10

CVaR Minimizations in Support Vector Machines

Jun-ya Gotoh[1] and Akiko Takeda[2]
[1]*Chuo University, Japan*
[2]*The University of Tokyo, Japan*

How to measure the riskiness of a random variable has been a major concern of financial risk management. Among the many possible measures, *conditional value at risk* (*CVaR*) is viewed as a promising functional for capturing the characteristics of the distribution of a random variable. CVaR has attractive theoretical properties, and its minimization with respect to involved parameters is often tractable. In portfolio selection especially, the minimization of the empirical CVaR is a linear program. On the other hand, machine learning is based on the so-called regularized empirical risk minimization, where a surrogate of the empirical error defined over the in-sample data is minimized under some regularization of the parameters involved. Considering that both theories deal with empirical risk minimization, it is natural to look at their interaction. In fact, a variant of support vector machine (SVM) known as ν-SVM implicitly carries out a certain CVaR minimization, though the relation to CVaR is not clarified at the time of the invention of ν-SVM.

This chapter overviews the connections between SVMs and CVaR minimization and suggests further interactions beyond their similarity in appearance. Section 10.1 summarizes the definition and properties of CVaR. The authors wish this section to be a quick introduction for those who are not familiar with CVaR. Section 10.2 collects basic formulations of various SVMs for later reference. Those who are familiar with SVM formulations can skip this section and consult it when succeeding sections refer to the formulations therein. Section 10.3 provides CVaR minimization-based representations of the SVM formulations given in Section 10.2. Section 10.4 is devoted to dual formulations of CVaR-based formulations. Section 10.5 describes two robust optimization extensions of these formulations. For further

Financial Signal Processing and Machine Learning, First Edition.
Edited by Ali N. Akansu, Sanjeev R. Kulkarni and Dmitry Malioutov.

study, Section 10.6 briefly overviews the literature regarding the interaction between risk measure theory and SVM.

10.1 What Is CVaR?

This section introduces CVaR as a risk measure of random variables and describes its relation to VaR and other statistics.

10.1.1 Definition and Interpretations

Let $\tilde{L}$ be a random variable defined on a sample space Ω (i.e., $\tilde{L} : \Omega \to \mathbb{R}$), and let it represent a quantity that we would like to minimize, such as payments, costs, damages, or error. We figuratively refer to such random variables as *losses*.[1] In financial risk management, each element $\omega \in \Omega$ can be interpreted as a future state or scenario. Furthermore, let us suppose that all random variables are associated with a probability measure $\mathbb{P}$ on Ω (and a set of events), satisfying $\mathbb{E}_{\mathbb{P}}[|\tilde{L}|] < +\infty$. In most parts of this chapter, this assumption is satisfied because losses are associated with empirical distributions based on finite observations, and each loss is defined on a finite sample space (i.e., $\Omega = \{\omega_1, \ldots, \omega_m\}$).

For a loss $\tilde{L}$, a *risk measure* is a functional that maps $\tilde{L}$ to $\mathbb{R} \cup \{\infty\}$, expressing how risky $\tilde{L}$ is. Among the many risk measures, VaR is popular because it is easy to interpret.

Definition 10.1 (VaR (value-at-risk) or α-quantile of loss) *The VaR of $\tilde{L}$ at a significant level $\alpha \in (0, 1)$ is defined as*

$$VaR_{(\alpha,\mathbb{P})}[\tilde{L}] := \min_c \{c : \mathbb{P}\{\tilde{L} \le c\} \ge \alpha\}.$$

The parameter α is determined by decision makers such as fund managers. To capture the possibility of a large loss that could occur with a small probability, the parameter α is typically set close to 1 in financial risk management (e.g., $\alpha = 0.99$ or 0.999). While VaR can capture the upper tail of $\tilde{L}$, it fails to capture the impact of a loss beyond VaR. In addition, the lack of convexity makes VaR intractable in risk management.[2]

CVaR has gained growing popularity as a convex surrogate to VaR.

Definition 10.2 (CVaR or α-superquantile of loss) *The conditional value-at-risk (CVaR) of $\tilde{L}$ with a significant level $\alpha \in [0, 1)$ is defined as*

$$CVaR_{(\alpha,\mathbb{P})}[\tilde{L}] := \inf_c \left\{ G(c) := c + \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}}[\max\{\tilde{L} - c, 0\}] \right\}. \tag{10.1}$$

$\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$ is nondecreasing in α since the function G is nondecreasing in α.

