(z) := -v(-z)$.

Table 3: Examples of risk functionals $\mathcal{F}_{\boldsymbol{p}}$ satisfying (14) and their $\mathcal{I}_{v^*}$, v^* and v

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

4.1 Formulations and duality of convex risk functional-based SVMs

A pair of primal and dual formulations. Let $\mathcal{F}(\cdot)$ be a risk functional that is proper, l.s.c. and convex on $\mathbb{R}^m$. We consider the empirical risk which is defined with $\mathcal{F}$ on some loss $\boldsymbol{L}$ associated with discriminating hyperplane. In this paper, we suppose that the loss $\boldsymbol{L}$ is represented by a linear function with respect to parameters, which are decision variables in primal formulations. In particular, we consider the loss of the following form

$$\boldsymbol{L} = -(\boldsymbol{G}^\top \boldsymbol{w} - \boldsymbol{y}b), \tag{15}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. With $\boldsymbol{G} = \boldsymbol{Y}\boldsymbol{X}$, (15) becomes

$$\boldsymbol{L} = -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}_m b) \quad \leftrightarrow \quad L_i = -y_i(\boldsymbol{x}_i^\top \boldsymbol{w} - b),\ i = 1, ..., m. \tag{16}$$

Namely, the negative of the margin is employed as the loss.

We should note that since $\boldsymbol{L}$ is linear in parameters $(\boldsymbol{w}, b)$, $\mathcal{F}(-(\boldsymbol{G}^\top \boldsymbol{w} - \boldsymbol{y}b))$ is convex (positively homogeneous) with respective to $(\boldsymbol{w}, b)$ if $\mathcal{F}$ is a convex (positively homogeneous, resp.) risk functional. On the other hand, the monotonicity or the translation invariance is not preserved, in general.

It should be noted that the so-called kernelized representation of the margin-based loss can also be rewritten by a linear function with respect to involved parameters. Let $k(x, z)$ be a kernel function and let $\boldsymbol{K} = (y_i y_j k(x_i, x_j))_{i,j=1,...,m} \in \mathbb{R}^{m\times m}$. With substitution $\boldsymbol{K} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\alpha} \in \mathbb{R}^m$ in place of $\boldsymbol{G} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$, respectively, the loss (15) becomes

$$\boldsymbol{L} = -(\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}b) \quad \leftrightarrow \quad L_i = -y_i(\sum_{h=1}^{m} y_h k(\boldsymbol{x}_h, \boldsymbol{x}_i)\alpha_h - b),\ i = 1, ..., m. \tag{17}$$

The restriction to linear function (15) may seem to excessively limit the generality. However, it is noteworthy that we are still able to incorporate various nonlinear empirical risk by carefully choosing a nonlinear convex risk functionals. In addition, a merit of limiting to linear loss can be found in the relation to the use of the so-called (geometric) margin. Indeed, we can show that the maximal margin criteria (or hard margin SVM) corresponds to the max functional[17] (see Gotoh, Takeda, and Yamamoto, 2013, for the details).

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

$$p^\star := \inf_{\boldsymbol{w},b}\ \mathcal{F}(-(\boldsymbol{G}^\top \boldsymbol{w} - \boldsymbol{y}b)) + \gamma(\boldsymbol{w}). \tag{18}$$

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

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

Note that if γ is a regularizer, so is γ^* (Rifkin and Lippert, 2007).[18] In particular, we below consider the

[17] See Appendix A.1 for the definition and the relation to CVaR.

[18] Such regularizer include the simple regularizers listed below, and it also can represent more sophisticated regularizers. For example, $\gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \delta_{\mathbb{R}^n_+}(\boldsymbol{w})$ represents a regularizer, explicitly given by

$$\gamma(\boldsymbol{w}) = \begin{cases} \frac{1}{2}\|\boldsymbol{w}\|_2^2, & \text{if } \boldsymbol{w} \geq \boldsymbol{0}, \\ +\infty, & \text{otherwise.} \end{cases}$$

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 $\iota : \mathbb{R}_+ \to [0, +\infty]$ is non-decreasing and convex. In the following, we pay special attention to the following three regularizers.

- $\gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 \leftrightarrow \gamma^*(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$;
- $\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\| \leftrightarrow \gamma^*(\boldsymbol{w}) = \delta_{\|\cdot\|^\circ \leq 1}(\boldsymbol{w})$;
- $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\| \leq 1}(\boldsymbol{w}) \leftrightarrow \gamma^*(\boldsymbol{w}) = \|\boldsymbol{w}\|^\circ$,

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

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

Note that $\|\cdot\|$ denotes an arbitrary norm and $\|\cdot\|^\circ$ denotes its dual norm, while $\|\cdot\|_2$ denotes the ℓ_2-norm.

The first and second cases are categorized as the Tikhonov regularization, while the third one is categorized as the Ivanov regularizer. 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 risk functional employed. This will be analyzed in Section 4.3.

In spite of the fact that we restrict the loss to be linear with respect to parameters, the general formulations cover a variety of optimization problem formulations for binary classification.

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

Here we emphasize that (18) covers non-decomposable empirical risk functionals such as CVaR and LSE.

The dual problem to problem (18) is

$$d^\star := \sup_{\boldsymbol{\lambda}} \quad -\gamma^*(\boldsymbol{G}\boldsymbol{\lambda}) - \mathcal{F}^*(\boldsymbol{\lambda}) - \delta_{\boldsymbol{y}^\top(\cdot)=0}(\boldsymbol{\lambda}). \tag{19}$$

Given a pair of the two formulations (18) and (19), we can prove the weak and strong duality theorems, as follows.

Proposition 2 (Weak duality)**.** *The weak duality holds between (18) and (19), i.e., we have* $p^\star \geq d^\star$.

Proof. Straightforward from the Fenchel's inequality (see Sections 12 and 31 of Rockafellar, 1970). □

Theorem 4 (Strong duality)**.** *The strong duality holds between (18) and (19), 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 $\boldsymbol{w} \in \mathrm{ri}(\mathrm{dom}\,\gamma)$ and $-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y} b \in \mathrm{ri}(\mathrm{dom}\,\mathcal{F})$.*

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

*Under (a), the supremum in (19) is attained at some $\boldsymbol{\lambda}$, while under (b), the infimum in (18) is attained at some $(\boldsymbol{w}, b)$. In addition, if $\mathcal{F}$ (or equivalently, $\mathcal{F}^*$) is polyhedral, "*ri*" can be omitted.*

Theorem 4 can be obtained from the Fenchel-Rockafellar duality theorem. See Appendix B.2 for the details.

4.2 Correspondence between risk functional properties and dual formulations

Next let us consider relations between the property of the empirical risk function $\mathcal{F}$ and the dual formulation (19).

Based on Theorem 12, we can associate the constraints of dual formulations with functional properties of the risk functionals $\mathcal{F}$ employed in the primal formulations.

Proposition 3. *1. If $\mathcal{F}$ is monotonic, the dual problem (19) can be represented as*

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

2. If $\mathcal{F}$ is translation invariant, the dual problem (19) can be represented as

$$\sup_{\boldsymbol{\lambda}} \; -\gamma^*(\boldsymbol{G}\boldsymbol{\lambda}) - \mathcal{F}^*(\boldsymbol{\lambda}) - \delta_C(\boldsymbol{\lambda}) \text{ with } C = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{y}^\top \boldsymbol{\lambda} = 0, \boldsymbol{1}_m^\top \boldsymbol{\lambda} = 1\};$$

3. If $\mathcal{F}$ is positively homogeneous, the dual problem (19) can be represented as

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

Corollary 3. *If $\mathcal{F}$ is monotonic and translation invariant, the dual problem (19) can be rewritten by*

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

Furthermore, if $\mathcal{F}$ is coherent, the third statement of Proposition 3 is valid with $\mathcal{Q}_\mathcal{F}$ such that $\mathcal{Q}_\mathcal{F} \subset \mathbb{\Pi}^m$.

Accordingly, we can interpret the dual variable as a probability distribution. Especially, on the basis of Theorem 3, we can obtain an explicit representation by connecting a class of risk functionals and φ-divergence.

Corollary 4. *If the risk functional $\mathcal{F}$ is written by (14) with v^* which satisfies $\mathrm{dom}\, v^* \subset \mathbb{R}_+$, then the dual problem (19) is represented by*

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

Looking at Table 3, we see that $\mathrm{dom}\, v^*$ of CVaR and LSE are in the nonnegative while MV is not. As will be demonstrated in Section 5, we can find interesting interpretations of the dual problems if $\mathcal{F}$ is monotonic and translation invariant, or $\mathcal{F}$ is coherent on the basis of Proposition 3 and Corollaries 3 and 4.

4.3 General formulations with non-ℓ_2-norm regularizers

To explore a tractable formulation with each non-ℓ_2-norm, we start with the following observation.

Proposition 4. *Suppose that both regularizer $\gamma(\cdot)$ and risk functional $\mathcal{F}(\cdot)$ are positively homogeneous. Then the primal (18) either has an optimal solution $(\boldsymbol{w}^\star, b^\star)$ such that $\boldsymbol{w}^\star = \mathbf{0}$, or results in an unbounded solution such that $p^\star = -\infty$.*

Proof. Note that $(\boldsymbol{w}, b) = \mathbf{0}$ is feasible to (18) 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 $(\boldsymbol{w}, b)$. □

The above observation indicates that the combination of a positively homogeneous functional $\mathcal{F}$ and a regularizer of the form $\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\|$ is not qualified for the task in hand. On the other hand, with a non-homogeneous ι, the regularizer given in the form $\gamma(\boldsymbol{w}) = \iota(\|\boldsymbol{w}\|)$ makes sense.

Indeed, the case where $\gamma(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, corresponding to $\iota(z) = \frac{1}{2}z^2$, is central in the SVM. Similarly, we can employ a non-ℓ_2-norm for the positively homogeneous risk minimization with the Tikhonov regularization if a non-positively homogeneous function such as $\iota(z) = \frac{1}{2}z^2$ is applied. However, such a strategy results in a non-linear formulation even when a polyhedral norm such as the ℓ_1- or the ℓ_∞-norm is employed as $\|\cdot\|$. Accordingly, we further consider the case of the Ivanov regularization, i.e., $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\| \leq 1}(\boldsymbol{w})$.[19]

Namely, the primal formulation (18) and the dual formulation (19) become

$$p^\star := \inf_{\boldsymbol{w}, b} \ \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) + \delta_{\|\cdot\| \leq 1}(\boldsymbol{w}), \tag{20}$$

and

$$d^\star := \sup_{\boldsymbol{\lambda}} \ -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) - \delta_{\boldsymbol{y}^\top(\cdot)=0}(\boldsymbol{\lambda}), \tag{21}$$

respectively. Formulations (20) and (21) are given in the "inf / sup" form so as to fit the development of duality theory later. For the sake of practical implementation, let us consider the corresponding primal and dual pair in the "min / max" form:

$$p^\star := \left| \begin{array}{ll} \underset{\boldsymbol{w}, b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{w}\| \leq 1; \end{array} \right. \tag{22}$$

$$d^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0. \end{array} \right. \tag{23}$$

We denote by $(\mathcal{F}, \|\cdot\|)$ the pair of the primal and dual formulations (22) and (23) for an SVM.

Example. Employing the convex risk functional $\mathrm{LSE}_{(t,\boldsymbol{p})}(\cdot)$, we have a (nonlinear) convex risk functional-based SVM $(\mathrm{LSE}_{(t,\boldsymbol{p})}, \|\cdot\|)$, where its dual is obtained as

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

[19] The Tikhonov and the Ivanov regularizations are often considered identical regularizations. However, as the above example indicates, a careful treatment is required in viewing the 'equivalence' especially if there is a chance that the optimality is not attained.

Optimality conditions. In the context of SVM, more important is the correspondence between primal and dual solutions, i.e., $(\boldsymbol{w}^\star, b^\star)$ and $\boldsymbol{\lambda}^\star$, as given by (4) for the standard SVMs. The following theorem provides a general form of the dual correspondence between the two solutions.

Theorem 5 (Optimality condition)**.** *In order that* $(\boldsymbol{w}^\star, b^\star)$ *and* $\boldsymbol{\lambda}^\star$ *be vectors such that*

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

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

$$\boldsymbol{G}\boldsymbol{\lambda}^\star \in \mathcal{N}(\boldsymbol{w}^\star), \quad \|\boldsymbol{w}^\star\| \le 1, \quad \boldsymbol{y}^\top \boldsymbol{\lambda}^\star = 0, \quad -\boldsymbol{G}^\top \boldsymbol{w}^\star + \boldsymbol{y}b^\star \in \partial\mathcal{F}^*(\boldsymbol{\lambda}^\star), \tag{25}$$

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.,* $\partial\mathcal{F}^*(\boldsymbol{\lambda}^\star) := \{\boldsymbol{L} : \mathcal{F}^*(\boldsymbol{L}) \ge \mathcal{F}^*(\boldsymbol{\lambda}^\star) + \boldsymbol{L}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star),$ *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 (25) can be decomposed into the following three cases: Case i) $\boldsymbol{w}^\star = \boldsymbol{0}$. Then we have $\boldsymbol{G}\boldsymbol{\lambda}^\star = \boldsymbol{0}$; Case ii) $\boldsymbol{G}\boldsymbol{\lambda}^\star = \boldsymbol{0}$. Then we have $\|\boldsymbol{w}^\star\| \le 1$; Case iii) $\boldsymbol{G}\boldsymbol{\lambda}^\star \neq \boldsymbol{0}$. Then we have $\|\boldsymbol{w}^\star\| = 1$. Cases ii) and iii) can be rewritten by

$$\boldsymbol{w}^\star \in \arg\max_{\boldsymbol{w}} \{(\boldsymbol{\lambda}^\star)^\top \boldsymbol{G}^\top \boldsymbol{w} : \|\boldsymbol{w}\| \le 1\}. \tag{26}$$

In particular, if we employ the ℓ_2-norm, Case iii) implies that

$$\boldsymbol{w}^\star = \frac{\boldsymbol{G}\boldsymbol{\lambda}^\star}{\|\boldsymbol{G}\boldsymbol{\lambda}^\star\|_2}. \tag{27}$$

Accordingly, as long as the ℓ_2-norm is employed, the resultant classifier is the same as the standard Tikhonov regularization. This condition, which claims a parallelism between $\boldsymbol{w}^\star$ and $\boldsymbol{\lambda}^\star$, corresponds to the one given in (4), which is a key to the kernel trick for the ordinary SVMs (e.g., Burges, 1998, or Section D for a brief explanation).

Contrarily, if we employ a non-ℓ_2-norm, we have to pay attention to the deviation from (27). We will discuss this, especially the use of a parameterized class of LP-representable norms, in Section 6.

4.4 Positively homogeneous convex risk functional-based SVMs

If we limit our attention to positively homogeneous risk functionals, the dual formulation (21) can be simplified with the help of its risk envelope. Indeed, we can obtain a suggestive primal and dual pair of formulations with the help of the risk envelope $\mathcal{Q}_\mathcal{F}$:

$$p^\star := \left| \begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \sup_{\boldsymbol{q}}\{-\boldsymbol{q}^\top(\boldsymbol{G}^\top\boldsymbol{w} - \boldsymbol{y}b) : \boldsymbol{q} \in \mathcal{Q}_\mathcal{F}\} \\ \text{subject to} & \|\boldsymbol{w}\| \le 1. \end{array} \right. \tag{28}$$

and

$$d^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0, \quad \boldsymbol{\lambda} \in \mathcal{Q}_\mathcal{F}. \end{array} \right. \tag{29}$$

Here note that there is a symmetric dual correspondence between the primal (28) and the dual (29). In fact, the primal (28) has a norm constraint with $\|\cdot\|$ while its dual norm $\|\cdot\|^\circ$ appears in the objective of the dual (29); the positively homogeneous convex risk functional $\mathcal{F}$ in the primal's objective corresponds to its risk envelope $\mathcal{Q}_\mathcal{F}$ in the dual's constraint.[20]

[20] (28) with $\mathcal{Q}_\mathcal{F} \subset \mathbb{\Pi}^m$ is a convex relaxation of the formulation developed by Gotoh, Takeda, and Yamamoto (2013). On the other hand, the dual formulation (29) is not mentioned in their paper since theirs includes some nonconvexity.

Remark 1. Tsyurmasto, Gotoh, and Uryasev (2013) show that under some mild conditions, the following formulations are equivalent in the sense that all of them provide the same (set of) classifiers.

$$\left| \begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{w}\|_2 \leq E, \end{array} \right. \tag{30}$$

$$\left| \begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & C \cdot \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) + \frac{1}{2}\|\boldsymbol{w}\|_2^2, \end{array} \right. \tag{31}$$

$$\left| \begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} & \mathcal{F}(-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y}b) \leq -D, \end{array} \right. \tag{32}$$

where E, C and D are positive constants. Besides, the equivalence is independent of E, C and D, and therefore, we can set $E = C = D = 1$. This is another virtue of the positive homogeneity of the risk functional. Indeed, when the ℓ_2-norm is employed, the quadratically constrained formulation (30) can be replaced with quadratic programming formulations (31) and (32). Note that solver softwares which can optimize (30) is less available than that for (31) and (32). Even with solver softwares which can solve both, the latter formulations can be more stably solved.

On the other hand, without positive homogeneity, the above equivalence does not hold. Indeed, with $\mathcal{F}(\boldsymbol{L}) = \mathrm{Hinge1}_{(1,\mathbf{1}_m C/m)}$, the (31) is equal to C-SVM (1). However, the equivalence among (30) to (32) depends on E and D.

5 Interpretations of formulations from various viewpoints

In this section, we demonstrate connections between the general formulations and the existing papers, using the notions introduced so far.

5.1 A connection to geometric interpretation

It is known that dual problems of a couple of SVMs can be interpreted as the problem of finding two nearest points over two separate sets, each corresponding to the data points having the same label. With the dual formulation (19), 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 (implicit) constraints, $\boldsymbol{y}^\top \boldsymbol{\lambda} = 0$, of (19) 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 $-\gamma^*(\sum_{i \in I_+} \boldsymbol{x}_i \lambda_i - \sum_{h \in I_-} \boldsymbol{x}_h \lambda_h)$, where $I_+ := \{i \in \{1, ..., m\} : y_i = +1\}$ and $I_- := \{i \in \{1, ..., m\} : y_i = -1\}$. Consequently, with a change of variables $\mu_{+,i} := 2\lambda_i$ for $i \in I_+$ and $\mu_{-,i} := 2\lambda_i$ for $i \in I_-$, (19) can be represented as

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

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

For further concrete interpretation, let us consider the case where $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\| \leq 1}(\boldsymbol{w})$ and $\mathcal{F}$ is given in the form of (9) with monotonic v. Then (19) is represented as

$$-d^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\mu}_+, \boldsymbol{\mu}_-}{\text{minimize}} & \frac{1}{2} \left\| \sum\limits_{i \in I_+} \boldsymbol{x}_i \mu_{+,i} - \sum\limits_{i \in I_-} \boldsymbol{x}_i \mu_{-,i} \right\|^\circ + \mathcal{I}_{v^*}(\frac{1}{2}\boldsymbol{\mu}, \boldsymbol{p}) \\ \text{subject to} & \boldsymbol{\mu}_+ \in \mathbb{\Pi}^{|I_+|}, \ \boldsymbol{\mu}_- \in \mathbb{\Pi}^{|I_-|}, \end{array} \right. \tag{33}$$

where $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ are vectors consisting of $\mu_{+,i}$ and $\mu_{-,i}$, respectively. It is noteworthy that the formulation (33) is close to what Kanamori, Takeda, and Suzuki (2013) demonstrate. Indeed, employing

the ℓ_2-norm, they virtually present the minimization of the empirical risk 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 $\|\boldsymbol{w}\|_2^2 \leq t$ with $t > 0$. It is easy to see that (33) 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 $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\|\leq 1}(\boldsymbol{w})$ 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

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

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 $\boldsymbol{\mu}_-$. $\mathcal{Q}_{\mathrm{CVaR}(\alpha, 2\boldsymbol{p}_+)} \subset \mathrm{\Pi}^{|I_+|}$ and $\mathcal{Q}_{\mathrm{CVaR}(\alpha, 2\boldsymbol{p}_-)} \subset \mathrm{\Pi}^{|I_-|}$.[21] 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 functional and a general norm, which is parallel to Kanamori, Takeda, and Suzuki (2013).

We would also emphasize that (33) is derived in a mechanical manner based on the Fenchel duality, while the derivation by all the aforementioned papers is based on the Lagrangean dual for QPs. This is also advantageous in that non-ℓ_2-norm can be similarly treated.

In addition, the formulation (33) bridges the geometric interpretation and an information theoretic interpretation. Noting the relation $\mathcal{I}_{v^*}(\frac{1}{2}\boldsymbol{\mu}, \boldsymbol{p}) = \sum_{i\in I_+} p_i v^*(\frac{\mu_i}{2p_i}) + \sum_{h\in I_-} p_h v^*(\frac{\mu_h}{2p_h})$, the second term of the objective 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 $\sum_{i\in I_+} p_i = \sum_{h\in I_-} p_h (= \frac{1}{2})$, the penalty term is rewritten as $\mathcal{I}_\varphi(\boldsymbol{\mu}_+, \boldsymbol{r}_+) + \mathcal{I}_\varphi(\boldsymbol{\mu}_-, \boldsymbol{r}_-)$ where $r_i = p_i / \sum_{h\in I_+} p_h$ for $i \in I_+$ and $r_i = p_i / \sum_{h\in I_-} p_h$ for $i \in I_-$.[22] Namely, we can symbolically recast (33) as

$$-\underset{\boldsymbol{\mu}\in \mathrm{\Pi}^m}{\text{minimize}} \ \ \frac{1}{2}\|\mathbb{E}_{\boldsymbol{\mu}}(\boldsymbol{x}|y=+1) - \mathbb{E}_{\boldsymbol{\mu}}(\boldsymbol{x}|y=-1)\|^\circ + \mathcal{I}_\varphi(\boldsymbol{\mu}, \boldsymbol{p}|y=+1) + \mathcal{I}_\varphi(\boldsymbol{\mu}, \boldsymbol{p}|y=-1),$$

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 $a=-1$.

5.2 Interpretation of regularizers based on robust optimization modeling

Xu, Caramanis, and Mannor (2009) show that the regularized hinge loss minimization can be interpreted as a robust optimization modeling. They suppose that each sample of the given data set $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)\}$ suffers from some perturbation of the form

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

Under this uncertainty, they consider to minimize the worst-case empirical hinge loss. Namely, they consider to minimize $\sup_{\boldsymbol{\Delta}\in\mathcal{T}} \mathrm{Hinge1}_{(m, \mathbf{1}_m/m)}(-\boldsymbol{Y}\{(\boldsymbol{X} - \boldsymbol{\Delta})\boldsymbol{w} - \mathbf{1}_m b\})$, where $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)^\top$.

Theorem 6 (Xu, Caramanis, and Mannor (2009))**.** *Suppose that the training samples are linearly non-separable. Then the following two optimization problems on $(\boldsymbol{w}, b)$ are equivalent.*

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

[21] We admit an abuse of the notation. Indeed, it is not valid that $2\boldsymbol{p}_+ \in \mathrm{\Pi}^{|I_+|}$ or $2\boldsymbol{p}_- \in \mathrm{\Pi}^{|I_-|}$. However, we put $2\boldsymbol{p}_+$ or $2\boldsymbol{p}_-$ in the place of a probability measure $\boldsymbol{p}$.

[22] This condition may be excessively strong in application. However, it is worth mentioning that the approach of Bennett and Mangasarian (1992) virtually employs this assumption and attains a nice performance in the breast cancer data set (see Wolberg, Street, and Mangasarian, 1995).

$$\underset{\boldsymbol{w},b}{\text{minimize}} \;\; C\|\boldsymbol{w}\| + \sum_{i=1}^{m}[1 - y_i\{\boldsymbol{w}^\top(\boldsymbol{x}_i - \boldsymbol{\delta}_i) - b\}]^+.$$

In this subsection, we derive a similar results for the case of monotonic and translation invariant risk functionals. In place of $\mathcal{T}$, we consider the following uncertainty.

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

Note that $\mathcal{S}$ is called the box uncertainty in Xu, Caramanis, and Mannor (2009), and $\mathcal{S} \supset \mathcal{T}$ holds.

Theorem 7. *Let the risk functional $\mathcal{F}$ be monotonic and translation invariant as well as proper, l.s.c. and convex. Then, for any $(\boldsymbol{w}, b)$, we have*

$$\max_{\boldsymbol{\Delta}\in\mathcal{S}} \; \mathcal{F}(-\boldsymbol{Y}\{(\boldsymbol{X} - \boldsymbol{\Delta})\boldsymbol{w} - \boldsymbol{1}_m b\}) = C\|\boldsymbol{w}\| + \mathcal{F}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}_m b)), \tag{34}$$

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

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

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

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

Accordingly, the Tikhonov regularization $\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\|$ can be interpreted as a robustification technique not only for the hinge loss, but also for a variety of other risk functionals. However, there is a fact which is noteworthy. As shown in Proposition 4, if $\mathcal{F}$ is coherent, the unconstrained minimization of (34) is not adequate for the classification task. In order the minimization of (34) to make sense, the addition of an Ivanov regularization is required. Indeed, a primal formulation with a monotonic translation invariant risk functional and a regularizer $\gamma(\boldsymbol{w}) = \|\boldsymbol{w}\| + \delta_{\|\cdot\|'\le 1}(\boldsymbol{w})$ where $\|\cdot\|'$ is another norm can be viewed as a robustified version of $\gamma(\boldsymbol{w}) = \delta_{\|\cdot\|'\le 1}(\boldsymbol{w})$.

5.3 Distributionally robust SVMs

Different from the robust optimization modeling described in the previous section, the so-called *distributionally robust optimization* is also popular in the literature (see, e.g., Wiesemann, Kuhn, and Sim, 2013, for the recent development). In this subsection, we show that a class of generalized SVM formulations described in this paper fits also to the extension of this another robust optimization modeling.

In existing SVMs, the m samples, $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)$, are usually supposed to be independently drawn from an unknown distribution, and the empirical expectation in the empirical risk is taken under the uniform probability distribution, $\boldsymbol{p} = \boldsymbol{1}_m/m$. However, the 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 the reference probability is under uncertainty of the form: $\boldsymbol{p}+\boldsymbol{\delta} \in P$ with some P satisfying $\boldsymbol{p} \in P \subset \mathbb{\Pi}^m$. Similarly to the previous section, one reasonable strategy is to consider the worst case over the set P. Let us list examples of the uncertainty set P.

- $P = \mathcal{Q}_{\mathrm{Fi}(\boldsymbol{p}_1,...,\boldsymbol{p}_K)} := \{\boldsymbol{p}_1, ..., \boldsymbol{p}_K\}$, with $\boldsymbol{p}_1, ..., \boldsymbol{p}_K \in \mathbb{\Pi}^m$;
- $P = \mathcal{Q}_{\mathrm{Dist}(\|\cdot\|', \boldsymbol{A}, \boldsymbol{p})} := \{\boldsymbol{\pi} \in \mathbb{\Pi}^m : \boldsymbol{\pi} = \boldsymbol{p} + \boldsymbol{A}\boldsymbol{\zeta}, \|\boldsymbol{\zeta}\|' \leq 1\}$, with $\boldsymbol{A} \in \mathbb{S}^m_{++}$, $\|\cdot\|'$: a norm;
- $P = \mathcal{Q}_{\mathcal{I}_\varphi(t, \boldsymbol{p})} := \{\boldsymbol{\pi} \in \mathbb{\Pi}^m : \mathcal{I}_\varphi(\boldsymbol{\pi}, \boldsymbol{p}) \leq t\}$, 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 $\boldsymbol{p}_1, ..., \boldsymbol{p}_K$ 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.[23]

Let us consider that the risk functional $\mathcal{F}_{\boldsymbol{p}}$ has the form (14). Then, the worst-case risk functional is given as

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

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.[24]

Proposition 5. *If P is a convex set and $\mathcal{I}_\varphi$ is a φ-divergence, the distributionally robust version of the generalized dual formulation is a convex minimization*

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

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

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

Although we describe the case of $P = \mathcal{Q}_{\mathrm{Fi}(\boldsymbol{p}_1,...,\boldsymbol{p}_K)}$ 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 functional $\mathcal{F}_{\boldsymbol{p}}$ is positively homogeneous. Indeed, we then have

$$\text{Worst-}\mathcal{F}_P(\boldsymbol{L}) = \max_{\boldsymbol{p} \in P} \max_{\boldsymbol{q} \in \mathcal{Q}_{\mathcal{F}}(\boldsymbol{p})} \boldsymbol{q}^\top \boldsymbol{L} = \sup_{\boldsymbol{q}} \{-\boldsymbol{L}^\top \boldsymbol{q} : \boldsymbol{q} \in \underset{\boldsymbol{p} \in P}{\cup} \mathcal{Q}_{\mathcal{F}}(\boldsymbol{p})\}. \tag{35}$$

Note that the union, $\cup_{\boldsymbol{p} \in P} \mathcal{Q}_{\mathcal{F}}(\boldsymbol{p})$, in (35) can be a nonconvex set. However, the convex hull of the union provides an equivalent coherent risk functional. Namely, we have $\text{Worst-}\mathcal{F}_P(\boldsymbol{L}) = \max_{\boldsymbol{q}}\{\boldsymbol{q}^\top \boldsymbol{L} : \boldsymbol{q} \in \mathrm{conv}(\cup_{\boldsymbol{p} \in P} \mathcal{Q}_{\mathcal{F}}(\boldsymbol{p}))\}$. Since the convex hull of the risk envelopes becomes another (possibly, larger) risk envelope, the distributionally robust coherent risk functional-based SVM is also another coherent risk functional-based SVM. Below are examples of the uncertainty set P with which we can achieve tractable convex optimization formulations for distributionally robust SVMs.

The distributionally robust version of the dual form (29) is thus rewritten by

$$d^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0,\ \boldsymbol{\lambda} \in \mathrm{conv}(\{\mathcal{Q}_{\mathcal{F}}(\boldsymbol{p}) : \boldsymbol{p} \in P\}). \end{array} \right. \tag{36}$$

[23] $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|, \boldsymbol{A}, \boldsymbol{p})}$ is considered as a set of probability measures which are away from a reference probability measure $\boldsymbol{p}$ with distance at most 1 under a norm $\|\cdot\|$ and a metric $(\boldsymbol{A}^{-1})^2$. If $\|\cdot\| = \|\cdot\|_\infty$ and $\boldsymbol{A} = \mathrm{diag}(\overline{\boldsymbol{\zeta}})$ with some $\overline{\boldsymbol{\zeta}} \geq \mathbf{0}$, $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|, \boldsymbol{A}, \boldsymbol{p})}$ forms a box-type constraint, i.e., $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|_\infty, \mathrm{diag}(\overline{\boldsymbol{\zeta}}), \boldsymbol{p})} = \{\boldsymbol{q} \in \mathbb{\Pi}^m : \boldsymbol{q} \in [\boldsymbol{p} - \overline{\boldsymbol{\zeta}}, \boldsymbol{p} + \overline{\boldsymbol{\zeta}}]\}$.

[24] If $\varphi(q)$ is convex in q, then $p\varphi(q/p)$ is convex in (p, q) for $p > 0$ (see Section 3.2.6 of Boyd and Vandenberghe, 2004, for the details).

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] $P = \mathcal{Q}_{\mathrm{Fi}(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K)}$ $\leftrightarrow$

$$\begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{\pi},\boldsymbol{\tau}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \mathbf{1}_m^\top\boldsymbol{\lambda} = 1,\ \mathbf{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \boldsymbol{\pi} = \sum_{k=1}^{K} \tau_k \boldsymbol{p}_k,\ \mathbf{1}_K^\top\boldsymbol{\tau} = 1,\ \boldsymbol{\tau} \ge \mathbf{0}; \end{array}$$

[Distance-based uncertainty] $P = \mathcal{Q}_{\mathrm{Dist}(\|\cdot\|',\boldsymbol{A},\mathbf{1}_m/m)}$ $\leftrightarrow$

$$\begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{\pi},\boldsymbol{\zeta}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \mathbf{1}_m^\top\boldsymbol{\lambda} = 1,\ \mathbf{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \boldsymbol{\pi} = \frac{1}{m}\mathbf{1}_m + \boldsymbol{A}\boldsymbol{\zeta},\ \mathbf{1}_m^\top\boldsymbol{A}\boldsymbol{\zeta} = 0,\ \|\boldsymbol{\zeta}\|' \le 1; \end{array}$$

[Entropy-based uncertainty] $P = \mathcal{Q}_{\mathrm{KL}(t,\mathbf{1}_m/m)}$ $\leftrightarrow$

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

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

It is noteworthy that the distributional robustification technique above incorporates prior knowledge on the distribution without significant increase of complexity of optimization problem 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 modelings often lead to excessively conservative results, adding experts' knowledge as constraints can be helpful to escape from those situations.

6 SVMs with general norms and application to parametrized families of polyhedral norms

In the literature, the ℓ_2-norm is sometimes replaced with non-ℓ_2-norms. However, justification of the replacement is ambiguous although, intuitively, we can expect $\|\cdot\|_1 \approx \|\cdot\|_2$. This section treats arbitrary norms with the general formulations (22) and (23).

6.1 Proximity of SVMs

First note that even in our formulations, the ℓ_2-norm is still a benchmark because only the ℓ_2-norm can simultaneously appear both in the primal and dual formulations due to the self-duality, i.e., $\|\cdot\| = \|\cdot\|^\circ$. In addition, the so-called representer theorem based on the parallelism between primal solution $\boldsymbol{w}$ and

dual solution $\boldsymbol{\lambda}$ holds only with the ℓ_2-norm. Thus, in order to recognize the difference between the ℓ_2-based SVMs and the non-ℓ_2-based SVMs, let us introduce the following notations:

$$\begin{aligned}
\phi^{\mathrm{E}} &:= \min_{\boldsymbol{w},b}\{\mathcal{F}(\boldsymbol{y}b - \boldsymbol{G}^\top \boldsymbol{w}) : \|\boldsymbol{w}\|_2 \le 1\} \;\equiv\; \max_{\boldsymbol{\lambda}}\{-\|\boldsymbol{G}\boldsymbol{\lambda}\|_2 - \mathcal{F}^*(\boldsymbol{\lambda}) : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0\}, && (\ell_2\text{-norm})\\
\hat{\phi} &:= \min_{\boldsymbol{w},b}\{\mathcal{F}(\boldsymbol{y}b - \boldsymbol{G}^\top \boldsymbol{w}) : \;\|\boldsymbol{w}\| \le 1\} \;\equiv\; \max_{\boldsymbol{\lambda}}\{-\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0\}, && (\text{arbitrary norm})\\
\phi_s &:= \min_{\boldsymbol{w},b}\{\mathcal{F}(\boldsymbol{y}b - \boldsymbol{G}^\top \boldsymbol{w}) : s\|\boldsymbol{w}\| \le 1\} \;\equiv\; \max_{\boldsymbol{\lambda}}\{-\tfrac{1}{s}\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) : \boldsymbol{y}^\top\boldsymbol{\lambda} = 0\}, && \text{for some } s > 0.
\end{aligned}$$

Note that $\phi_1 = \hat{\phi}$.

Let us consider positive constants L and U such that for any $\boldsymbol{x} \in \mathbb{R}^n$, $L\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\| \le U\|\boldsymbol{x}\|_2$. Also, observe that it is valid $(1/U)\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|^\circ \le (1/L)\|\boldsymbol{x}\|_2$ at the same time. We write these relations as

$$\forall \boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}, \quad L \le \left\{ \begin{array}{c} \|\boldsymbol{x}\|/\|\boldsymbol{x}\|_2 \\ \|\boldsymbol{x}\|_2/\|\boldsymbol{x}\|^\circ \end{array} \right\} \le U. \tag{37}$$

To make (37) the most informative, let us suppose that the constants L and U are tight, i.e., L is the largest among the lower bounds, while U is the smallest among the upper bounds. For example, when $\|\cdot\| = \|\cdot\|_1$, it is valid

$$1 \le \left\{ \begin{array}{c} \|\boldsymbol{x}\|_1/\|\boldsymbol{x}\|_2 \\ \|\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_\infty \end{array} \right\} \le \sqrt{n}.$$

Based on such tight bounds, we can evaluate the proximity between $(\mathcal{F}, \|\cdot\|)$ and the ℓ_2-formulation $(\mathcal{F}, \|\cdot\|_2)$.

Theorem 8. (**Proximity between the SVM with the ℓ_2-norm and that with a general norm**) *Suppose that there are two positive constants L and U satisfying the relation (37). Further suppose that the primal (22) and the dual (23) have optimal solutions $(\hat{\boldsymbol{w}}, \hat{b})$ and $\hat{\boldsymbol{\lambda}}$, respectively, i.e., $\hat{\phi} = \mathcal{F}(\boldsymbol{y}\hat{b} - \boldsymbol{G}^\top\hat{\boldsymbol{w}}) = -\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|^\circ - \mathcal{F}^*(\hat{\boldsymbol{\lambda}})$. Then we have,*

$$\phi_{\frac{L}{U}} \le \hat{\phi} \le \phi_{\frac{U}{L}}, \quad \phi_{\frac{1}{U}} \le \phi^{\mathrm{E}} \le \phi_{\frac{1}{L}}.$$

Especially if $L = 1$, then we have $\phi_{\frac{1}{U}} \le \phi^{\mathrm{E}} \le \hat{\phi} \le \phi_U$. On the other hand, if $U = 1$, then we have $\phi_L \le \hat{\phi} \le \phi^{\mathrm{E}} \le \phi_{\frac{1}{L}}$.

Furthermore, if $\mathcal{F}$ is positively homogeneous (as well as proper, l.s.c. and convex), then $(\hat{\boldsymbol{w}}, \hat{b})/\|\hat{\boldsymbol{w}}\|_2$ and $\hat{\boldsymbol{\lambda}}$ are feasible to (28) and (29) with $\|\cdot\|_2$, respectively, and we have

$$U\hat{\phi} \le -\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2 \le \phi^{\mathrm{E}} \le \frac{\hat{\phi}}{\|\hat{\boldsymbol{w}}\|_2} \le L\hat{\phi}. \tag{38}$$

Note that (38) implies

$$L \le \frac{\phi^{\mathrm{E}}}{\hat{\phi}} \le U, \quad 1 \le \frac{\phi^{\mathrm{E}}}{\hat{\phi}/\|\hat{\boldsymbol{w}}\|_2} \le \frac{U}{L}, \quad 1 \le \frac{-\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2}{\phi^{\mathrm{E}}} \le \frac{U}{L}. \tag{39}$$

The first pair of inequalities in (39) imply that when the risk functional is positively homogeneous, the tight bounds (37) also measure the proximity between the optimal values of $\|\cdot\|_2$- and $\|\cdot\|$- (or $\|\cdot\|^\circ$-) formulations. On the other hand, the second and the third pairs imply that $(\hat{\boldsymbol{w}}, \hat{b})/\|\hat{\boldsymbol{w}}\|_2$ and $\hat{\boldsymbol{\lambda}}$ are guaranteed with the ratio U/L to be close to the optimal solution to the ℓ_2-norm case. In this sense, the ratio U/L measures the closeness between $(\mathcal{F}, \|\cdot\|)$ and $(\mathcal{F}, \|\cdot\|_2)$ when $\mathcal{F}$ is positively homogeneous.

For later reference, we here also examine the optimality condition (25) on the basis of the proximity between $\|\cdot\|_2$ and $\|\cdot\|$. From Theorem 5, we have $\boldsymbol{u}^\top\boldsymbol{w}^\star = \|\boldsymbol{u}\|^\circ, \|\boldsymbol{w}^\star\| = 1$ with $\boldsymbol{u} := \boldsymbol{G}\boldsymbol{\lambda}^\star \neq \boldsymbol{0}$, for optimal solutions $(\boldsymbol{w}^\star, b^\star)$ and $\boldsymbol{\lambda}^\star$ to the primal and dual formulations. However, note that unless the

ℓ_2-norm is employed, we cannot conclude the relation $\boldsymbol{w}^\star = \eta \boldsymbol{G}\boldsymbol{\lambda}^\star$ for some η. Therefore, to gauge the deviation from the condition under a general norm $\|\cdot\|$, we consider the ratio:

$$\frac{\boldsymbol{u}^\top \boldsymbol{w}^\star}{\|\boldsymbol{u}\|_2 \|\boldsymbol{w}^\star\|_2}.$$

From the Cauchy-Schwarz's inequality, this ratio is bounded above by 1, which is achieved if and only if $\boldsymbol{w}^\star = \eta \boldsymbol{u}$ for some $\eta > 0$ as long as $\boldsymbol{u} \neq \boldsymbol{0}$ and $\boldsymbol{w}^\star \neq \boldsymbol{0}$. In this sense, the closer to 1 the ratio is, the more justifiable the parallelism is.