[1] In SVM and other statistical learning contexts, the word *loss* takes on a specific meaning. However, in this chapter, we will use this word in a different and more general manner for the sake of consistency with risk measure theory.
[2] A functional $\mathcal{F}$ is *convex* if for all $\tilde{L}, \tilde{L}', \tau \in (0, 1)$, $(1-\tau)\mathcal{F}[\tilde{L}] + \tau\mathcal{F}[\tilde{L}'] \ge \mathcal{F}[(1-\tau)\tilde{L} + \tau\tilde{L}']$.

Note that for $\alpha \in (0,1)$, "inf" in the above formula is replaced with "min." Indeed, for $\alpha \in (0,1)$, Theorem 10 of Rockafellar and Uryasev (2002) shows that

$$\arg\min_c G(c) = [\mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}], \mathrm{VaR}^+_{(\alpha,\mathbb{P})}[\tilde{L}]], \tag{10.2}$$

where $\mathrm{VaR}^+_{(\alpha,\mathbb{P})}[\tilde{L}] := \inf_c \{c : \mathbb{P}\{\tilde{L} \le c\} > \alpha\}$. The relation (10.2) implies that any minimizer in the formula (10.1) can be an approximate VaR and that if the minimizer is unique,[3] it is equal to VaR.

The definition (10.1) of CVaR implies that VaR is always bounded above by CVaR, that is,

$$\mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}] \le \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}],$$

and $\mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}] < \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$ unless there is no chance of a loss greater than VaR. More specifically, CVaR is represented as a convex combination of VaR and $\mathbb{E}_{\mathbb{P}}[\tilde{L}|\tilde{L} > \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]]$.[4] Indeed, with $t_\alpha := \mathbb{P}\{\tilde{L} > \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]\}/(1-\alpha) \in [0,1]$, it is true that

$$\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] = t_\alpha \mathbb{E}_{\mathbb{P}}[\tilde{L}|\tilde{L} > \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]] + (1-t_\alpha)\mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]. \tag{10.3}$$

When $\tilde{L}$ follows a parametric distribution, CVaR may be explicitly given a closed formula.

Example 10.1 (CVaR under normal distribution) *When $\tilde{L}$ follows a normal distribution $N(\mu,\sigma^2)$, CVaR can be explicitly expressed as*

$$CVaR_{(\alpha,\mathbb{P})}[\tilde{L}] = \mu + \frac{1}{(1-\alpha)\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{\Psi^{-1}(\alpha)\right\}^2\right)\cdot\sigma, \tag{10.4}$$

where Ψ is the cumulative distribution function of $N(0,1)$ (i.e., $\Psi(z) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp(-\frac{1}{2}t^2)dt$). On the other hand, $VaR_{(\alpha,\mathbb{P})}[\tilde{L}] = \mu + \Psi^{-1}(\alpha)\cdot\sigma$.

Note that (10.4) is equal to the conditional expectation of a loss exceeding VaR, that is,

$$\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] = \mathbb{E}_{\mathbb{P}}[\tilde{L}|\tilde{L} > \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]] = \mathbb{E}_{\mathbb{P}}[\tilde{L}|\tilde{L} \ge \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]].$$

This relation also holds for other distributions as long as there is no probability atom at $\mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$.

However, for general loss distributions and arbitrary α, the above equalities only hold in an approximate manner. That is, we have

$$\mathbb{E}_{\mathbb{P}}[\tilde{L}|\tilde{L} \ge \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]] \le \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] \le \mathbb{E}_{\mathbb{P}}[\tilde{L}|\tilde{L} > \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]].$$

Usually, the loss distributions are not known, so an empirical distribution is used as a surrogate of the true distribution. In such a case, there is a positive probability atom at $\mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$ for all $\alpha \in (0,1)$, and either of the equalities does not hold (see Proposition 5 of Rockafellar and Uryasev (2002) for the details).

[3] The minimizer is unique if and only if $\mathrm{VaR}^+_{(\alpha,\mathbb{P})}[\tilde{L}] = \mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$ (i.e., there is no probability atom at $\mathrm{VaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$).

[4] $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$ is also considered to be the mean of the α-tail distribution of $\tilde{L}$. See Proposition 6 of Rockafellar and Uryasev (2002) for the details.