Theorem 9 (Proximity for parallelism between $\boldsymbol{w}^\star$ and $\boldsymbol{G}\boldsymbol{\lambda}^\star$)**.** *Suppose that there are two positive constants L and U satisfying the condition (37). Let $(\boldsymbol{w}^\star, b^\star)$ and $\boldsymbol{\lambda}^\star$ be optimal solutions to the primal (22) and dual formulations (23), respectively. Suppose that the solutions satisfy the strong duality and $\boldsymbol{G}\boldsymbol{\lambda}^\star \neq \boldsymbol{0}$. Then,*

$$\frac{L}{U} \leq \frac{(\boldsymbol{G}\boldsymbol{\lambda}^\star)^\top \boldsymbol{w}^\star}{\|\boldsymbol{G}\boldsymbol{\lambda}^\star\|_2 \|\boldsymbol{w}^\star\|_2} \leq 1. \tag{40}$$

This theorem states that the lower (or higher) the ratio U/L (or L/U, respectively) is, the more the parallelism between $\boldsymbol{w}^\star$ and $\boldsymbol{G}\boldsymbol{\lambda}^\star$ fits. Namely, the ratio U/L measures the degree of congruency of the classifier with the ℓ_2-classifier, independently of the data set.

Note that either $\|\cdot\| = \|\cdot\|_1$ or $\|\cdot\| = \|\cdot\|_\infty$ leads to $U/L = \sqrt{n}$. On the other hand, if the ℓ_p-norm, i.e., $\|\boldsymbol{x}\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$, is employed, smaller ratio U/L can be obtained for $p \neq 1$ or ∞. However, $(\mathcal{F}, \|\cdot\|_p)$ becomes harder than $(\mathcal{F}, \|\cdot\|_2)$, $((\mathcal{F}, \|\cdot\|_1)$ or $(\mathcal{F}, \|\cdot\|_\infty))$.

In the following two subsection, we introduce two pairs of polyhedral norms analyzed in Gotoh and Uryasev (2013), with which we can attain smaller U/L (or larger L/U) without significant increase of the complexity of the optimization problem from the ℓ_1-norm or ℓ_∞-norm case.

6.2 Case: ν-SVMs with CVaR norm and its dual norm

As explained in Pavlikov and Uryasev (2013) and Gotoh and Uryasev (2013), the notion of CVaR can provide a parametrized family of polyhedral norms. Indeed, we can define the *CVaR norm* by

$$\langle\!\langle \boldsymbol{w} \rangle\!\rangle_\alpha := n(1-\alpha)\mathrm{CVaR}_{(\alpha, \mathbf{1}_n/n)}(|\boldsymbol{w}|),$$

where $\alpha \in [0, 1)$.[25] The CVaR norm includes the so-called *largest-k norm* as a special case. Indeed, for $k = 1, 2, ..., n$, the CVaR norm with $\alpha = 1 - k/n$ is equal to the largest-k norm. Especially, if $\alpha = 0$, the CVaR norm is equal to the ℓ_1-norm, while if $\alpha = \frac{n-1}{n}$, it is equal to the ℓ_∞-norm. Note that for any $\alpha \in [0, 1)$, minimization of the CVaR norm can be written by an LP, and it can be employed in the general formulations $(\mathcal{F}, \langle\!\langle \cdot \rangle\!\rangle_\alpha)$ without significant increase of computational complexity from the ℓ_1- or ℓ_∞-formulation.

Its dual norm is given by

$$\langle\!\langle \boldsymbol{w} \rangle\!\rangle_\alpha^\circ \equiv \max\{\|\boldsymbol{w}\|_1/\{n(1-\alpha)\}, \|\boldsymbol{w}\|_\infty\}.$$

See Bertsimas, Pachamanova, and Sim (2004) for the derivation. The dual CVaR norm also includes the ℓ_1- and ℓ_∞-norms as special limiting cases, i.e., $\langle\!\langle \boldsymbol{w} \rangle\!\rangle_0^\circ = \|\boldsymbol{w}\|_\infty$ and $\langle\!\langle \boldsymbol{w} \rangle\!\rangle_{\frac{n-1}{n}}^\circ = \|\boldsymbol{w}\|_1$.

According to Gotoh and Uryasev (2013), the CVaR and dual CVaR norms have the tight bounds as follows.

$$\min\{1, \sqrt{n}(1-\alpha)\} \leq \left\{ \begin{array}{c} \langle\!\langle \boldsymbol{w} \rangle\!\rangle_\alpha / \|\boldsymbol{x}\|_2 \\ \|\boldsymbol{x}\|_2 / \langle\!\langle \boldsymbol{w} \rangle\!\rangle_\alpha^\circ \end{array} \right\} \leq \sqrt{\lfloor \kappa \rfloor + (\kappa - \lfloor \kappa \rfloor)^2},$$

[25] With a reparametrization, it is equal to the D-norm introduced by Bertsimas, Pachamanova, and Sim (2004) except $\alpha \in (\frac{n-1}{n}, 1)$.

where $\kappa = n(1-\alpha)$. Further, they show that the ratio, $U/L = \sqrt{\lfloor\kappa\rfloor + (\kappa - \lfloor\kappa\rfloor)^2}/\min\{1, \kappa/\sqrt{n}\}$, is minimized at $\kappa = \sqrt{n}$, or equivalently, at $\alpha = 1 - \frac{1}{\sqrt{n}} =: \alpha^\star$, and accordingly, the minimum value is given by $\sqrt{\lfloor\sqrt{n}\rfloor + (\sqrt{n} - \lfloor\sqrt{n}\rfloor)^2}$. See Gotoh and Uryasev (2013) for the proof.

The smallest ratio U/L is approximately in the order of $n^{1/4}$. (Recall that the smallest ratio of the ℓ_1- and ℓ_∞-norms is $n^{1/2}$.)

Example: $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha)$**.** When we employ the CVaR norm as the regularizing norm, i.e., $\|\boldsymbol{x}\| = \langle\!\langle\boldsymbol{x}\rangle\!\rangle_\alpha$, the ν-SVM can be symbolically written by

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \langle\!\langle\boldsymbol{w}\rangle\!\rangle_\alpha \equiv n(1-\alpha)\mathrm{CVaR}_{(\alpha,\mathbf{1}_n/n)}(|\boldsymbol{w}|) \le 1. \end{array}\right. \tag{41}$$

An equivalent LP is given by

$$\left|\begin{array}{lllllll} \underset{\boldsymbol{w}^+,\boldsymbol{w}^-,b,\rho,\boldsymbol{z},c,\boldsymbol{\zeta}}{\text{minimize}} & & & & -\rho + \frac{1}{m\nu}\mathbf{1}_m^\top\boldsymbol{z} & & \downarrow \text{dual var.} \downarrow \\ \text{subject to} & -\boldsymbol{G}^\top\boldsymbol{w}^+ & +\boldsymbol{G}^\top\boldsymbol{w}^- & +\boldsymbol{y}b + \mathbf{1}_m\rho & -\boldsymbol{z} & \le \mathbf{0}, & \leftarrow \boldsymbol{\lambda} \ge \mathbf{0} \\ & \boldsymbol{w}^+ & +\boldsymbol{w}^- & & & -\mathbf{1}_n c - \boldsymbol{\zeta} \le \mathbf{0}, & \leftarrow \boldsymbol{\mu} \ge \mathbf{0} \\ & & & & n(1-\alpha)c + \mathbf{1}_n^\top\boldsymbol{\zeta} & \le 1, & \leftarrow \theta \ge 0 \\ & (\boldsymbol{w}^+, \boldsymbol{w}^-, \boldsymbol{z}, \boldsymbol{\zeta}) \ge \mathbf{0}, & (b, \rho, c : \text{free variable}). & & & & \end{array}\right. \tag{42}$$

With an optimal solution $((\boldsymbol{w}^+)^\star, (\boldsymbol{w}^-)^\star, b^\star, \rho^\star, \boldsymbol{z}^\star, c^\star, \boldsymbol{\zeta}^\star)$, to (42), the decision function is given by $d(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{x}^\top((\boldsymbol{w}^+)^\star - (\boldsymbol{w}^-)^\star) - b^\star)$.

On the other hand, the dual to (42) is given by

$$\left|\begin{array}{lllll} \underset{\boldsymbol{\lambda},\boldsymbol{\mu},\theta}{\text{maximize}} & & & -\theta & \downarrow \text{dual var.} \downarrow \\ \text{subject to} & \boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & \le \mathbf{0}, & \leftarrow \boldsymbol{w}^+ \ge \mathbf{0} \\ & -\boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & \le \mathbf{0}, & \leftarrow \boldsymbol{w}^- \ge \mathbf{0} \\ & -\boldsymbol{y}^\top\boldsymbol{\lambda} & & = 0, & \leftarrow b \\ & -\mathbf{1}_m^\top\boldsymbol{\lambda} & & = -1, & \leftarrow \rho \\ & \boldsymbol{\lambda} & & \le \mathbf{1}_m/(m\nu) & \leftarrow \boldsymbol{z} \ge \mathbf{0} \\ & & \mathbf{1}_n^\top\boldsymbol{\mu} - n(1-\alpha)\theta & = 0, & \leftarrow c \\ & & \boldsymbol{\mu} - \mathbf{1}_n\theta & \le \mathbf{0}, & \leftarrow \boldsymbol{\zeta} \ge \mathbf{0} \\ & (\boldsymbol{\lambda}, \boldsymbol{\mu}, \theta) \ge \mathbf{0}, & & & \end{array}\right. \tag{43}$$

which is symbolically written by

$$\left|\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\langle\!\langle\boldsymbol{G}\boldsymbol{\lambda}\rangle\!\rangle_\alpha^\circ \equiv -\max\{n(1-\alpha)\|\boldsymbol{G}\boldsymbol{\lambda}\|_1, \|\boldsymbol{G}\boldsymbol{\lambda}\|_\infty\} \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{\lambda} \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}. \end{array}\right. \tag{44}$$

In addition to the CVaR risk functional, which corresponds to the ν-SVM, many popular convex risk functionals are LP-representable, as listed in Section 3.1 and Appendix A.2. Thus, we can reformulate as LPs both the primal and dual formulations of an SVM $(\mathcal{F}, \|\cdot\|)$, as long as we employ an LP-representable risk functional $\mathcal{F}$ and an LP-representable norm $\|\cdot\|$.

A merit of LP representations, such as (42) or (43), is wide availability of efficient LP solvers such as IBM ILOG CPLEX, Gurobi and FICO-Xpress-MP. Those general LP solvers provide an easy way to implement those customized SVMs.

Besides, such solver software packages attain the primal and dual solutions, i.e., $(\boldsymbol{w}^\star, b^\star)$ and $\boldsymbol{\lambda}^\star$, simultaneously. Recall that we cannot further expect the parallelism $\boldsymbol{w} = \eta\boldsymbol{G}\boldsymbol{\lambda}$ with a constant $\eta > 0$

when we employ a non-ℓ_2-norm. However, we can still make use of the decision function of the primal form (2) once the dual problem (43) is solved by a state-of-the-art LP solver.[26]

Another well-known (but likely to be understated) merit of LP formulations is its efficiency and stability in computation. Although we may not find any (significant) difference between LP and QP or SOCP in computational complexity theory, the computation of LP is rather faster and more stable than that of those quadratic counterparts in practice.

Also, availability of simplex algorithms for LP is advantageous over interior point algorithms in the so-called post-optimization and/or warm-start methods. For example, by using the so-called parametric simplex method, the so-called regularization path (Hastie, Rosset, and Tibshirani, 2004) can be efficiently implemented (for example, with respect to ν for the ν-SVM).

On the other hand, a merit of the symbolic representation, such as (41) or (44), can be found in the recent development of optimization softwares. In fact, for example, PSG (American Optimal Decisions, Inc., 2009) builds in several risk functional listed in this paper, while CVX (Grant and Boyd, 2012) builds in a norm function which can deal with $\ell_1, \ell_2, \ell_\infty$-norms, etc. By using such a user-friendly interface, the general SVM can be implemented by writing a couple of sentences in a source code. See a pair of primal and dual formulation examples of PSG in Appendix C.

6.3 Case: ν-SVM with a convex combination of ℓ_1- and ℓ_∞-norms

Another pair of parametrized families of LP-representable norms investigated in Gotoh and Uryasev (2013) is the convex combination of the ℓ_1- and ℓ_∞-norms, i.e.,

$$(\!(\boldsymbol{x})\!)_\tau := (1-\tau)\|\boldsymbol{x}\|_1 + \tau\|\boldsymbol{x}\|_\infty,$$

where $\tau \in [0,1]$, and its dual norm, $(\!(\boldsymbol{x})\!)^\circ_\tau$, which is represented in a succinct form:

$$(\!(\boldsymbol{x})\!)^\circ_\tau = \max\left\{ |x_{(1)}|, \frac{|x_{(1)}|+|x_{(2)}|}{2-\tau}, ..., \frac{|x_{(1)}|+\cdots+|x_{(n)}|}{n-(n-1)\tau} \right\},$$

where $x_{(i)}$ is the element whose absolute value is the i-th largest among the n elements of $\boldsymbol{x} \in \mathbb{R}^n$, i.e., $|x_{(1)}| \equiv \|\boldsymbol{x}\|_\infty \geq |x_{(2)}| \geq \cdots \geq |x_{(n)}|$. We call $(\!(\boldsymbol{x})\!)_\tau$ and $(\!(\boldsymbol{x})\!)^\circ_\tau$ the *deltoidal norm* and the *dual deltoidal norm*, respectively. Note that $(\!(\boldsymbol{x})\!)^\circ$ can be represented by the maximum of n CVaR norms:

$$(\!(\boldsymbol{x})\!)^\circ_\tau = \max\left\{ \langle\!\langle \boldsymbol{x} \rangle\!\rangle_{\frac{n-1}{n}}, \frac{\langle\!\langle \boldsymbol{x} \rangle\!\rangle_{\frac{n-2}{n}}}{2-\tau}, ..., \frac{\langle\!\langle \boldsymbol{x} \rangle\!\rangle_0}{n-(n-1)\tau} \right\}. \tag{45}$$

Norms $(\!(\boldsymbol{x})\!)_\tau$ and $(\!(\boldsymbol{x})\!)^\circ_\tau$ include the ℓ_1- and ℓ_∞-norms ($\tau = 0, 1$, respectively). According to Gotoh and Uryasev (2013), the tight bounds, L and U, are given by

$$g_n(\tau) \leq \left\{ \begin{array}{c} (\!(\boldsymbol{x})\!)_\tau / \|\boldsymbol{x}\|_2 \\ \|\boldsymbol{x}\|_2 / (\!(\boldsymbol{x})\!)^\circ_\tau \end{array} \right\} \leq \sqrt{(n-1)(1-\tau)^2+1},$$

where

$$g_n(\tau) := \left\{ \begin{array}{ll} \min\left\{ \frac{(1-\tau)\lfloor \frac{\tau}{1-\tau} \rfloor + \tau}{\sqrt{\lfloor \frac{\tau}{1-\tau} \rfloor}}, \frac{(1-\tau)\lfloor \frac{\tau}{1-\tau} \rfloor + 1}{\sqrt{\lfloor \frac{\tau}{1-\tau} \rfloor + 1}} \right\}, & 0 \leq \tau < \frac{n}{n+1}, \\ \frac{n-(n-1)\tau}{\sqrt{n}}, & \frac{n}{n+1} \leq \tau \leq 1. \end{array} \right.$$

The ratio U/L, i.e., $h_n(\tau)$, is minimized approximately at $\tau \approx \sqrt{n}/(\sqrt{n}+1)$ and the minimum ratio is $\sqrt{\frac{\sqrt{n}+1}{2}}$.[27] See Gotoh and Uryasev (2013).

[26] On the other hand, we should note that the decision function on the basis of the parallelism between $\boldsymbol{\lambda}^\star$ and $\boldsymbol{w}^\star$, i.e., $d(\boldsymbol{x}) = \mathrm{sign}(\eta \boldsymbol{x}^\top \boldsymbol{G}\boldsymbol{\lambda}^\star - b^\star)$, is now just an approximation. This point will be elaborated on again in Section D.

[27] Let $\tau_n^\star$ denote the (unique) minimizer of $h_n(\tau) := \sqrt{(n-1)(1-\tau)^2+1}/g_n(\tau)$

Table 4: Comparison of the best proximities of CVaR norms and deltoidal norms.

	ℓ_1- and ℓ_∞-norms $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_\infty$			CVaR norms with $\alpha^\star$ $\langle\!\langle \boldsymbol{x} \rangle\!\rangle_{\alpha^\star}$ and $\langle\!\langle \boldsymbol{x} \rangle\!\rangle^\circ_{\alpha^\star}$				deltoidal norms with $\tau^\star$ $(\!(\boldsymbol{x})\!)_{\tau^\star}$ and $(\!(\boldsymbol{x})\!)^\circ_{\tau^\star}$			
n	U/L	L	U	$\alpha^\star$	U/L	L	U	$\tau^\star$	U/L	L	U
2	1.4142	1	1.4142	0.2929	1.0824	1.0000	1.0824	0.5858	1.0824	1.0000	1.0824
3	1.7321	1	1.7321	0.4226	1.2393	1.0000	1.2393	0.5858	1.1589	1.0000	1.1589
5	2.2361	1	2.2361	0.5528	1.4338	1.0000	1.4338	0.7101	1.2673	0.9121	1.1559
10	3.1623	1	3.1623	0.6838	1.7396	1.0000	1.7396	0.7760	1.4412	0.8360	1.2048
100	10.0000	1	10.0000	0.9000	3.1623	1.0000	3.1623	0.9091	2.3452	0.5750	1.3484
1000	31.6228	1	31.6228	0.9684	5.6025	1.0000	5.6025	0.9692	4.0386	0.3454	1.3950

Table 4 demonstrates the values of U/L, U and L together with the corresponding parameters $\alpha^\star$ and $\tau^\star$ for several n. In general, the deltoidal norms attain smaller U/L than the CVaR norms, which implies that the former is more likely to be closer to the ℓ_2-norm solution. On the other hand, L attaining the minimum U/L is no longer equal to 1 for the deltoidal norms, whereas the L of the CVaR norms is 1 independently of n. This indicates that with CVaR norm, we can evaluate the optimal value proximity even if $\mathcal{F}$ is not positively homogeneous, as suggested in Theorem 8.

6.4 Numerical example

This section demonstrates with a numerical example the theory developed in this paper for the LP-representable norms.

Proximity values. Table 5 shows the four types of ratios, $\phi^{\mathrm{E}}/\hat{\phi}$, $\phi^{\mathrm{E}}/(\hat{\phi}/\|\hat{\boldsymbol{w}}\|_2)$ and $-\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2/\phi^{\mathrm{E}}$, in the inequalities (39) and (40) when the ν-SVMs with six LP-representable norms, i.e., the ℓ_1-, ℓ_∞-, CVaR, dual CVaR, deltoidal and dual deltoidal norms with $\alpha^\star$ or $\tau^\star$, are applied to the WDBC breast cancer data set (Wolberg, Street, and Mangasarian, 1995), which consists of 569(= m) samples and 30(= n) attributes. The table reports the statistics of the ratios over 14 different values of ν, as well as the theoretical bounds U, L or U/L.

Attained values are ranged within intervals as the theory tells. The realized values with each norm tend to be in the same order as in that of the theoretical bounds. In fact, if we employ either of deltoidal norms with $\tau = \tau^\star$, i.e., $(\!(\cdot)\!)_{\tau^\star}$ or $(\!(\cdot)\!)^\circ_{\tau^\star}$, the resulting classifiers are expected to much closer to the usual ℓ_2-norm case than the ℓ_1- or ℓ_∞-regularization, $\|\cdot\|_1$ or $\|\cdot\|_\infty$.

Table 6 shows the correlation coefficients between the optimal solutions to the ℓ_2-case, $\boldsymbol{w}_2$, and an LP-representable norm case, $\hat{\boldsymbol{w}}$. We can confirm that obtained solutions with deltoidal norm (or CVaR norm) stably attained classifiers resembling the ℓ_2-case, i.e., the standard ν-SVM.

Regularizer tuning. Another use of the generalized norms, such as $\langle\!\langle\cdot\rangle\!\rangle_\alpha, \langle\!\langle\cdot\rangle\!\rangle^*_\alpha, (\!(\cdot)\!)_\tau, (\!(\cdot)\!)^*_\tau$, can be found in the tuning of the regularizers. Namely, finding a better α or τ depending on the data at hand may improve the out-of-sample performance.

Table 7 reports results of leave-one-out cross validation, comparing the ℓ_2-norm and the CVaR norm tuning for the WDBC breast cancer data set Wolberg, Street, and Mangasarian (1995). The parameter α of the CVaR norm was chosen from among $0.000, 0.817$ and 0.967, each corresponding to the ℓ_1-norm,

Case 1) n satisfies $\hat{\tau}_{k-1} \le \frac{n-k}{n-1} \le \hat{\tau}_k$ with some $k \in \{1, ..., n-1\}$. $h_n(\tau)$ uniquely attains its minimum value $\sqrt{k(n-1)/\{n+k(k-2)\}}$ at $\tau^\star = \frac{n-k}{n-1}$, where $\hat{\tau}_k := \{(k+1)\sqrt{k} - k\sqrt{k+1}\}/\{k\sqrt{k} - (k-1)\sqrt{k+1}\}$, $k = 0, 1, ..., n-1$.

Case 2) n satisfies $\frac{n-1-k}{n-1} \le \hat{\tau}_k \le \frac{n-k}{n-1}$ with some some $k \in \{1, ..., n-1\}$. $h_n(\tau)$ uniquely attains its minimum value $\sqrt{(n-1)\left\{(k+1)\sqrt{k} - k\sqrt{k+1}\right\}^2 + \left\{k\cdot\sqrt{k} - (k-1)\sqrt{k+1}\right\}^2}$ at $\tau = \hat{\tau}_k$.

Table 5: Proximity values via $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \|\cdot\|)$ ($\nu = 0.05, 0.10, ..., 0.70$; the WDBC breast cancer data set (Wolberg, Street, and Mangasarian, 1995))

(a) $\phi^{\mathrm{E}}/\hat{\phi}$ or $\hat{\phi}/\phi^{\mathrm{E}}$

$\|\cdot\|$	$\|\cdot\|_1$	$\|\cdot\|_\infty$	$\langle\!\langle\cdot\rangle\!\rangle_{\alpha^\star}$	$\langle\!\langle\cdot\rangle\!\rangle^\circ_{\alpha^\star}$	$(\!(\cdot)\!)_{\alpha^\star}$	$(\!(\cdot)\!)^\circ_{\alpha^\star}$
U	5.4772		2.2864		1.3005	
max.	2.9804	3.2972	1.7967	1.9792	1.0190	1.0472
med.	1.1491	1.5330	1.1491	1.5044	0.8655	0.9765
min.	1.0191	1.3777	1.0191	1.3209	0.8144	0.8961
L	1.0000		1.0000		0.7234	

(b) $\phi^{\mathrm{E}}/(\hat{\phi}/\|\hat{\boldsymbol{w}}\|_2)$

U/L	5.4772	5.4772	2.2864	2.2864	1.7977	1.7977
max.	1.5959	3.9757	1.2948	1.7309	1.1853	1.3407
med.	1.1409	3.5729	1.1286	1.5198	1.0574	1.2959
min.	1.0191	1.1567	1.0191	1.0989	1.0141	1.1317

(c) $-\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2/\phi^{\mathrm{E}}$

U/L	5.4772	5.4772	2.2864	2.2864	1.7977	1.7977
max.	1.5670	1.2945	1.1622	1.3193	1.1713	1.0993
med.	1.0039	1.0038	1.0039	1.0037	1.0025	1.0026
min.	1.0013	1.0003	1.0013	1.0001	1.0001	1.0001

(d) $\boldsymbol{w}^\top\boldsymbol{G}\hat{\boldsymbol{\lambda}}/(\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2\|\hat{\boldsymbol{w}}\|_2)$

max.	0.9795	0.7774	0.9795	0.8976	0.9848	0.8051
med.	0.8734	0.2797	0.8734	0.6577	0.9404	0.7602
min.	0.3999	0.2475	0.7397	0.5688	0.7607	0.7451
L/U	0.1826	0.1826	0.4374	0.4374	0.5562	0.5562

This table reports the four types of ratios, $\phi^{\mathrm{E}}/\hat{\phi}$, $\phi^{\mathrm{E}}/(\hat{\phi}/\|\hat{\boldsymbol{w}}\|_2)$, $-\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2/\phi^{\mathrm{E}}$ and $\hat{\boldsymbol{w}}^\top\boldsymbol{G}\hat{\boldsymbol{\lambda}}/(\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2\|\hat{\boldsymbol{w}}\|_2)$. The first three ratios are associated with the inequalities (39), while the last one are with (40). The theoretical ranges are conservative, but their order reflects the order of the numerical results. Indeed, results by the deltoidal norms show smaller variability than the other pairs.

the CVaR norm with $\alpha^\star$ and the ℓ_∞-norm, respectively, so that average error in the validation samples would be minimized.

For each ν, the CVaR norm found classifiers which attained lower out-of-sample error than the ℓ_2-norm case. Important is that the total sum of elapsed time for solving three times LPs of the CVaR norm tuning is almost equal to that for solving a QP of the ℓ_2-norm case.[28] In this sense, there is no reason for persisting in the use of the ℓ_2-norm for the linear classification problems.

7 Concluding remarks

This paper studies formulations of SVMs for binary classification in a unified way, focusing on the capability of non-ℓ_2-norms. To that purpose, we mainly employ the Ivanov regularization for the primal formulation, and characterize a pair of primal and dual formulations for a generalized SVM by a pair $(\mathcal{F}, \|\cdot\|)$, where $\mathcal{F}$ is a convex function for expressing the empirical risk to be minimized in the primal formulation, and $\|\cdot\|$ is an arbitrary norm for describing the regularizing constraint in the primal formulation. We admit a large freedom for the choice of $\mathcal{F}$, but we limit the loss to be linear with respect to the involved parameters.

With the formulation, we enjoy many features.

[28] For a fair comparison, we apply a decomposed formulation, which is applied to the LP-representable norms, also to the ℓ_2-norm case. (Indeed, this decomposition is obtained as (74) with $\boldsymbol{\Gamma} = \boldsymbol{G}$ and $r = m$ without eigenvalue computation.) Without this decomposition, i.e., putting the QP formulation (7) into the solver software as it is, it took about four times as long as the CVaR norm formulations.

Table 6: Correlation coefficients of the solutions between $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \|\cdot\|_2)$ and $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \|\cdot\|)$ ($\nu = 0.05, 0.10, ..., 0.70$; the WDBC breast cancer data set (Wolberg, Street, and Mangasarian, 1995))

(a) Correlation coefficient between $\boldsymbol{w}_2$ and $\hat{\boldsymbol{w}}$

$\|\cdot\|$	$\|\cdot\|_1$	$\|\cdot\|_\infty$	$\langle\!\langle\cdot\rangle\!\rangle_{\alpha^\star}$	$\langle\!\langle\cdot\rangle\!\rangle^\circ_{\alpha^\star}$	$(\!(\cdot)\!)_{\alpha^\star}$	$(\!(\cdot)\!)^\circ_{\alpha^\star}$
max.	0.9823	0.8671	0.9823	0.9165	0.9864	0.8971
med.	0.8837	0.2805	0.8902	0.6581	0.9499	0.7741
min.	0.6658	0.2562	0.7734	0.5876	0.8494	0.7460

(b) Correlation coefficient between $\boldsymbol{\lambda}_2$ and $\hat{\boldsymbol{\lambda}}$

max.	0.9910	0.9954	0.9910	0.9962	0.9948	0.9968
med.	0.9835	0.9851	0.9849	0.9877	0.9905	0.9869
min.	0.9378	0.9410	0.9741	0.9505	0.9668	0.9631

This table reports the correlation coefficients between the obtained solutions with the ℓ_2- and each of LP-representable norms. (a) reports the case for the primal solutions, i.e., the ℓ_2-norm solutions $\boldsymbol{w}_2$ vs. an LP-representable norm solutions $\hat{\boldsymbol{w}}$, while (b) for the dual solutions, i.e., $\boldsymbol{\lambda}_2$ vs. $\hat{\boldsymbol{\lambda}}$.

- Popular binary classification methods are characterized based on the function form of $\mathcal{F}$. Users can easily customize SVM by choosing a favorable combination of $\mathcal{F}$ and $\|\cdot\|$ from among a virtually infinite number of candidates.

- Most of results have been derived in a unified manner on the basis of the Fenchel duality. This approach is advantageous over the Lagrange duality, as several researches claim.

- We can see what properties of $\mathcal{F}$ or regularizer matters for the tractability of the resulting formulations. Especially, corresponding to functional properties of $\mathcal{F}$, we can express the dual formulation in an interpretable way. Indeed, with the general formulation, the existing geometric interpretations and the robust optimization modelings are extended.

- As non-ℓ_2-regularizer, we elaborate on the use of two parametrized families of polyhedral norms. With those families, tuning of the regularizer can be implemented without significant increase of the computational complexity from the ℓ_1-norm regularization. Especially, this is advantageous over the ℓ_p-norms.

- If the proposed (Ivanov regularization-based) formulation employs the ℓ_2-norm, it provides the same classifier as the standard (Tikhonov regularization-based) formulation does. In this sense, the ℓ_2-norm remains to be central. Based on the tight bounds of the proximity between a non-ℓ_2-norm and the ℓ_2-norm, i.e., U/L, we can guarantee the closeness of the associated solutions to the ℓ_2-norm case. We show that the ratio U/L indicates the closeness to the ℓ_2-formulations, even when the kernelization is incorporated (see Appendix D for the treatment of the kernelization with non-ℓ_2-norms).

Using softwares which build in risk functionals and norms, the generalized formulation enriches applications of SVMs in practice. On the other hand, in general, such a generic algorithm or software may be less attractive than existing algorithms which have been developed for a long time and made use of the special structure of $\mathcal{F}$ and/or $\iota(\|\cdot\|)$. However, the analysis focusing on the functional properties, as developed in this paper, may be helpful to enhance the development of new algorithms.

In this paper, we supposed that the argument of $\mathcal{F}$ was of the form $\boldsymbol{L} = -(\boldsymbol{G}\boldsymbol{\theta} - \boldsymbol{y}b)$ and regularization was added for $\boldsymbol{\theta}$. Nevertheless, we can treat a variety of nonlinear classification criteria. On the other hand, excluded class of risk functional 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 $y_i = +1$, with $\Omega = \{\omega_{(i,j)} : y_i = -1, y_j = +1, i, j = 1, ..., m\}$,

remains to be investigated.

Table 7: Leave-one-out cross-validation: ℓ_2-norm vs. CVaR norm with tuning $\alpha \in \{0.000, 0.817, 0.967\}$ (the WDBC breast cancer data set (Wolberg, Street, and Mangasarian, 1995))

(a) ℓ_2-norm (ordinary ν-SVM)

	LOO ave. errors		time [sec.]	
ν	in sample	out-of-sample	sum	ave.
0.01	0.05%	5.45%	21.286	0.0374
0.05	1.24%	4.39%	21.474	0.0377
0.10	3.37%	4.75%	20.493	0.0360
0.30	9.49%	9.49%	15.935	0.0280
0.50	13.18%	13.18%	17.739	0.0312
0.70	21.74%	21.97%	32.899	0.0578

(b) CVaR norm with tuning α

	chosen	LOO ave. errors		time [sec.]	
ν	α	in sample	out-of-sample	sum	ave.
0.01	0.817	0.00%	5.10%	27.580	0.0162
0.05	0.967	1.23%	3.69%	28.253	0.0166
0.10	0.000	3.52%	4.39%	28.032	0.0164
0.30	0.967	9.14%	9.14%	19.611	0.0115
0.50	0.000	13.05%	13.18%	16.832	0.0099
0.70	0.967	20.92%	21.27%	16.661	0.0098

The results in the above table are based on solutions to the (reduced) dual forms of $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \|\cdot\|_2)$ and $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha)$. All the computation was done on a laptop PC with Intel(R) Core(TM) i7-3612QM CPU, 2.10 GHz, 8.00 GB RAM, Windows 7 (64 bit) OS, and was implemented in MATLAB, and IBM ILOG CPLEX 12.5 was employed for solving the quadratic programs (for the ℓ_2-norm case) and LPs (for the six LP-representable norm cases). 'sum' denotes the total computation time in second for solving 569×3 instances at each ν, while 'ave.' denotes the average for each instance. This table reveals that using a class of LP-representable norms with a parameter, α, can find a better classifier (the column 'out-of-sample') than the ℓ_2-norm case (i.e., ordinary ν-SVM), which is implemented by solving QP, in a comparably efficient manner (the column 'sum') even with tuning of the parameter.

The framework can be extended to the other types of machine learning tasks, such as one-class and multi-class 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.

Acknowledgment. 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 a MEXT Grant-in-Aid for Young Scientists (B) 23710176. 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 Prof. Mike Zabarankin for his suggestion on the risk envelope of mixed CVaR.

A A collection of risk functionals

A.1 Basic properties of CVaR (Rockafellar and Uryasev, 2000)

By the duality theory of LP, CVaR is rewritten by

$$\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \max_{\boldsymbol{q}} \{\mathbb{E}_{\boldsymbol{q}}(\boldsymbol{L}) : \mathbf{1}_m^\top \boldsymbol{q} = 1, \mathbf{0} \le \boldsymbol{q} \le \tfrac{1}{1-\alpha}\boldsymbol{p}\}. \tag{46}$$

This is exactly the dual representation (Corollary 2) of CVaR, and explains its risk envelope (13). Observing that (46) is a so-called continuous knapsack problem, we have

$$\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \tfrac{1}{1-\alpha}\sum_{i=1}^{k} p_{(i)}L_{(i)} + \tfrac{1}{1-\alpha}(1-\alpha-\sum_{i=1}^{k} p_{(i)})L_{(k+1)},$$

where $L_{(i)}$ indicates the i-th largest component of $\boldsymbol{L}$, i.e., $L_{(1)} \geq L_{(2)} \geq \cdots \geq L_{(m)}$, $p_{(i)}$ is the corresponding reference probability, and k is integer satisfying $\sum_{i=1}^{k} p_{(i)} \leq 1-\alpha < \sum_{i=1}^{k+1} p_{(i)}$. By omitting $1-\alpha-\sum_{i=1}^{k} p_{(i)}$, CVaR is approximately equal to the mean of the largest $100(1-\alpha)$ % losses $L_{(1)}, ..., L_{(k)}$. More precisely, by defining the value-at-risk (VaR) by

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{VaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) := \min_{c}\{c : \mathbb{E}_{\boldsymbol{p}}([\boldsymbol{L} \leq c\mathbf{1}_m]) \geq \alpha\}, \quad \text{with } \alpha \in (0,1),$$

CVaR can also be rephrased to be approximately equal to the conditional expectation of loss above VaR. In fact, the minimizer, $c^\star$, in the definition of CVaR is known to provide an approximate VaR (see Rockafellar and Uryasev, 2002, for the details). While CVaR is convex, VaR is not and is not treated in this paper.

CVaR includes the mean loss and the maximum loss, as special limiting cases. Indeed, if $\alpha = 0$, it is equal to the mean loss, i.e., $\mathrm{CVaR}_{(0,\boldsymbol{p})}(\boldsymbol{L}) = \mathrm{Mean}_{\boldsymbol{p}}(\boldsymbol{L})$; If $\alpha \geq 1 - \min_i p_i$, it is equal to the maximum loss, i.e., $\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \mathrm{Max}(\boldsymbol{L})$. These facts can also be confirmed based on the inclusion of the risk envelopes, i.e., $\{\boldsymbol{p}\} \equiv \mathcal{Q}_{\mathrm{Mean}(\boldsymbol{p})} = \mathcal{Q}_{\mathrm{CVaR}(0,\boldsymbol{p})} \subset \mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})} \subset \mathcal{Q}_{\mathrm{CVaR}(\alpha',\boldsymbol{p})} = \mathcal{Q}_{\mathrm{Max}} \equiv \mathbb{\Pi}^m$ for $\alpha' \geq 1 - \min_i p_i$ (see Section A.2 for $\mathcal{Q}_{\mathrm{Mean}(\boldsymbol{p})}$ and $\mathcal{Q}_{\mathrm{Max}}$).

A.2 Further examples of convex risk functionals.

A number of risk functionals are widely used in various fields and it is impossible to provide all of them. We below list several important functionals and suggest how to create another one by applying basic operations for combination, which might indicate the way to create a new SVM.

1. Maximum loss. The max loss can be linked to the maximum margin SVM (or hard-margin SVM) (Boser, Guyon, and Vapnik, 1992).

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{Max}(\boldsymbol{L}) := \max\{L_1, ..., L_m\}. \tag{47}$$

It is not hard to see that this measure is coherent.

$$\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{Max}} := \left\{\boldsymbol{q} \in \mathbb{R}^m : \mathbf{1}_m^\top \boldsymbol{q} = 1, \boldsymbol{q} \geq \mathbf{0}\right\} \equiv \mathbb{\Pi}^m. \tag{48}$$

In contrast to the other functionals, this does not require the reference probability $\boldsymbol{p}$.

2. Mean loss. Mean is also a coherent risk functional and it is used in various situations.[29]

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{Mean}_{\boldsymbol{p}}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) = \boldsymbol{p}^\top \boldsymbol{L},$$

$$\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{Mean}(\boldsymbol{p})} := \{\boldsymbol{p}\}. \tag{49}$$

In this sense, the mean loss has the smallest (nonempty) risk envelope among the coherent risk functionals.

3. Above-target-loss (Below-target-return (Fishburn, 1977)). If a fixed target, say t, is given for the loss, upper partial moment gives a monotonic convex risk functional.

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{ATL}_{(t,\boldsymbol{p})} := \mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L} + t\mathbf{1}_m)_+), \text{ with } t \in \mathbb{R}. \tag{50}$$

We can associate this functional with the C-SVM (1). On the other hand, this functional lacks translation-invariance and positive homogeneity, and thus not coherent.[30] This risk functional may be difficult to interpret unless the target t of loss can be clearly interpreted.

[29] Mean is a coherent risk functional, but not a (regular) measure of risk in the sense of the risk quadrangle developed in Rockafellar and Uryasev (2013) since it does not satisfy the aversion property.

[30] Besides, it is not a (regular) measure of risk in the risk quadrangle sense (Rockafellar and Uryasev, 2013).

4. Deviation-type functionals. In the case where the dispersion of losses is of central concern, the variance or standard deviation is a popular choice for the risk functional.

$$\mathcal{F}(\boldsymbol{L}) = \mathbb{V}_{\boldsymbol{p}}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L} - \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L})\mathbf{1}_m)^{\cdot 2}); \quad \mathrm{SD}_{\boldsymbol{p}}(\boldsymbol{L}) := \sqrt{\mathbb{V}(\boldsymbol{L})}.$$

However, the variance does not satisfy the monotonicity, translation invariance or positive homogeneity (and therefore not coherent). On the other hand, standard deviation satisfy the positive homogeneity, although neither monotonicity nor translation invariance is satisfied. As an alternative to variance or standard deviation, absolute deviation is sometimes used due to the tractability in its minimization.