Example 10.2 (CVaR under finite scenarios) *Suppose that $\tilde{L}$ is defined on a finite sample space $\Omega = \{\omega_1, \dots, \omega_m\}$ equipped with $\mathbb{P}\{\omega = \omega_i\} = p_i > 0$, $i = 1, \dots, m$. For $\alpha \in [0, 1)$, CVaR is given by*

$$CVaR_{(\alpha,\mathbb{P})}[\tilde{L}] = CVaR_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) := \min_{c} \left\{ c + \frac{1}{1-\alpha} \sum_{i=1}^{m} p_i \max\{L_i - c, 0\} \right\},$$

where $\boldsymbol{L} := (L_1, \dots, L_m)^\top$ and $\boldsymbol{p} := (p_1, \dots, p_m)^\top$.[5] *In this case, CVaR can be represented with a linear program (LP).*

$$CVaR_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \begin{vmatrix} \underset{c,z}{\text{minimize}} & c + \frac{1}{1-\alpha} \sum_{i=1}^{m} p_i z_i, \\ \text{subject to} & z_i \geq L_i - c, i = 1, \dots, m, \\ & z_i \geq 0, i = 1, \dots, m. \end{vmatrix} \tag{10.5}$$

Note that (10.5) is feasible for any $\boldsymbol{L}$, $\boldsymbol{p}$, and α (e.g., $(c, z_1, \dots, z_m) = (\bar{c}, L_1 - \bar{c}, \dots, L_m - \bar{c})$ with $\bar{c} = \min\{L_1, \dots, L_m\}$ is a feasible solution). The dual problem of (10.5) is derived as

$$\begin{vmatrix} \underset{\boldsymbol{q}}{\text{maximize}} & \sum_{i=1}^{m} q_i L_i, \\ \text{subject to} & \sum_{i=1}^{m} q_i = 1, \\ & 0 \leq q_i \leq \frac{p_i}{1-\alpha}, i = 1, \dots, m. \end{vmatrix} \tag{10.6}$$

The solution $\boldsymbol{q} = \boldsymbol{p}$ is feasible for (10.6) for any $\alpha \in [0, 1)$, and therefore, the optimal value of (10.6) is equal to $CVaR_{(\alpha,\boldsymbol{p})}(\boldsymbol{L})$ because of the strong duality of LP (see, e.g., Vanderbei, 2014).

As will be elaborated on later, (10.6) suggests that CVaR can be interpreted as the worst expected loss over a set of probability measures, $\{\boldsymbol{q} : \sum_{i=1}^{m} q_i = 1, 0 \leq q_i \leq \frac{p_i}{1-\alpha}, i = 1, \dots, m\}$. Note also that (10.6) can be viewed as a variant of the continuous (or fractional) knapsack problem,[6] *and its solution is obtained in a greedy manner as follows:*[7]

1. *Sort the loss scenarios, $L_1, \dots, L_m$, in descending order, and let $L_{(i)}$ denote the i-th largest component (i.e., $L_{(1)} \geq L_{(2)} \geq \cdots \geq L_{(m)}$), and $p_{(i)}$, denote the reference probability corresponding to $L_{(i)}$.*
2. *Find the integer k satisfying $\frac{1}{1-\alpha} \sum_{i=1}^{k} p_{(i)} \leq 1 < \frac{1}{1-\alpha} \sum_{i=1}^{k+1} p_{(i)}$. Let $q_{(i)}$ denote the element of the solution vector $\boldsymbol{q}$, corresponding to $L_{(i)}$ in the objective of (10.6), and set $q_{(i)} = \frac{p_{(i)}}{1-\alpha}$ for $i = 1, \dots, k$, $q_{(k+1)} = 1 - \frac{1}{1-\alpha} \sum_{i=1}^{k} p_{(i)}$, and $q_{(i)} = 0$ for $i = k+2, \dots, m$. Consequently, CVaR is given by the formula*

$$CVaR_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \frac{1}{1-\alpha} \sum_{i=1}^{k} p_{(i)} L_{(i)} + (1 - \frac{1}{1-\alpha} \sum_{i=1}^{k} p_{(i)}) L_{(k+1)}. \tag{10.7}$$

[5] Since the random variable $\tilde{L}$ and probability measure $\mathbb{P}$ can be expressed as vectors $\boldsymbol{L} := (L_1, \dots, L_m)^\top$ and $\boldsymbol{p} := (p_1, \dots, p_m)^\top$, we denote $\text{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$ by $\text{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L})$.

[6] While the *knapsack problem* is formulated as a binary integer program $\max\{\sum_{i=1}^{m} b_i x_i : \sum_{i=1}^{m} c_i x_i \leq C, \boldsymbol{x} \in \{0, 1\}^m\}$, where $c_i, b_i, C > 0$, the corresponding continuous knapsack problem (CKP) is formulated as an LP: $\max\{\sum_{i=1}^{m} b_i x_i : \sum_{i=1}^{m} c_i x_i = C, \boldsymbol{x} \in [0, 1]^m\}$. With a change of variables $x_i = (1 - \alpha) q_i / p_i$, (10.6) can be rewritten into an LP with $b_i = p_i L_i / (1 - \alpha)$, $c_i = p_i$, and $C = 1 - \alpha$. Strictly speaking, (10.6) is not a CKP since L_i, and thus b_i, can be negative. However, the same greedy algorithm is applicable.