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{AD}_{\boldsymbol{p}}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(|\boldsymbol{L} - \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L})\mathbf{1}_m|).$$

An interesting class of deviation-type functionals is characterized by the notion of the so-called *M-estimator*, which is originally developed in the context of robust statistics (e.g., Huber, 1981). An M-estimator risk is defined with a convex function ρ by

$$\mathcal{F}(\boldsymbol{L}) = \min_{\theta} \boldsymbol{p}^{\top} \rho(\boldsymbol{L} - \theta\mathbf{1}_m) \equiv \min_{\theta} \sum_{i=1}^{m} p_i \rho(L_i - \theta).$$

If ρ is given by $\rho(L) = |L|$, i.e., $\mathcal{F}(\boldsymbol{L}) = \mathrm{MeD}_{\boldsymbol{p}}(\boldsymbol{L}) := \min_{\theta} \mathbb{E}_{\boldsymbol{p}}(|\boldsymbol{L} - \theta\mathbf{1}_m|)$, it is known that $\mathcal{F}(\boldsymbol{L})$ represents the absolute deviation of $\boldsymbol{L}$ around the median (under the distribution $\boldsymbol{p}$), whereas if ρ is given by $\rho(L) = |L|^2$, i.e., $\mathcal{F}(\boldsymbol{L}) = \min_{\theta} \mathbb{E}_{\boldsymbol{p}}(|\boldsymbol{L} - \theta\mathbf{1}_m|^{\cdot 2})$, it is known that the optimal θ defines the mean of $\boldsymbol{L}$ and, accordingly, $\mathcal{F}(\boldsymbol{L})$ itself is equivalent to $\mathbb{V}_{\boldsymbol{p}}(\boldsymbol{L})$. Another interesting example of this kind is known as the Huber loss, which can be given (with a slight generalization) with $\rho = h_{(l,r)}$:

$$h_{(l,r)}(L) = \begin{cases} r(L - \frac{r}{2}), & L \geq r, \\ L^2/2, & l \leq L \leq r, \\ l(-L - \frac{l}{2}), & l \leq L, \end{cases} \tag{51}$$

with two constants l and r such that $l < r$. We denote this by $\mathrm{Huber}_{(l,r,\boldsymbol{p})}(\boldsymbol{L})$. Especially if $l < 0 < r$, this function can be a composite of $\mathbb{V}_{\boldsymbol{p}}(\boldsymbol{L})$ and $\mathrm{MeD}_{\boldsymbol{p}}(\boldsymbol{L})$ because quadratic penalty is assigned near a "center," i.e., the optimal θ^*, while linear penalty is assigned on the loss L above r or below l. In the context of SVM, this functional has been used (e.g., Chapelle, 2007, where, however, $l = 0$ is used[31]).

7. Gini mean difference and mixed CVaR. The Gini's mean difference is popular in measuring the inequality of income distribution in a society and is proposed to be used also in the financial optimization (e.g., Yitzhaki, 1982). Although several equivalent representations are known, with our notation it can be simply defined by

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{Gini}_{\boldsymbol{p}}(\boldsymbol{L}) := \tfrac{1}{2} \sum_{h=1}^{m} \sum_{i=1}^{m} |L_i - L_j| p_i p_j.$$

We can readily observe that this risk functional is positively homogeneous convex, but not monotonic or translation invariant. Following Zabarankin and Uryasev (2013), the Gini mean difference can be linked to the *mixed CVaR*, which is defined by

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{mixCVaR}_{(\tau(\cdot),\boldsymbol{p})}(\boldsymbol{L}) := \int_0^1 \mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L})\, \mathrm{d}\tau(\alpha),$$

with a function $\tau(\cdot)$ satisfying $\int_0^1 \mathrm{d}\tau(\alpha) = 1$ and $\mathrm{d}\tau(\alpha) \geq 0$. Indeed, with $\tau(\alpha) = -(1-\alpha)^2$, we have $\mathrm{Gini}_{\boldsymbol{p}}(\boldsymbol{L}) = \mathrm{mixCVaR}_{(\tau(\cdot),\boldsymbol{p})}(\boldsymbol{L}) - \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L})$.[32]

[31] This monotonic Huber-type is mentioned in Rockafellar and Uryasev (2013) as a regular measure of risk in the Truncated-Mean-Based Quadrangle.

[32] This relation implies that the Gini mean difference is the deviation measure forming the risk quadrangle

Convexity-preserving operations. Many new convex risk functionals can be created by applying the convexity-preserving and/or convexity-creating operations (e.g., Rockafellar, 1970; Boyd and Vandenberghe, 2004). For example, for any two convex risk functionals $\mathcal{F}_1, \mathcal{F}_2$ and any nonnegative constants a_1, a_2, the weighted sum $a_1\mathcal{F}_1 + a_2\mathcal{F}_2$ is also a convex risk functional; the max operation, $\max\{\mathcal{F}_1, \mathcal{F}_2\}$, also creates another convex risk functional.

8. Mean-standard deviation. In financial context, the composite of the mean and standard deviation plays an important role (Markowitz, 1952).

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{MSD}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + t \cdot \mathrm{SD}_{\boldsymbol{p}}(\boldsymbol{L}), \text{ with } t \geq 0. \tag{52}$$

9. Mean-semideviation, MASD and MAD. With variance (or standard deviation), the lower deviation from the mean is also considered to be averse. In order to fix this excessive aversion, the lower semi-deviation functionals have been developed.

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{M}k\mathrm{SD}_{(k,t)}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + t(\mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L} - \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L})\mathbf{1}_m)_+{}^{\cdot k}))^{1/k}, \text{ with } t \geq 0;\ k \geq 1.$$

The case, $k = 1$ is known as the mean-absolute deviation (MAD) or the mean-absolute semideviation (MASD) model in the financial context (Konno and Yamazaki, 1991).

$$\begin{aligned}
\mathcal{F}(\boldsymbol{L}) = \mathrm{MASD}_{(t,\boldsymbol{p})}(\boldsymbol{L}) &:= \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + t \cdot \mathbb{E}_{\boldsymbol{p}}((\boldsymbol{L} - \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L})\mathbf{1}_m)_+), \quad \text{with } t \geq 0, \\
&\equiv\ \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + \tfrac{t}{2} \cdot \mathbb{E}_{\boldsymbol{p}}(|\boldsymbol{L} - \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L})\mathbf{1}_m|) =: \mathrm{MAD}_{(t/2,\boldsymbol{p})}(\boldsymbol{L}); \\
\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{MASD}(t,\boldsymbol{p})} &= \{\boldsymbol{q} \in \mathbb{R}^m : \boldsymbol{q} = \boldsymbol{p} + \boldsymbol{u} - \boldsymbol{p}\mathbf{1}_m^\top \boldsymbol{u} \text{ for some } \boldsymbol{u} \text{ s.t. } \mathbf{0} \leq \boldsymbol{u} \leq \boldsymbol{p}\}.
\end{aligned}$$

This envelope monotonically increases in t, i.e., $\mathcal{Q}_{\mathrm{MASD}(t_1,\boldsymbol{p})} \subset \mathcal{Q}_{\mathrm{MASD}(t_2,\boldsymbol{p})}$ with $t_1 < t_2$, and $\mathcal{Q}_{\mathrm{MASD}(0,\boldsymbol{p})} = \{\boldsymbol{p}\}$. For $t \in [0, 1]$, MASD is coherent.

10. Mean-Gini. As the above Examples 8 and 9, the addition of the mean to the Gini creates a coherent risk functional. In fact, with $t \in [0, 1]$, the mean-Gini functional (Yitzhaki, 1982):

$$\mathcal{F}(\boldsymbol{L}) = \mathrm{MGini}_{(t,\boldsymbol{p})}(\boldsymbol{L}) := \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + t \cdot \mathrm{Gini}_{\boldsymbol{p}}(\boldsymbol{L})$$

is coherent. From the relation between the Gini mean difference and the mixed CVaR with $\tau(\alpha) = -(1-\alpha)^2$, we have $\mathrm{mixCVaR}_{(-(1-(\cdot))^2,\boldsymbol{p})}(\boldsymbol{L}) = \sum_{i=1}^m c_i L_{(i)}$ with $c_1 = p_{(1)}(1 + \sum_{i=2}^m p_{(i)})$; $c_m = p_{(m)} - \sum_{i=2}^m p_{(i-1)}p_{(i)}$; $c_i = p_{(i)}(1 + \sum_{j=i+1}^m p_{(j)}) - \sum_{j=2}^i p_{(j-1)}p_{(j)}$ for $i = 2, ..., m-1$. This implies that the risk envelope of the mixed CVaR with $\tau(\alpha) = -(1-\alpha)^2$, i.e., $\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{mixCVaR}(-(1-(\cdot))^2,\boldsymbol{p})}$, is given by the set of probabilities assigning the weight c_i on the i-th largest component of a loss $\boldsymbol{L}$. Since $\mathcal{F}(\boldsymbol{L}) = \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + t \cdot \mathrm{Gini}_{\boldsymbol{p}}(\boldsymbol{L}) = (1-t)\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + t \cdot \mathrm{mixCVaR}_{(-(1-\cdot)^2,\boldsymbol{p})}$, the risk envelope of Mean-Gini is given by $\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{MGini}(t,\boldsymbol{p})} = (1-t)\boldsymbol{p} + t\mathcal{Q}_{\mathrm{mixCVaR}(-(1-(\cdot))^2,\boldsymbol{p})}$.

It is noteworthy that above examples show that addition of a monotonic risk functional can provide a deviation-type (nonmonotonic) one with monotonicity. This implies a way to create a new SVM with monotonic risk functional on the basis of deviation-type functionals.

A.3 Risk envelope-induced risk functionals.

The first statement of Corollary 2 suggests that any convex set in $\mathbb{R}^m$ can induce a positively homogeneous convex risk functional.

(Rockafellar and Uryasev, 2013) with the risk measure $\mathrm{mixCVaR}_{(-(1-(\cdot))^2,\boldsymbol{p})}(\boldsymbol{L})$. To measure the inequality of income distribution, say, $\boldsymbol{W}$, the formula should be applied to $\boldsymbol{L} = -\boldsymbol{W}$. The Gini mean difference of $\boldsymbol{W}$ is computed by $\mathrm{Gini}_{\boldsymbol{p}}(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{W}) - \mathrm{mixCVaR}_{(\tau(\cdot),\boldsymbol{p})}(\boldsymbol{W})$, and the so-called Gini index is given by dividing the Gini mean difference by the average income, i.e., $\mathrm{Gini}_{\mathbf{1}_m/m}(\boldsymbol{W})/\mathbb{E}_{\mathbf{1}_m/m}(\boldsymbol{W})$.

1. **Distance-based envelope.** Consider $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|,\boldsymbol{A},\boldsymbol{p})}$, which is defined in Section 5.3.

   As risk envelope, $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|,\boldsymbol{A},\boldsymbol{p})}$corresponds to the risk functional of the form: $\mathcal{F}(\boldsymbol{L}) = \min_{\boldsymbol{\lambda},\theta}\{\mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}+\boldsymbol{\lambda}) + \|\boldsymbol{A}^\top(\boldsymbol{L}+\boldsymbol{\lambda}-\theta\mathbf{1}_m)\|^\circ : \boldsymbol{\lambda}\geq\mathbf{0}\}$, where $\|\cdot\|^\circ$ is the dual norm to $\|\cdot\|$. Note that if the non-negativity is dropped off from the definition of $\mathcal{Q}_{\mathrm{Dist}(\|\cdot\|,\boldsymbol{A},\boldsymbol{p})}$, i.e., $\mathcal{Q}_{\mathcal{F}} = \{\boldsymbol{q}\in\mathbb{R}^m : \mathbf{1}_m^\top\boldsymbol{q}=1, \boldsymbol{q} = \boldsymbol{p}+\boldsymbol{A}\boldsymbol{\zeta}, \|\boldsymbol{\zeta}\|\leq 1\}$, we instead obtain $\mathcal{F}(\boldsymbol{L}) = \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + \min_\theta\{\|\boldsymbol{A}^\top(\boldsymbol{L}-\theta\mathbf{1}_m)\|^\circ\}$, which lacks the monotonicity. In particular, if we set $\|\cdot\| = \|\cdot\|_2, \boldsymbol{A} = \sqrt{t}\cdot\mathrm{diag}(\sqrt{\boldsymbol{p}})$, this results in the mean-standard deviation $\mathrm{MSD}_{(t,\boldsymbol{p})}$, defined by (52). If the constraint $\mathbf{1}_m^\top\boldsymbol{q}=1$ is further dropped off, we obtain $\mathcal{F}(\boldsymbol{L}) = \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + \|\boldsymbol{A}^\top\boldsymbol{L}\|^\circ$.

2. **φ-divergence envelope.** Also, considering $\mathcal{Q}_{\mathcal{I}_\varphi(t,\boldsymbol{p})}$, which is also defined in Section 5.3, we can construct a positively homogeneous risk functional. In particular, the KL divergence, i.e., $\mathcal{Q}_{\mathrm{KL}(t,\boldsymbol{p})}$, can be related to $\mathrm{LSE}_{(t,\boldsymbol{p})}$ since $\mathcal{Q}_{\mathrm{KL}(t,\boldsymbol{p})}$ is defined with its conjugate $\mathrm{KL}_{(t,\boldsymbol{p})}$. It contrasts with that $\mathrm{LSE}^*_{(t,\boldsymbol{p})}\neq 0$.

A.4 Risk envelope-preserving operations.

In general, given a set of convex sets, we can create another convex set via some operation and, accordingly, obtain a corresponding proper l.s.c. positively homogeneous convex risk functional. For example, the convex combination of any finite number of coherent risk functionals is coherent. Let $\mathcal{F}_1,...,\mathcal{F}_K$ be proper l.s.c. positively homogeneous convex risk functionals, and $\mathcal{Q}_1,...,\mathcal{Q}_K$ be the set of corresponding risk envelopes. Then the convex combination of the sets, i.e.,

$$\sum_{h=1}^K \tau_h\mathcal{Q}_h := \{\sum_{h=1}^K \tau_h\boldsymbol{q}_h : \boldsymbol{q}_h\in\mathcal{Q}_h, h=1,2,...,K\},\ \text{with some } \boldsymbol{\tau}\in\mathbb{\Pi}^K,$$

yields another convex set. The corresponding risk functional is the convex combination of the risk functionals $\mathcal{F}_1,...,\mathcal{F}_K$. Namely, $\mathcal{Q}_{\mathcal{F}} = \sum_{h=1}^K \tau_h\mathcal{Q}_h \ \leftrightarrow\ \mathcal{F} = \sum_{h=1}^K \tau_h\mathcal{F}_h$. This implies that $\mathcal{F}$ is monotonic if $\mathcal{F}_1,...,\mathcal{F}_K$ are monotonic, and $\mathcal{F}$ is translation invariant if $\mathcal{F}_1,...,\mathcal{F}_K$ are translation invariant. Especially when $\mathcal{Q}_1,...,\mathcal{Q}_K\subset\mathbb{\Pi}^m$, their convex combination is also in $\mathbb{\Pi}^m$ and, accordingly, the resultant risk functional is coherent.

For example, the convex combination of the risk envelopes of the maximum loss and the mean loss yields another risk envelope, and the corresponding risk functional is the convex combination of the maximum loss and the mean loss: $\mathcal{Q}_{\mathcal{F}} = (1-\tau)\mathcal{Q}_{\mathrm{Max}}+\tau\mathcal{Q}_{\mathrm{Mean}(\boldsymbol{p})} \equiv \{\boldsymbol{q}\in\mathbb{\Pi}^m : \boldsymbol{q}\geq(1-\tau)\boldsymbol{p}\} \ \leftrightarrow\ \mathcal{F}(\boldsymbol{L}) = (1-\tau)\mathrm{Max}(\boldsymbol{L}) + \tau\mathrm{Mean}_{\boldsymbol{p}}(\boldsymbol{L})$, where $\tau\in[0,1]$.

Besides, convex hull of risk envelopes defines another risk envelope, i.e.,

$$\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{conv}(\mathcal{Q}_1,...,\mathcal{Q}_K)} := \{\boldsymbol{q}\in\mathbb{R}^m : \boldsymbol{q} = \sum_{h=1}^K \tau_h\boldsymbol{q}_h, \mathbf{1}_K^\top\boldsymbol{\tau}=1, \boldsymbol{\tau}\geq\mathbf{0}, \boldsymbol{q}_h\in\mathcal{Q}_h, h=1,...,K\}.$$

The corresponding risk functional is given by

$$\mathcal{F}(\boldsymbol{L}) = \max_{\boldsymbol{\tau}}\{\sum_{h=1}^K \tau_h\mathcal{F}_h(\boldsymbol{L}) : \mathbf{1}_K^\top\boldsymbol{\tau}=1, \boldsymbol{\tau}\geq\mathbf{0}\} \equiv \max\{\mathcal{F}_1(\boldsymbol{L}),...,\mathcal{F}_K(\boldsymbol{L})\}.$$

Especially when $\mathcal{Q}_1,...,\mathcal{Q}_K\in\mathbb{\Pi}^m$, their convex hull is also in $\mathbb{\Pi}^m$. For example, for $\mathcal{Q}_h = \{\boldsymbol{p}_h\}, h=1,...,K$ with $\boldsymbol{p}_h\in\mathbb{\Pi}^m$, the set

$$\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathrm{conv}(\{\boldsymbol{p}_1\},...,\{\boldsymbol{p}_K\})} \equiv \{\boldsymbol{q} : \boldsymbol{q} = \sum_{h=1}^K \tau_h\boldsymbol{p}_h, \mathbf{1}_K^\top\boldsymbol{\tau}=1, \boldsymbol{\tau}\geq\mathbf{0}\}$$

is again in $\mathbb{\Pi}^m$. The corresponding risk functional is given by

$$\mathcal{F}(\boldsymbol{L}) = \max\{\mathbb{E}_{\boldsymbol{p}_1}(\boldsymbol{L}),...,\mathbb{E}_{\boldsymbol{p}_K}(\boldsymbol{L})\}.$$

Applying the above operations repeatedly and/or in a combined manner, we can create an infinite number of risk functionals.

Table 8: Properties of risk functionals

risk functionals $\mathcal{F}$	convex	monotonic	trans. invar.	posi. homo.	coherent	polyhedral
Max	yes	yes	yes	yes	yes	yes
$\mathrm{Mean}_{\boldsymbol{p}}$	yes	yes	yes	yes	yes	yes
$\mathrm{ATL}_{(t,\boldsymbol{p})}$	yes	yes	no	no	no	yes
$\mathrm{VaR}_{(\alpha,\boldsymbol{p})}$	no	yes	yes	yes	no	no
$\mathbb{V}_{\boldsymbol{p}}$	yes	no	no	no	no	no
$\mathrm{SD}_{\boldsymbol{p}}$	yes	no	no	yes	no	no
$\mathrm{AD}_{\boldsymbol{p}}$	yes	no	no	yes	no	yes
$\mathrm{MeD}_{\boldsymbol{p}}$	yes	no	no	yes	no	yes
$\mathrm{Huber}_{(\boldsymbol{p},l,r)}$	yes	$l>0$	no	no	no	no
$\mathrm{Gini}_{\boldsymbol{p}}$	yes	no	no	yes	no	yes
$\mathrm{MSD}_{(t,\boldsymbol{p})}$	yes	no	yes	yes	no	no
$\mathrm{MAD}_{(t,\boldsymbol{p})}$	yes	$t\in[0,\frac{1}{2}]$	yes	yes	$t\in[0,\frac{1}{2}]$	yes
$\mathrm{MMeD}_{(t,\boldsymbol{p})}$	yes	$t\in[0,1]$	yes	yes	$t\in[0,1]$	yes
$\mathrm{MASD}_{(t,\boldsymbol{p})}$	yes	$t\in[0,1]$	yes	yes	$t\in[0,1]$	yes
$\mathrm{MGini}_{(t,\boldsymbol{p})}$	yes	$t\in[0,1]$	yes	yes	$t\in[0,1]$	yes
$\mathrm{mixCVaR}_{(-(1-(\cdot))^2,\boldsymbol{p})}$	yes	yes	yes	yes	yes	yes

B Sketches of Proofs

B.1 Proof of Theorem 3

Proof. To show the first statement, it suffices to confirm the boundedness of $\mathcal{F}_{\boldsymbol{p}}$. Indeed, we see that $\mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L}) \geq \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) + B > -\infty$. The l.s.c convexity is obvious. To see (14), observe that the right-hand side is equal to $\inf_c \sup_{\boldsymbol{\lambda}}\{\sum_{i=1}^m (L_i\lambda_i - p_i v^*(\frac{\lambda_i}{p_i})) - c(\sum_{i=1}^m \lambda_i - 1)\} = \inf_c\{c + \sum_{i=1}^m p_i \sup_{\lambda_i}\{\frac{\lambda_i}{p_i}(L_i - c) - v^*(\frac{\lambda_i}{p_i})\}\} = \inf_c\{c + \sum_{i=1}^m p_i v(L_i - c)\} = \mathcal{F}_{\boldsymbol{p}}(\boldsymbol{L})$.[33] 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 2 to v and Quadrangle Theorem (c) of Rockafellar and Uryasev (2013). □

B.2 Proof of Theorem 4

To prove Theorem 4 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*

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

[33] The expression (14) 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 (14), the proof is given for completeness.

if either of the following conditions is satisfied:
(a) There exists an $\boldsymbol{z} \in \mathrm{ri}(\mathrm{dom}\, f)$ *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 $b \neq 0$. This implies that $\boldsymbol{y}^\top \boldsymbol{\lambda} = 0$ must holds in order to maximize $-f^*(-\boldsymbol{A}^\top \boldsymbol{\lambda})$. □

Corollary 5 (Strong duality for the Ivanov regularization case)**.** *The strong duality holds between (20) and (21), 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* $\|\boldsymbol{w}\| < 1$ *and* $-\boldsymbol{G}^\top \boldsymbol{w} + \boldsymbol{y} b \in \mathrm{ri}(\mathrm{dom}\, \mathcal{F})$.
(b) There exists a $\boldsymbol{\lambda} \in \mathrm{ri}(\mathrm{dom}\, \mathcal{F}^*)$ *such that* $\boldsymbol{y}^\top \boldsymbol{\lambda} = 0$,
where $\mathrm{ri}(\cdot)$ *denotes the interior points of a set. Under (a), the supremum in the dual is attained at some* $\boldsymbol{\lambda}$*, while under (b), the infimum is attained at some* $\boldsymbol{w}, b$*. In addition, if* $\mathcal{F}$ *(or equivalently,* $\mathcal{F}^*$*) is polyhedral, "*ri*" can be omitted.*

Proof. We reach the conclusion just by observing that with $f(\boldsymbol{z}, z_0) = \delta_{\{(\boldsymbol{u}, v) : \|\boldsymbol{u}\| \leq 1\}}(\boldsymbol{z}, z_0)$, $f^*(\boldsymbol{w}, b) = \|\boldsymbol{w}\|^\circ$ if $b = 0$; $+\infty$ if $b \neq 0$. (Note that the conjugate of an indicator function of a set is its support function.) □

With a positively homogeneous risk functional $\mathcal{F}$, (21) is rewritten by

$$d^\star := \sup_{\boldsymbol{\lambda}} \quad -\|\boldsymbol{G}\boldsymbol{\lambda}\|^\circ - \delta_{\mathcal{Q}_\mathcal{F}}(\boldsymbol{\lambda}) - \delta_{\boldsymbol{y}^\top(\cdot)=0}(\boldsymbol{\lambda}). \tag{53}$$

Correspondingly, the condition (b) of Corollary 5 can be replaced by
(b) There exists a $\boldsymbol{\lambda}$ such that $\boldsymbol{y}^\top \boldsymbol{\lambda} = 0$ and $\boldsymbol{\lambda} \in \mathrm{ri}\mathcal{Q}_\mathcal{F}$.

B.3 Proof of Theorem 5

Similarly to Theorem 4, in order to prove Theorem 5, 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:*

$$-\boldsymbol{A}^\top \boldsymbol{\lambda}^\star \in \partial f(\boldsymbol{z}^\star), \quad \boldsymbol{A}\boldsymbol{z}^\star \in \partial g^*(\boldsymbol{\lambda}^\star). \tag{54}$$

For (20) and (21), the KKT condition (54) can be explicitly written by

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

Note that

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

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

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

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 $\|\boldsymbol{u}\|^\circ = \boldsymbol{u}^\top \boldsymbol{w}^\star$. □

Example: Optimality condition for $(\mathrm{LSE}_{(t,\boldsymbol{p})}, \|\cdot\|)$**.** Let us consider the optimality condition (25) for $(\mathrm{LSE}_{(t,\boldsymbol{p})}, \|\cdot\|)$. Noting that at any $\boldsymbol{\lambda} \in \mathbb{\Pi}_+^m$, the function $\mathcal{F}^*(\boldsymbol{\lambda}) = \mathrm{KL}_{(t,\boldsymbol{p})}(\boldsymbol{\lambda}) = \frac{1}{t}\boldsymbol{\lambda}^\top \ln(\boldsymbol{\lambda}./\boldsymbol{p}) + \delta_{\mathbb{\Pi}_+^m}(\boldsymbol{\lambda})$ has subdifferential $\partial\mathcal{F}^*(\boldsymbol{\lambda}) = \{\nabla \frac{1}{t}\boldsymbol{\lambda}^\top \ln(\boldsymbol{\lambda}./\boldsymbol{p})(\boldsymbol{\lambda}) + k\mathbf{1}_m : k \in \mathbb{R}\}$, the optimality condition is then explicitly given by

$$(\boldsymbol{\lambda}^\star)^\top \boldsymbol{G}^\top \boldsymbol{w}^\star = \|\boldsymbol{G}\boldsymbol{\lambda}^\star\|^\circ, \quad \|\boldsymbol{w}^\star\| \le 1, \quad \boldsymbol{y}^\top \boldsymbol{\lambda}^\star = 0,$$
$$-\boldsymbol{G}^\top \boldsymbol{w}^\star + \boldsymbol{y} b^\star = \tfrac{1}{t}(\ln \boldsymbol{\lambda}^\star./\boldsymbol{p} + \mathbf{1}_m) + \mathbf{1}_m k^\star, \quad \mathbf{1}_m^\top \boldsymbol{\lambda}^\star = 1 \quad \boldsymbol{\lambda}^\star > \mathbf{0}.$$

Furthermore, consider the situation where ℓ_2-norm is employed in $(\mathrm{LSE}_{(t,\boldsymbol{p})}, \|\cdot\|_2)$, and there exists a solution $\boldsymbol{\lambda}^\star > \mathbf{0}$ 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[34]:

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

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

If $\mathcal{F}$ is positively homogeneous, the condition (25) of Theorem 5 can be rewritten as follows:

$$\boldsymbol{G}\boldsymbol{\lambda}^\star \in \mathcal{N}(\boldsymbol{w}^\star), \quad \|\boldsymbol{w}^\star\| \le 1, \quad \boldsymbol{y}^\top \boldsymbol{\lambda}^\star = 0, \quad -\boldsymbol{G}^\top \boldsymbol{w}^\star + \boldsymbol{y} b^\star \in N_{\mathcal{Q}_\mathcal{F}}(\boldsymbol{\lambda}^\star), \quad \boldsymbol{\lambda}^\star \in \mathcal{Q}_\mathcal{F}, \tag{55}$$

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.,* $N_{\mathcal{Q}_\mathcal{F}}(\boldsymbol{\lambda}^\star) := \{\boldsymbol{L} \in \mathbb{R}^m : \boldsymbol{L}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star) \le 0,$ *for all* $\boldsymbol{\lambda} \in \mathcal{Q}_\mathcal{F}\}$.

The change in the final part of (55) 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).

B.4 Proof of Theorem 8

Noting the relation $L/U \le 1 \le U/L$, we can observe that $\{\boldsymbol{w} : (L/U)\|\boldsymbol{w}\| \le 1\} \supset \{\boldsymbol{w} : \|\boldsymbol{w}\| \le 1\} \supset \{\boldsymbol{w} : (U/L)\|\boldsymbol{w}\| \le 1\}$, which indicates an inclusion relation of the feasible regions of the optimization problems. Therefore, $\phi_{L/U} \le \hat{\phi} \le \phi_{U/L}$.

From (37), we have $(1/U)\|\boldsymbol{w}\| \le \|\boldsymbol{w}\|_2 \le (1/L)\|\boldsymbol{w}\|$, and accordingly, $\{\boldsymbol{w} : (1/U)\|\boldsymbol{w}\| \le 1\} \supset \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \le 1\} \supset \{\boldsymbol{w} : (1/L)\|\boldsymbol{w}\| \le 1\}$, which indicates an inclusion relation of the feasible regions again. Accordingly, $\phi_{1/U} \le \phi^{\mathrm{E}} \le \phi_{1/L}$.

When $L = 1$, the above two relations become $\phi_{1/U} \le \hat{\phi} \le \phi_U$ and $\phi_{1/U} \le \phi^{\mathrm{E}} \le \phi_1$. Taking into account $\hat{\phi} = \phi_1$, we have $\phi_{1/U} \le \phi^{\mathrm{E}} \le \hat{\phi} \le \phi_U$. Likewise, when $U = 1$, combining the relations $\phi_L \le \hat{\phi} \le \phi_{1/L}$, $\phi_1 \le \phi^{\mathrm{E}} \le \phi_{1/L}$ and $\hat{\phi} = \phi_1$, we have $\phi_L \le \hat{\phi} \le \phi^{\mathrm{E}} \le \phi_{1/L}$.

Finally, let us consider the case where $\mathcal{F}$ is positively homogeneous. In such a case, observing that $\phi_{1/L} = L\hat{\phi}$ (and $\phi_{1/U} = U\hat{\phi}$), we have the worst-case upper bound $\phi^{\mathrm{E}} \le \phi_{1/L} = L\hat{\phi}$. From (37), we have $-U\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|^\circ = U\hat{\phi} \le -\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2$. Since $\hat{\boldsymbol{\lambda}}$ is feasible to the dual problem under any norm $\|\cdot\|^\circ$, we have $-\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2 \le \phi^{\mathrm{E}}$. Similarly, $(\hat{\boldsymbol{w}}, \hat{b})/\|\hat{\boldsymbol{w}}\|_2$ is feasible to the primal problem with $\|\boldsymbol{w}\|_2$, we have $\phi^{\mathrm{E}} \le \hat{\phi}/\|\hat{\boldsymbol{w}}\|_2$. Summing up all the relations, we have (38). □

[34] Note that we have $\lambda^\star = \boldsymbol{p}./\exp(-\mathbf{1}_m - tk^\star \mathbf{1}_m - t\boldsymbol{G}^\top \boldsymbol{w}^\star + \boldsymbol{y} b^\star)$ from the condition $-\boldsymbol{G}^\top \boldsymbol{w}^\star + \boldsymbol{y} b^\star = \frac{1}{t}(\ln \boldsymbol{\lambda}^\star./\boldsymbol{p} + \mathbf{1}_m) + \mathbf{1}_m k^\star$. Therefore, substituting this to the other equalities, the system is reduced to $n + 2$ equalities.

B.5 Proof of Theorem 9

Let $\boldsymbol{u} = \boldsymbol{\Gamma}\boldsymbol{\lambda}$. From the optimality condition (25), if $\boldsymbol{u} \neq \mathbf{0}$, then $\|\boldsymbol{w}^\star\| = 1$, and $\|\boldsymbol{u}\|^\circ = \|\boldsymbol{u}\|^\circ\|\boldsymbol{w}\| = \boldsymbol{u}^\top \boldsymbol{w}^\star \leq \|\boldsymbol{u}\|_2\|\boldsymbol{w}\|_2$. Therefore, $1 \geq \frac{\boldsymbol{u}^\top \boldsymbol{w}^\star}{\|\boldsymbol{u}\|_2\|\boldsymbol{w}^\star\|_2} = \frac{\|\boldsymbol{u}\|^\circ\|\boldsymbol{w}^\star\|}{\|\boldsymbol{u}\|_2\|\boldsymbol{w}^\star\|_2}$. From the condition (37), it is valid that $\|\boldsymbol{u}\|^\circ/\|\boldsymbol{u}\|_2 \geq (1/U)$ and $\|\boldsymbol{w}^\star\|/\|\boldsymbol{w}^\star\|_2 \geq L$, and accordingly, the ratio on the right-hand side above is bounded below L/U. □

C LP and symbolical formulations of ν-SVM with $\langle\!\langle\cdot\rangle\!\rangle_\alpha^\circ$, $(\!(\cdot)\!)_\tau$ and $(\!(\cdot)\!)_\tau^\circ$

This section provides three examples of the ν-SVM with the parametrized families of polyhedral norms, $\langle\!\langle\cdot\rangle\!\rangle_\alpha^\circ$, $(\!(\cdot)\!)_\tau$ and $(\!(\cdot)\!)_\tau^\circ$, and a PSG code example.

Example: $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha^\circ)$**.** The primal and dual pair of the SVM with the dual CVaR norm regularizer are given by, respectively,

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \langle\!\langle\boldsymbol{w}\rangle\!\rangle_\alpha^\circ \leq 1, \end{array}\right| \left|\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\langle\!\langle\boldsymbol{G}\boldsymbol{\lambda}\rangle\!\rangle_\alpha \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{\lambda} \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}, \end{array}\right.$$

which are presented by LPs

$$\left|\begin{array}{lrrrrll} \underset{\boldsymbol{w}^+,\boldsymbol{w}^-,b,\rho,\boldsymbol{z}}{\text{minimize}} & & & & -\rho + \frac{1}{m\nu}\mathbf{1}_m^\top\boldsymbol{z} & & \downarrow \text{dual var.} \downarrow \\ \text{subject to} & -\boldsymbol{G}^\top\boldsymbol{w}^+ & +\boldsymbol{G}^\top\boldsymbol{w}^- & +\boldsymbol{y}b + \mathbf{1}_m\rho & -\boldsymbol{z} \leq \mathbf{0}, & & \leftarrow \boldsymbol{\lambda} \geq \mathbf{0} \\ & \boldsymbol{w}^+ & +\boldsymbol{w}^- & & \leq \mathbf{1}_n, & & \leftarrow \boldsymbol{\mu} \geq \mathbf{0} \\ & \mathbf{1}_n^\top\boldsymbol{w}^+ & +\mathbf{1}_n^\top\boldsymbol{w}^- & & \leq n(1-\alpha), & & \leftarrow \theta \geq 0 \\ & (\boldsymbol{w}^+, \boldsymbol{w}^-, \boldsymbol{z}) \geq \mathbf{0}, & (b, \rho: \text{ free variable}) & & & & . \end{array}\right. \tag{56}$$

and

$$\left|\begin{array}{lrrrll} \underset{\boldsymbol{\lambda},\boldsymbol{\mu},\theta}{\text{maximize}} & & -\mathbf{1}_n^\top\boldsymbol{\mu} & -n(1-\alpha)\theta & & \downarrow \text{dual var.} \downarrow \\ \text{subject to} & \boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & -\mathbf{1}_n\theta \leq \mathbf{0}, & & \leftarrow \boldsymbol{w}^+ \geq \mathbf{0} \\ & -\boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & -\mathbf{1}_n\theta \leq \mathbf{0}, & & \leftarrow \boldsymbol{w}^- \geq \mathbf{0} \\ & -\boldsymbol{y}^\top\boldsymbol{\lambda} & & = 0, & & \leftarrow b \\ & -\mathbf{1}_m^\top\boldsymbol{\lambda} & & = -1, & & \leftarrow \rho \\ & \boldsymbol{\lambda} & & \leq \mathbf{1}_m/(m\nu) & & \leftarrow \boldsymbol{z} \geq \mathbf{0} \\ & (\boldsymbol{\lambda}, \boldsymbol{\mu}, \theta) \geq \mathbf{0}, & & & & \end{array}\right. \tag{57}$$

respectively. As in the case of $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha)$, for a primal optimal solution $((\boldsymbol{w}^+)^\star, (\boldsymbol{w}^-)^\star, b^\star, \rho^\star, \boldsymbol{z}^\star)$, which can be obtained via the dual formulation as mentioned before, we can obtain the decision function (2).

Example: $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, (\!(\cdot)\!)_\tau)$**.** The primal and dual pair of the ν-SVM with the deltoidal norm regularization can be represented by, respectively,

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & (\!(\boldsymbol{w})\!)_\tau \leq 1, \end{array}\right| \left|\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -(\!(\boldsymbol{G}\boldsymbol{\lambda})\!)_\tau^\circ \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{\lambda} \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}, \end{array}\right.$$

which are presented by LPs

$$\left|\begin{array}{lrrrrrrrl}
\underset{\boldsymbol{w}^+,\boldsymbol{w}^-,b,\rho,\boldsymbol{z},s}{\text{minimize}} & & & & & -\rho+\frac{1}{m\nu}\mathbf{1}_m^\top\boldsymbol{z} & & & \downarrow \text{dual var.} \downarrow \\
\text{subject to} & -\boldsymbol{G}^\top\boldsymbol{w}^+ & +\boldsymbol{G}^\top\boldsymbol{w}^- & +\boldsymbol{y}b & +\mathbf{1}_m\rho & -\boldsymbol{z} & \le \mathbf{0}, & \leftarrow \boldsymbol{\lambda}\ge\mathbf{0} \\
& \boldsymbol{w}^+ & +\boldsymbol{w}^- & & & & -\mathbf{1}_n s\le\mathbf{0}, & \leftarrow \boldsymbol{\mu}\ge\mathbf{0} \\
& (1-\tau)\mathbf{1}_n^\top\boldsymbol{w}^+ & +(1-\tau)\mathbf{1}_n^\top\boldsymbol{w}^+ & & & & +\tau s\le 1, & \leftarrow \theta\ge 0 \\
& (\boldsymbol{w}^+,\boldsymbol{w}^-,\boldsymbol{z})\ge\mathbf{0},\ (b,\rho,s:\text{ free variable}) & & & & & . &
\end{array}\right. \tag{58}$$

and

$$\left|\begin{array}{lrrrrl}
\underset{\boldsymbol{\lambda},\boldsymbol{\mu},\theta}{\text{maximize}} & & & -\theta & & \downarrow \text{dual var.} \downarrow \\
\text{subject to} & \boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & -(1-\tau)\mathbf{1}_m\theta\le\mathbf{0}, & & \leftarrow \boldsymbol{w}^+\ge\mathbf{0} \\
& -\boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & -(1-\tau)\mathbf{1}_m\theta\le\mathbf{0}, & & \leftarrow \boldsymbol{w}^-\ge\mathbf{0} \\
& -\boldsymbol{y}^\top\boldsymbol{\lambda} & & =0, & & \leftarrow b \\
& -\mathbf{1}_m^\top\boldsymbol{\lambda} & & =-1, & & \leftarrow \rho \\
& \boldsymbol{\lambda} & & \le\mathbf{1}_m/(m\nu) & & \leftarrow \boldsymbol{z}\ge\mathbf{0}, \\
& & \mathbf{1}_n^\top\boldsymbol{\mu} & -\tau\theta=0, & & \leftarrow c \\
& (\boldsymbol{\lambda},\boldsymbol{\mu},\theta)\ge\mathbf{0}, & & & &
\end{array}\right. \tag{59}$$

respectively. As in the case of the CVaR norms, that is, $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha)$ and $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha^\circ)$, for a primal optimal solution $((\boldsymbol{w}^+)^\star,(\boldsymbol{w}^-)^\star,b^\star,\rho^\star,\boldsymbol{z}^\star,s^\star)$, which can be obtained via the dual formulation as mentioned before, we can obtain the decision function (2). As the proximity theory tells, with $\tau=\tau^\star$, the solution is more alike the ℓ_2-solution than the CVaR case in the sense of Theorems 8 and 9.

Example: $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, (\!(\cdot)\!)_\tau^\circ)$**.** Likewise, the ν-SVM with the dual deltoidal norm regularization can be written by, respectively,

$$\left|\begin{array}{ll}
\underset{\boldsymbol{w},b}{\text{minimize}} & \mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{G}^\top\boldsymbol{w}+\boldsymbol{y}b) \\
\text{subject to} & (\!(\boldsymbol{w})\!)_\tau^\circ\le 1,
\end{array}\right.
\left|\begin{array}{ll}
\underset{\boldsymbol{\lambda}}{\text{maximize}} & -(\!(\boldsymbol{G}\boldsymbol{\lambda})\!)_\tau \\
\text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda}=0,\ \boldsymbol{\lambda}\in\mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)},
\end{array}\right.$$

which are presented by LPs

$$\left|\begin{array}{lrrrrrrrl}
\underset{\boldsymbol{w}^+,\boldsymbol{w}^-,b,\rho,\boldsymbol{z},\boldsymbol{\zeta}}{\text{minimize}} & & & & & -\rho+\frac{1}{m\nu}\mathbf{1}_m^\top\boldsymbol{z} & & & \downarrow \text{dual var.} \downarrow \\
\text{subject to} & -\boldsymbol{G}^\top\boldsymbol{w}^+ & +\boldsymbol{G}^\top\boldsymbol{w}^- & +\boldsymbol{y}b & +\mathbf{1}_m\rho & -\boldsymbol{z} & \le\mathbf{0}, & \leftarrow \boldsymbol{\lambda}\ge\mathbf{0} \\
& \boldsymbol{w}^+ & +\boldsymbol{w}^- & & & & -\{(1-\tau)\boldsymbol{E}_n+\tau\boldsymbol{I}_n\}\boldsymbol{\zeta}\le\mathbf{0}, & \leftarrow \boldsymbol{\mu}\ge\mathbf{0} \\
& & & & & & \mathbf{1}_n^\top\boldsymbol{\zeta}\le 1, & \leftarrow \theta\ge 0 \\
& (\boldsymbol{w}^+,\boldsymbol{w}^-,\boldsymbol{z},\boldsymbol{\zeta})\ge\mathbf{0},\ (b,\rho:\text{ free variable}) & & & & & & &
\end{array}\right. . \tag{60}$$

and

$$\left|\begin{array}{lrrrrl}
\underset{\boldsymbol{\lambda},\boldsymbol{\mu},\theta}{\text{maximize}} & & & -\theta & & \downarrow \text{dual var.} \downarrow \\
\text{subject to} & \boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & \le\mathbf{0}, & & \leftarrow \boldsymbol{w}^+\ge\mathbf{0} \\
& -\boldsymbol{G}\boldsymbol{\lambda} & -\boldsymbol{\mu} & \le\mathbf{0}, & & \leftarrow \boldsymbol{w}^-\ge\mathbf{0} \\
& -\boldsymbol{y}^\top\boldsymbol{\lambda} & & =0, & & \leftarrow b \\
& -\mathbf{1}_m^\top\boldsymbol{\lambda} & & =-1, & & \leftarrow \rho \\
& \boldsymbol{\lambda} & & \le\mathbf{1}_m/(m\nu) & & \leftarrow \boldsymbol{z}\ge\mathbf{0}, \\
& & \{(1-\tau)\boldsymbol{E}_n+\tau\boldsymbol{I}_n\}\boldsymbol{\mu} & -\mathbf{1}_n\theta\le\mathbf{0}, & & \leftarrow \boldsymbol{\zeta}\ge\mathbf{0} \\
& (\boldsymbol{\lambda},\boldsymbol{\mu},\theta)\ge\mathbf{0}, & & & &
\end{array}\right. \tag{61}$$

respectively. Similarly to the three aforementioned examples above, for a primal optimal solution $((\boldsymbol{w}^+)^\star, (\boldsymbol{w}^-)^\star, b^\star, \rho^\star, \boldsymbol{z}^\star, \boldsymbol{\zeta})$, which can be obtained via the dual formulation, we can obtain the decision function (2). As in the case of $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, (\!(\cdot)\!)_\tau)$, with $\tau = \tau^\star$, the solution is more alike the ℓ_2-solution than the CVaR case.

Compatibility with optimization software packages. For last decade, solver softwares having user-friendly interfaces are increasingly available. For example, PSG (American Optimal Decisions, Inc., 2009) and CVX (Grant and Boyd, 2012) have the ℓ_2-, ℓ_1-, ℓ_∞-norms, and CVaR norm as built-in functions. In particular, the former builds in several risk functional listed in this paper. For example, CVaR is coded in PSG just by its name and the matrix name containing data for the CVaR function. Let us consider ℓ_∞-norm-regularized ν-SVM, i.e., $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \|\cdot\|_\infty)$, whose pair of the primal and dual is formulated in a symbolic manner as

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{G}^\top\boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{w}\|_\infty \le 1; \end{array}\right| \left|\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{G}\boldsymbol{\lambda}\|_1 \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{\lambda} \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}. \end{array}\right.$$

By using the built-in CVaR and ℓ_∞-norm functions, the PSG script for implementing $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \|\cdot\|_\infty)$, is written (in simplified form) as follows:

```
minimize                                  | maximize
  cvar_risk (0.42,matrix_YX_b)            |   -polynom_abs (matrix_polynom_abs_q)
Constraint: <= 1                          | Constraint: == 0
  max_comp_abs (matrix_max_comp_abs)      |   linearmulti(matrix_YXT_Lam_q)
                                          | Constraint: == 0
                                          |   linear(matrix_yi)
                                          | Constraint: == 1
                                          |   linear(matrix_unit)
                                          | Box: >= 0, <= point_upperbounds_0p42
```

where

- `cvar_risk (0.42, matrix_YX_b)` = CVaR with confidence level $1-\nu = 0.42$ defined on the matrix `matrix_YX_b` containing `data (Y X,b)`;
- `max_comp_abs (matrix_max_comp_abs)` = ℓ_∞-norm for vector $\boldsymbol{w}$ defined by matrix `matrix_max_comp_abs`;
- `polynom_abs (matrix_polynom_abs_q)` = ℓ_1-norm for vector $\boldsymbol{q}$ defined by matrix `matrix_max_comp_abs_q`;
- `linearmulti(matrix_YXT_Lam_q)` = system of equations $\boldsymbol{q} = \boldsymbol{G}\boldsymbol{\lambda}$ defined by matrix `matrix_YXT_Lam_q` `linear(matrix_yi)` = linear equation $\boldsymbol{y}^\top\boldsymbol{\lambda} = 0$ defined by matrix `matrix_yi`;
- `linear(matrix_unit)` = linear equation summing all components of vector $\boldsymbol{\lambda}$ to 1, which is included in the subset $\mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}$,
- `>= 0, <= point_upperbounds_0p42` define box for variables for vector $\boldsymbol{\lambda}$ in the subset $\mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}$.

PSG implementation for these two problems with codes, data, and solutions are posted as Problems 4a and 4b at this link.[35]

D Kernelization

One powerful feature of SVM is its extension to nonlinear classification based on the so-called *kernel trick*. This section explains how the kernel trick can be applied to the generalized formulations and give some remarks on the some incongruence when a non-ℓ_2-norm is employed for the kernelization.

Precisely, we consider the following three ways of incorporating the kernel.

1. The first option is to solve a kernelized dual which is obtained by simply replacing the matrix $\boldsymbol{G}$ with another matrix $\boldsymbol{\Gamma}$, and to apply the solution to the kernelized decision function.

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

2. The second option is to solve the primal where the parallelism $\boldsymbol{w} = \eta\boldsymbol{\Gamma\lambda}$ is embedded, and to apply the solution to the primal decision function (5).

3. The third option for kernelization is to replace the loss (16) with (17), which is defined with a kernel function, and to apply the decision function as if it is a linear classification.

In the following, we see pros and cons of the above three strategies for kernelizing the general formulations. For a benchmarking purpose, let us begin with the ℓ_2-case.

D.1 Kernel trick with ℓ_2-norm

One easy way for obtaining a nonlinear classification is just to replace the attribute vectors $\boldsymbol{X} := (\boldsymbol{x}_1, ..., \boldsymbol{x}_m)^\top \in \mathbb{R}^{m\times n}$ by mapped vectors $\phi(\boldsymbol{X}) := (\phi(\boldsymbol{x}_1), ..., \phi(\boldsymbol{x}_m))^\top \in \mathbb{R}^{m\times N}$, where ϕ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^N$, and to redefine the normal vector $\boldsymbol{w}$ as a vector in $\mathbb{R}^N$ to fit the corresponding mapped attributes. The primal formulation (22) is then modified by just replacing $\boldsymbol{X}$ with $\phi(\boldsymbol{X})$:

$$p^\star := \left| \begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{Y}(\phi(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{1}_m b)) \\ \text{subject to} & \|\boldsymbol{w}\| \le 1. \end{array} \right. \tag{62}$$

With an optimal solution $(\boldsymbol{w}^\star, b^\star)$, the decision function (5) is revised as $d(\boldsymbol{x}) = \mathrm{sign}(\phi(\boldsymbol{x})^\top \boldsymbol{w}^\star - b^\star)$.

When the ℓ_2-norm is employed for the regularizer, we can modify the dual formulation (23), corresponding to the above (cosmetic) change in the primal formulation (22). Indeed, by defining the kernel function by $k(\boldsymbol{x}, \boldsymbol{z}) := \phi(\boldsymbol{x})^\top \phi(\boldsymbol{z})$ and denoting $\boldsymbol{K} := \boldsymbol{Y}\phi(\boldsymbol{X})\phi(\boldsymbol{X})^\top \boldsymbol{Y}$ (i.e., $K_{ij} = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$), the objective function of the dual problem (23) can be rewritten by $-\sqrt{\boldsymbol{\lambda}^\top \boldsymbol{K}\boldsymbol{\lambda}} - \mathcal{F}^*(\boldsymbol{\lambda})$. Noting that $\boldsymbol{K}$ is a symmetric positive semidefinite matrix, consider the following decomposition $\boldsymbol{K} = \boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}$. Typically, $\boldsymbol{\Gamma}$ can be given by, for example, $\boldsymbol{\Gamma} := \mathrm{diag}(\sqrt{\kappa_1}, ..., \sqrt{\kappa_m})[\boldsymbol{\iota}_1, ..., \boldsymbol{\iota}_m]^\top$ with κ_i (nonnegative) eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{\iota}_i$ their corresponding orthonormal eigenvectors, $i = 1, ..., m$. With a decomposition $\boldsymbol{\Gamma}$, the objective is further represented by $-\|\boldsymbol{\Gamma\lambda}\|_2 - \mathcal{F}^*(\boldsymbol{\lambda})$ since $\|\boldsymbol{\Gamma\lambda}\|_2 = \sqrt{\boldsymbol{\lambda}^\top\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma\lambda}} = \sqrt{\boldsymbol{\lambda}^\top\boldsymbol{K\lambda}}$. Namely, by rewriting the objective function of the dual problem (23), we have a kernelized version of dual formulation:

$$d^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{\Gamma\lambda}\|_2 - \mathcal{F}^*(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0. \end{array} \right. \tag{63}$$

Since the formulations (62) and (63) remain unchanged from the formulations (22) and (23) except the cosmetic changes, all the theoretical results implied by linear classification are valid. For example, the optimality condition (25) implies that if the optimal solution $\boldsymbol{\lambda}^\star$ to the dual formulation (63) satisfies $\boldsymbol{\Gamma\lambda}^\star \neq \boldsymbol{0}$, the optimal solution $\boldsymbol{w}^\star$ to the corresponding primal formulation (62) with $\|\cdot\| = \|\cdot\|_2$ is given by

$$\boldsymbol{w}^\star = \frac{1}{\|\boldsymbol{\Gamma\lambda}^\star\|_2}\boldsymbol{\Gamma\lambda}^\star. \tag{64}$$

The decision function is then given with the kernel function $k(\cdot,\cdot)$ as follows.

$$d(\boldsymbol{x}) := \mathrm{sign}\left(\frac{1}{\|\boldsymbol{\Gamma\lambda}^\star\|_2}\sum_{i=1}^m \lambda_i^\star y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b^\star\right), \tag{65}$$

where $b^\star$ is defined so as to satisfy the fourth condition of the optimality condition (25).

D.2 Strategy 1: Simple replacement of ℓ_2-norm in dual and decision function

Next let us consider the kernelization for the case of non-ℓ_2-norms, i.e., $\|\cdot\| \neq \|\cdot\|_2$.

Replacing the ℓ_2-norm with $\|\cdot\|^\circ$ in (63), we have a kernelized version of the generalized formulation (23):

$$d^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{\Gamma}\boldsymbol{\lambda}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0. \end{array} \right. \tag{66}$$

The primal formulation corresponding to (66) is given as follows.

$$p^\star := \left| \begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{\Gamma}^\top \boldsymbol{w} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{w}\| \leq 1. \end{array} \right. \tag{67}$$

Using the approximation of the form $\|\boldsymbol{x}\|_2 \approx \|\boldsymbol{x}\|$, we can achieve a kernelized decision function via (66) in a similar manner to the ℓ_2-norm case. Indeed, with an optimal solution $\boldsymbol{\lambda}^\star$ to (66), an analogy to the ℓ_2-case brings the formula

$$\boldsymbol{w}^\star = \frac{1}{\|\boldsymbol{\Gamma}\boldsymbol{\lambda}^\star\|}\boldsymbol{\Gamma}\boldsymbol{\lambda}^\star, \tag{68}$$

and the decision function can be defined by

$$d(\boldsymbol{x}) := \mathrm{sign}\left(\frac{1}{\|\boldsymbol{\Gamma}\boldsymbol{\lambda}^\star\|}\sum_{i=1}^m \lambda_i^\star y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b^\star\right), \tag{69}$$

where $b^\star$ is a constant to be determined.[36]

We would like to again emphasize that the justification of this approximation highly depends on the parallelism (68). It is noteworthy that the approximation of the kernel trick depends on the formula (64) which can be derived only with the ℓ_2-norm.

Let us consider some justification of the use of the parallelism. Note that the size of the decision vector $\boldsymbol{w}$ in (67) is m. Also, note that, however, all the theoretical results including the duality theorems (e.g., Theorem 5), the optimality conditions (e.g., Theorem 5) and the proximity values (Theorems 8 and 9) remain to hold only with simple cosmetic changes in the matrix, i.e., $\boldsymbol{G}$ to $\boldsymbol{\Gamma}$, and in the size of the vector $\boldsymbol{w}$, i.e., n to m.

Corollary 6 (Proximity for parallelism between $\boldsymbol{w}^\star$ and $\boldsymbol{\Gamma}\boldsymbol{\lambda}^\star$)**.** *Suppose that there are two positive constants L and U satisfying the condition (37). Let $(\boldsymbol{w}^\star, b^\star)$ and $\boldsymbol{\lambda}^\star$ be optimal solutions to the primal and dual formulations, respectively. Suppose that the solutions satisfy the strong duality and $\boldsymbol{\Gamma}\boldsymbol{\lambda}^\star \neq \boldsymbol{0}$. Then we have*

$$\frac{L}{U} \leq \frac{(\boldsymbol{w}^\star)^\top \boldsymbol{\Gamma}\boldsymbol{\lambda}^\star}{\|\boldsymbol{\Gamma}\boldsymbol{\lambda}^\star\|_2 \|\boldsymbol{w}^\star\|_2} \leq 1. \tag{70}$$

As mentioned in Section 6.1, this states that the lower (or higher) the proximity U/L (or L/U, respectively) is, the more the parallelism (68) fits. This implies that with a norm with high ratio L/U, we can expect that the kernelized decision function works. At the same time, we should remember that the ratio U/L (or L/U) depends on the dimension. In this sense, the use of $\langle\!\langle\cdot\rangle\!\rangle_{\alpha^\star}$ is preferred to $\|\cdot\|_1$. However, the ratio in (70) can be very large since it depends on m, i.e., the number of data samples. Even with $\langle\!\langle\cdot\rangle\!\rangle_{\alpha^\star}$, the ratio can be far below 1 when m is large. Therefore, the simplistic application of the kernelization under a non-ℓ_2-norm needs more discussion.

Remark 2. *To mitigate the incongruence, we can estimate the threshold $b^\star$ on the basis of the decision function (69). Namely, after obtaining an optimal solution $\boldsymbol{\lambda}^\star$ to the dual (66), we can determine a $b^\star$ so that the in-sample error of the decision function is minimized. However, we should note that the logical incongruence would not fully disappear with this manipulation.*

[36] Term $b^\star$ can be attained on the basis of the optimality condition (25). Indeed, as mentioned in Sections 6.2 and 6.3, if an LP-representable norm is employed, any state-of-the-art solver yields $b^\star$ as a Lagrange multiplier. However, this substitution is justified only via an approximation of the optimality condition of the ℓ_2-case.

D.3 Strategy 2: Imposing the parallelism constraints to the primal formulation

Next we consider another kernelized formulation motivated by the formulations in the literature (e.g., Mangasarian, 2000; Zhou, Zhang, and Jiao, 2002; Chapelle, 2007) and describe similarity and difference between them and (66). As overviewed above, application of the kernelized decision (69) requires the parallelism (68). Therefore, by substituting the formula $\boldsymbol{w} = \boldsymbol{\Gamma}\boldsymbol{v}$ into the primal formulation (22), we can obtain a kernelized version of the primal formulation:

$$p^{\boldsymbol{K}} := \left| \begin{array}{ll} \underset{\boldsymbol{v},b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{K}\boldsymbol{v} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{\Gamma}\boldsymbol{v}\| \leq 1, \end{array} \right. \tag{71}$$

where $\boldsymbol{K} = \boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}$. For example, the formulation by Mangasarian (2000) corresponds to the case where $\mathcal{F} = \text{Hinge1}(t, \mathbf{1}_m/m)$ and $\|\cdot\| = \|\cdot\|_1$ are employed although the regularization term appears in the objective therein. Comparing (22) and (71), we see that $p^{\boldsymbol{K}} \geq p^\star = d^\star$ when $\boldsymbol{\Gamma} = \boldsymbol{G}$ and a non-ℓ_2-norm are employed. With an optimal solution $(\boldsymbol{v}^\star, b^\star)$, the decision function can be defined by

$$d(\boldsymbol{x}) := \text{sign}(\sum_{i=1}^m v_i^\star y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b^\star). \tag{72}$$

Let us emphasize the difference of the two kernelization approaches, i.e., (72) via (71) and (69) via (66). As for the former approach, the approximation of the parallelism is carried out in building the formulation (71), not in defining (72). On the other hand, the formulation (66) remains (substantially) the same as (23), while the approximation of the parallelism is carried out when an optimal solution $\boldsymbol{\lambda}^\star$ is substituted into the formula to obtain the decision function (69).

To make the difference clearer, let us introduce the dual formulation to (71) as follows:

$$d^{\boldsymbol{K}} := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{\xi}}{\text{maximize}} & -\|\boldsymbol{\Gamma}\boldsymbol{\lambda} - \boldsymbol{\xi}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{\Gamma}^\top\boldsymbol{\xi} = \boldsymbol{0}. \end{array} \right. \tag{73}$$

Theorem 10 (Strong duality for kernelized formulations)**.** *The strong duality holds between (71) and (73), i.e., we have $p^{\boldsymbol{K}} = d^{\boldsymbol{K}}$, if either of the following conditions is satisfied:*

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

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

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

Proof. Similarly to the proof of Theorem 5, we just need to confirm the correspondence of the notation between it and Theorem 31.3 of Rockafellar (1970) and to simplify the results. Also, note the relation $\sup_{\boldsymbol{w},\boldsymbol{\mu}}\{-\boldsymbol{\zeta}^\top\boldsymbol{\Gamma}^\top\boldsymbol{w} : \boldsymbol{\Gamma}\boldsymbol{\mu} = \boldsymbol{w}, \|\boldsymbol{w}\| \leq 1\} = \inf_{\boldsymbol{w}}\{\|\boldsymbol{\Gamma}\boldsymbol{\zeta} - \boldsymbol{w}\|^\circ : \boldsymbol{\Gamma}^\top\boldsymbol{w} = \boldsymbol{0}\}$. □

Since $(\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\boldsymbol{\lambda}^\star, \boldsymbol{0})$ is feasible to (66) with any optimal $\boldsymbol{\lambda}^\star$, we see that $d^\star \leq d^{\boldsymbol{K}}$ when $\boldsymbol{\Gamma} = \boldsymbol{G}$. Conversely, the difference arising from $\boldsymbol{\xi}$ can be considered as that from the approximation of the parallelism. In contrast to the approach via (23), the kernelization via (71) or their dual (73) is applicable also to non-ℓ_2-norm without a logical jump. However, we have to cope with larger optimization problem due to the added constraints, which can be considered as a price of the use of non-ℓ_2-norms for kernelization.

The following fact shows that two kernelized formulations coincide when the ℓ_2-norm is employed.

Proposition 6 (Equivalence between (66) and (73) with ℓ_2-norm)**.** *If the ℓ_2-norm is employed, then any optimal solution $(\boldsymbol{\lambda}^\star, \boldsymbol{\xi}^\star)$ to (73) attains $\boldsymbol{\xi}^\star = \boldsymbol{0}$. Namely, if $\boldsymbol{\Gamma} = \boldsymbol{G}$ is further employed, then (66) and (73) are equivalent.*

Proof. Note that regarding the variables $\boldsymbol{\xi}$, the objective of (73) is equivalent to minimize $\|\boldsymbol{\Gamma\lambda}-\boldsymbol{\xi}\|_2^2 = \|\boldsymbol{\Gamma\lambda}\|_2^2 - 2\boldsymbol{\lambda}^\top\boldsymbol{\Gamma}^\top\boldsymbol{\xi} + \|\boldsymbol{\xi}\|_2^2$. Since an optimal solution $\boldsymbol{\xi}^\star$ satisfies $\boldsymbol{\Gamma}^\top\boldsymbol{\xi}^\star = \mathbf{0}$, this function results in $\|\boldsymbol{\Gamma\lambda}^\star - \boldsymbol{\xi}^\star\|_2^2 = \|\boldsymbol{\Gamma\lambda}^\star\|_2^2 + \|\boldsymbol{\xi}^\star\|_2^2$, and accordingly, $\boldsymbol{\xi}^\star = \mathbf{0}$ must hold. □

Namely, when the ℓ_2-norm is employed, we can get rid of the additional constraints and the decision variables $\boldsymbol{\xi}$ from (73). This fact implies that the ℓ_2-norm has a computational advantage over non-ℓ_2-norms when m is very large.

Reduction with low-rank decomposition. Let r be the integer such that $r \leq m$, and let us denote by $\boldsymbol{\Gamma}_{[r]} \in \mathbb{R}^{r\times m}$ a low-rank decomposition of the kernel matrix $\boldsymbol{K} \in \mathbb{R}^{m\times m}$, i.e., $\boldsymbol{\Gamma}_{[r]}^\top\boldsymbol{\Gamma}_{[r]} \approx \boldsymbol{K}$. An example of $\boldsymbol{\Gamma}_{[r]}$ can be obtained by the eigenvalue decomposition, i.e., with the eigenvalues, $\kappa_1 \geq \cdots \geq \kappa_r \geq \cdots \geq \kappa_m$, of $\boldsymbol{K}$, a truncated version of $\boldsymbol{\Gamma}$ can be defined by $\boldsymbol{\Gamma}_{[r]} := \mathrm{diag}(\sqrt{\kappa_1}, ..., \sqrt{\kappa_r})[\boldsymbol{\iota}_1, ..., \boldsymbol{\iota}_r]^\top$, where $\boldsymbol{\iota}_i$ denotes the orthonormal eigenvector corresponding to κ_i.[37]

Note that the number of elements of this truncated matrix can be reduced proportionally to r. Besides, in the sense of the principal component analysis (PCA), it preserves the information contained in the largest r eigenvalues and vector of the original matrix $\boldsymbol{K}$.

Consequently, we can reduce the kernelized formulation (66) to

$$d_{[r]} := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda},\boldsymbol{u}}{\text{maximize}} & -\|\boldsymbol{u}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0,\ \boldsymbol{u} = \boldsymbol{\Gamma}_{[r]}\boldsymbol{\lambda},\ \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{u} \in \mathbb{R}^r. \end{array} \right. \tag{74}$$

Note that the number of constraints and variables can be significantly reduced. For example, when $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha, \boldsymbol{\Gamma})$ with $(n, m) = (50, 1000)$ is reduced to $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \langle\!\langle\cdot\rangle\!\rangle_\alpha, \boldsymbol{\Gamma}_{[r]})$ with $r = 5$, the number of equality and inequality constraints except the box-type constraints is reduced from $3m + 1 = 3001$ to $m + 2r + 1 = 1101$, while the number of decision variables is from $5m + 3 = 5003$ to $2m + 3r + 3 = 2018$.

It is noteworthy that the matrix $\boldsymbol{G} = \boldsymbol{X}^\top\boldsymbol{Y}$ can also be a rank-n decomposition of $\boldsymbol{K}$ (i.e., $r = n$). Indeed, this decomposition (74) can also be applied to the usual QP formulation such as (3) and (7), and shorten the computation time.

Figure 2 shows the average elapsed time for solving the two types of QP formulations: i) Decomposed formulation; ii) Standard formulation (7) with $\boldsymbol{\Gamma}$ replacing $\boldsymbol{X}^\top\boldsymbol{Y}$ of $(\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}, \|\cdot\|_2)$ with $\boldsymbol{\Gamma}_{[r]}$. using WDBC data set. WDBC data set and the RBF kernel is used for setting $\boldsymbol{\Gamma}_{[r]}$.

Improvement of proximity. Besides, the decomposition with small r reduces the incongruence with the parallelism between $\boldsymbol{w}$ and $\boldsymbol{\Gamma}_{[r]}\boldsymbol{\lambda}$. Indeed, when r is small, the ratio L/U can be close to 1.

Figure 3 reports the ratios which indicate the proximity to the ℓ_2-case when the ν-SVMs with six polyhedral norms are applied to the WDBC breast cancer data set by employing the rank-r decomposition with the eigenvalue decomposition of the RBF kernel with $\sigma = 100$ and $r = 569, 100, 50, 30, 20, 10, 5, 3$.

All the proximity values are improved as r decreases, as the theory predicts, which means the difference between the ℓ_2-norm case and any LP-representable norm become smaller under the reduction. In this sense, the reduction also reduces the incongruence associated with the approximation of the representer theorem by the non-ℓ_2-norms.

D.4 Strategy 3: Change the definition of loss

The third way for kernelization is to simply change the definition of the loss from (16) to (17), together with the change of primal variables from $\boldsymbol{w} \in \mathbb{R}^n$ to $\boldsymbol{\alpha} \in \mathbb{R}^m$. Namely, the primal and dual problems are given by

$$p^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\alpha},b}{\text{minimize}} & \mathcal{F}(-\boldsymbol{K\alpha} + \boldsymbol{y}b) \\ \text{subject to} & \|\boldsymbol{\alpha}\| \leq 1; \end{array} \right. \quad d^\star := \left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{K\lambda}\|^\circ - \mathcal{F}^*(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{y}^\top\boldsymbol{\lambda} = 0, \end{array} \right.$$

[37] Although making a truncated matrix itself may cause a large computational burden, we can mitigate it by applying any efficient algorithm for row rank matrix decomposition that has been proposed (e.g., Bach and Jordan, 2005; Fine and Scheinberg, 2001).

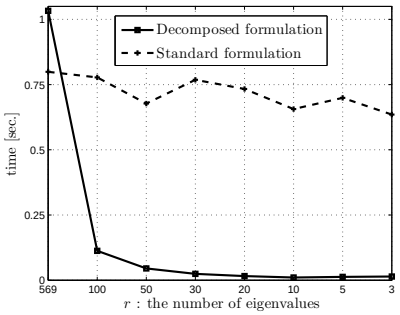


Figure 2: Average elapsed time for WDBC data set with RBF kernel

The average elapsed time in seconds for computing kernelized classifiers for WDBC data set via two formulations of quadratic program (QP): i) Decomposed formulation; ii) Standard formulation. Average was computed over 15 values of ν among $0.05, 0.10, ..., 0.70$. RBF kernel is used with parameter $\sigma = 100$. Decomposed formulation has r additional linear equality constraints and a diagonal matrix of size r ($\leq m$) as the Hessian for the quadratic term, whereas Standard formulation as no additional constraints and $m \times m$ dense Hessian, where m is the number of samples, i.e., 569. See the footnote of Table 7 for the computational environment.

where $\boldsymbol{K} \in \mathbb{R}^{m \times m}$. Note that the $\boldsymbol{K} = \boldsymbol{K}^\top$.

Looking at the dual, we see that the difference from (66) is the use of the kernel matrix $\boldsymbol{K}$ in place of $\boldsymbol{\Gamma} = \sqrt{\boldsymbol{K}}$. A merit of this kernelization strategy is that no logical incongruence is involved about the parallelism between the primal variable and dual variables. On the other hand, recalling that U/L grows as the dimension does, and that the dimension is now equal to the number of samples. Accordingly, the proximity can deteriorate, or deviation from the ℓ_2-case can be significant. Especially when we employ the ℓ_1-norm, it is noteworthy that optimal solution can be far away from the ℓ_2-case. Another possible demerit of this strategy is that the size of the reformulated optimization problem (e.g., LP) can involve a large number of constraints. However, as the above formulations look, if we can optimize the norm and the risk functional as they are, this problem can be mitigated. In this sense, recent development of indifferentiable optimization and the stochastic gradient-based optimization is expected.

References

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

Le Thi Hoai An, Pham Dinh Tao. The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems. *Annals of Operations Research*, 133:23–46, 2005.

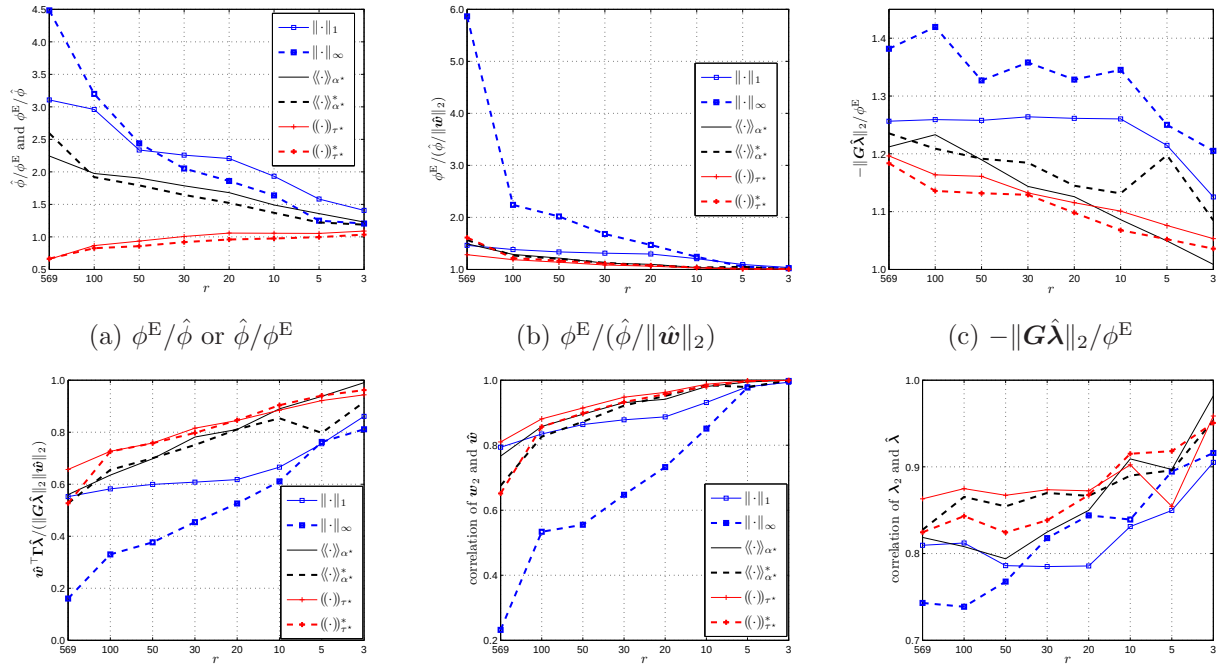
Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, David Heath. Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228, 1999.

Francis R. Bach, Michael I. Jordan. Predictive Low-rank Decomposition for Kernel Methods. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

Peter L. Bartlett, Michael I. Jordan, Jon D. McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Kristin P. Bennett, E.J. Bredensteiner. Duality and Geometry in SVM Classifiers. *Proceedings of the International Conference on Machine Learning*, 57–64, 2000.

Kristin P. Bennett, Olvi L. Mangasarian. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. *Optimization Methods and Software*, 1:23–34, 1992.



(a) $\phi^{\mathrm{E}}/\hat{\phi}$ or $\hat{\phi}/\phi^{\mathrm{E}}$ (b) $\phi^{\mathrm{E}}/(\hat{\phi}/\|\hat{\boldsymbol{w}}\|_2)$ (c) $-\|\boldsymbol{G}\hat{\boldsymbol{\lambda}}\|_2/\phi^{\mathrm{E}}$

(d) $\hat{\boldsymbol{w}}^\top\boldsymbol{\Gamma}\hat{\boldsymbol{\lambda}}/(\|\boldsymbol{\Gamma}\hat{\boldsymbol{\lambda}}\|_2\|\hat{\boldsymbol{w}}\|_2)$

Correlation coefficients

(e) between $\boldsymbol{w}_2$ and $\hat{\boldsymbol{w}}$ (f) between $\boldsymbol{\lambda}_2$ and $\hat{\boldsymbol{\lambda}}$

Figure 3: Improvement of proximity values to the ℓ_2-classifier (WDBC data set)

Aharon Ben-Tal, Marc Teboulle. An Old-new Concept of Convex Risk Measures: The Optimized Certainty Equivalent. *Mathematical Finance*, 17:449–476, 2007.

Dimitris Bertsimas, Dessislava Pachamanova, Melvyn Sim. Robust Linear Optimization under General Norms. *Operations Research Letters*, 32:510–516, 2004.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT92)*, pages 144–152, ACM Press, New York, New York, USA, 1992.

Stephen Boyd, Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

Emmanuel J. Candés and Yaniv Plan. Near-ideal Model Selection by ℓ_1 Minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

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

Pai-Hsuen Chen, Chih-Jen Lin, and Bernhard Schölkopf. A tutorial on ν-support vector machines. *Appl. Stochas. Models Bus. Indust.*, 21:111–136, 2005.

Andreas Christmann, Ingo Steinwart. On Robustness Properties of Convex Risk Minimization Methods for Pattern Recognition. *Journal of Machine Learning Research*, 5:1007–1034, 2004.

Ronan Collobert, Fabian Sinz, Jason Weston, Léon Bottou. Trading Convexity for Scalability. *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, 2006.

Corinna Cortes, Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

David J. Crisp, Christopher J.C. Burges. A Geometric Interpretation of ν-SVM Classifiers. In Sara A. Solla, Todd K. Leen and Klaus-Robert Müller, editors. *Advances in Neural Information Processing Systems 12*, pages 244–250. MIT Press, Cambridge, Massachusetts, USA, 2000.

I. Csiszár. Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungarica*, 2:299–318, 1967.

Fine S, Scheinberg K. Efficient SVM Training Using Low-Rank Kernel Representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

Peter C. Fishburn. Mean-Risk Analysis with Risk Associated with Below-Target Returns. *The American Economic Review*, 67(2):116–126, 1977.

Hans Föllmer, Alexander Schied. Convex Measures of Risk and Trading Constraints. *Finance and Stochastics*, 6(4):429–447.

Yoav Freund and Robert E. Schapire. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55:119–139.

Jun-ya Gotoh, Akiko Takeda. A Linear Classification Model Based on Conditional Geometric Score. *Pacific Journal of Optimization*, 1(2):277–296, 2005.

Jun-ya Gotoh, Akiko Takeda, Rei Yamamoto. Interaction between Financial Risk Measures and Machine Learning Methods. *Computational Management Science*, forthcoming. `DOI: 10.1007/s10287-013-0175-5`

Jun-ya Gotoh, Stan Uryasev. Two Pairs of Families of Polyhedral Norms Versus ℓ_p-Norms: Proximity and Applications in Optimization. Research report #2013-3, Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida, 2013. (Downloadable from `www.ise.ufl.edu/uryasev/publications/`)

Michael Grant and Stephen Boyd. CVX: MATLAB Software for Disciplined Convex Programming, Version 2.0 Beta. `cvxr.com/cvx`, 2012.

Trevor Hastie, Saharon Rosset, Robert Tibshirani. The Entire Regularization Path for the Support Vector Machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

Kaizhu Huang, Danian Zheng, Irwin King, Michael R. Lyu. Arbitrary Norm Support Vector Machines. *Neural Computation*, 21(2):560–582, 2009.

Peter J. Huber. *Robust Statistics*. Wiley, Hoboken, New Jersey, USA, 1981.

Takafumi Kanamori, Akiko Takeda, Taiji Suzuki. Conjugate Relation between Loss Functions and Uncertainty Sets in Classification Problems. *Journal of Machine Learning Research*, 14:1461–1504, 2013.

Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. 2007. An Interior-Point Method for Large-Scale ℓ_1-Regularized Logistic Regression *J. Mach. Learn. Res.*, 8, 1519–1555, 2007.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. 2011. ℓ_p-Norm Multiple Kernel Learning. *J. Mach. Learn. Res.*, 12, 953–997, 2011.

Hiroshi Konno, Hiroaki Yamazaki. Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock Market. *Management Science*, 37(5):519–531, 1991.

Roi Livni, Koby Crammer, and Amir Globerson. A Simple Geometric Interpretation of SVM using Stochastic Adversaries. Proceedings of the 15th International Conference on Artificial Intelligence and Statistics, 2012.

Olvi L. Mangasarian. Generalized Support Vector Machines. In: Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf and Dale Schuurmans, editors. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, Massachusetts, USA, 2000.

Olvi L. Mangasarian. Arbitrary-Norm Separating Plane. *Operations Research Letters*, 24:15–23, 1999.

Harry Markowitz. Portfolio Selection *The Journal of Finance*, 7(1):77–91, 1952.

Konstantin Pavlikov, Stan Uryasev. CVaR Norm and Applications in Optimization. *Optimization Letters*, forthcoming.

João P. Pedroso, Noboru Murata. Support Vector Machines with Different Norms: Motivation, Formulations and Results. *Pattern Recognition Letters*, 22:1263–1272, 2001.

Fernando Perez-Cruz, Jason Weston, Daniel Herrmann, Bernhard Schölkopf. Extension of the ν-SVM Range for Classification. In: Johan A.K. Suykens, G. Horvath, S. Basu, C. Micchelli and J. Vandewalle, editors. *Advances in Learning Theory: Methods, Models and Applications 190*, pages 179–196, IOS Press, Amsterdam, Netherlands, 2003.

Gunnar Rätsch, Bernhard Schölkopf, Alexander J. Smola, Sebastian Mika, Takashi Onoda, Klaus-Robert Müller. Robust Ensemble Learning. Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, D. Schuurmans, editors. *Advances in Large Margin Classifiers*, 207–219, MIT Press, Cambridge, MA, USA, 2000.

Mark D. Reid and Robert C. Williamson. Information, Divergence and Risk for Binary Experiments. *The Journal of Machine Learning Research* 12, 731–817, 2011.

Ryan M. Rifkin and Ross A. Lippert. Value regularization and Fenchel duality. *The Journal of Machine Learning Research* 8, 441–479, 2007.

R Terry Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

R Terry Rockafellar, Stan Uryasev. Optimization of Conditional Value-At-Risk. *The Journal of Risk*, 2(3):21–41, 2000.

R Terry Rockafellar, Stan Uryasev. Conditional Value-at-Risk for General Loss Distributions. *Journal of Banking and Finance* 26:1443–1471, 2002.

R Terry Rockafellar, Stan Uryasev. The Fundamental Risk Quadrangle in Risk Management, Optimization and Statistical Estimation. *Surveys in Operations Research and Management Science*, forthcoming.

Andrzej Ruszczyński, Alexander Shapiro. Optimization of risk measures. In Giuseppe Calafiore, Fabrizio Dabbene, editors. *Probabilistic and Randomized Methods for Design Under Uncertainty*, pages 117–158, Springer-Verlag, London, UK, 2005.

Andrzej Ruszczyński, Alexander Shapiro. Optimization of Convex Risk Functions. *Mathematics of Operations Research*, 31(3):433–452, 2006.

Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, Peter L. Bartlett. New Support Vector Algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

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

Akiko Takeda, H Mitsugi, Takafumi Kanamori. A Unified Classification Model Based on Robust Optimization. *Neural Computation*, 25(3):759–804, 2013.

Akiko Takeda, Masashi Sugiyama. ν-Support Vector Machine as Conditional Value-at-Risk Minimization. *Proceedings of the 25 th International Conference on Machine Learning*, 1056–1063, 2008.

Peter Tsyurmasto, Jun-ya Gotoh, Stan Uryasev. (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/`)

Ryota Tomioka, Taiji Suzuki, and Masashi Sugiyama. Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparsity Regularized Estimation. *The Journal of Machine Learning Research*, 12:1537-1586, 2011.

Yongqiao Wang. Robust ν-Support Vector Machine Based on Worst-case Conditional Value-at-Risk Minimization. *Optimization Methods and Software*, 27(6):1025–1038, 2012.

Wolfram Wiesemann, Daniel Kuhn, Melvyn Sim. Distributionally Robust Convex Optimization. Discussion Paper, *Optimization Online*. `www.optimization-online.org/DB_HTML/2013/02/3757.html` (accessed Jun 21, 2013)

William H. Wolberg, W. Nick Street, Olvi L. Mangasarian. Wisconsin Diagnostic Breast Cancer (WDBC) Data Set. `ftp.cs.wisc.edu/math-prog/cpo-dataset/ machine-learn/cancer/WDBC/` (accessed July 24, 2013)

Huan Xu, Constantine Caramanis, Shie Mannor and Sungho Yun. Risk Sensitive Robust Support Vector Machines. In *48th IEEE Conference on Decision and Control (CDC09)*, Shanghai, China, 2009.

Huan Xu, Constantine Caramanis and Shie Mannor. Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10 1485-1510, (2009).

Shlomo Yitzhaki. Stochastic Dominance, Mean Variance, and Gini's Mean Difference. *The American Economic Review*, 72(1):178–185, 1982.

Mike Zabarankin, Stan Uryasev (2013) *Statistical Decision Problems: Selected Concepts and Portfolio Safeguard Case Studies*. Springer.

Tong Zhang. Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *The Annals of Statistics*, 32(1):56–134.

Weida Zhou, Li Zhang, Licheng Jiao. Linear Programming Support Vector Machines. *Pattern Recognition*, 35:2927–2936, 2002.