[7] This procedure shows that the complexity of computing the CVaR of a loss is at most on the order of $m \log m$. If $\boldsymbol{p}$ is uniform (i.e., $p_i = 1/m$), we can use the so-called selection algorithm whose complexity is on the order of m.

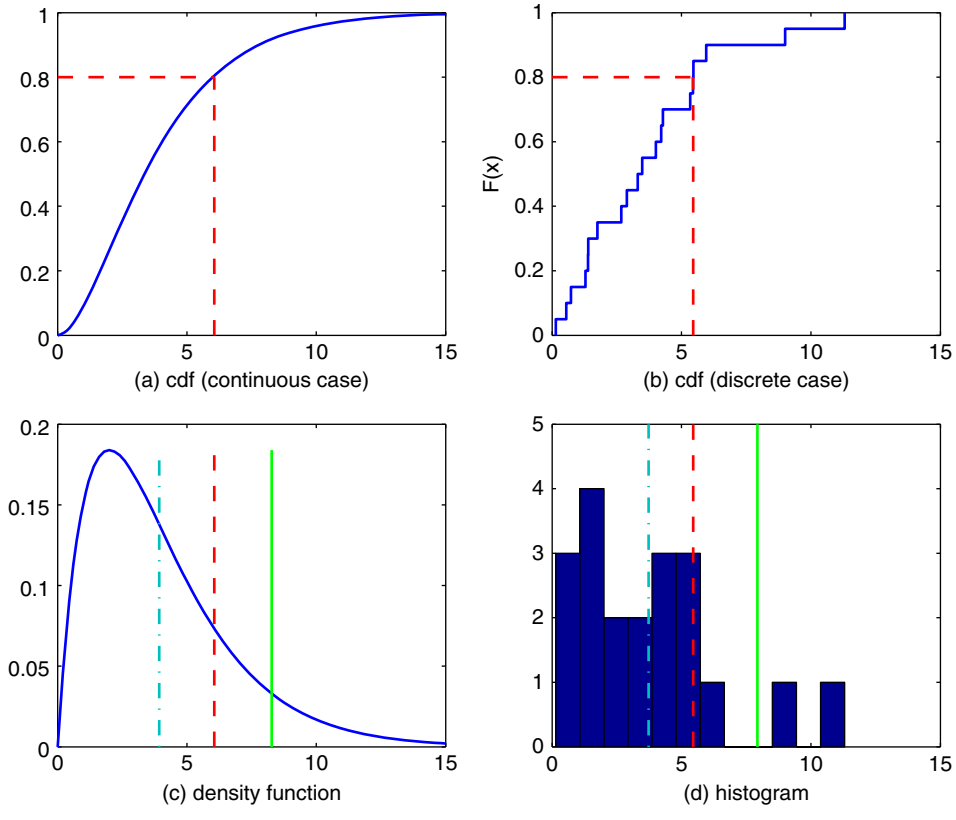

Figure 10.1 CVaR, VaR, mean, and maximum of distribution. (a, c) The cumulative distribution function (cdf) and the density of a continuous loss distribution; (b, d) the cdf and histogram of a discrete loss distribution. In all four figures, the location of VaR with $\alpha = 0.8$ is indicated by a vertical dashed line. In (c) and (d), the locations of CVaR and the mean of the distributions are indicated with vertical solid and dashed-dotted lines. In (b) and (d), the location of the maximum loss is shown for the discrete case.

Formula (10.7) implies that CVaR is approximately equal to the mean of the largest $100(1-\alpha)$ *% losses* $L_{(1)}, \ldots, L_{(k)}$ *if* $\sum_{i=1}^{k} p_{(i)} \approx 1-\alpha$ *(see Figures 10.1b and 10.1d).*

Now let us consider the complementarity condition. Namely, for any optimal solutions (c^*, z^*) *and* $\boldsymbol{q}^*$ *to (10.5) and (10.6), respectively, the condition*

$$\begin{cases} (z_i^* - L_i + c^*)q_i^* = 0, & i = 1, \ldots, m, \\ z_i^*\{p_i - (1-\alpha)q_i^*\} = 0, & i = 1, \ldots, m, \end{cases}$$

implies $z_{(i)}^* = L_{(i)} - c^*$, $i = 1, \ldots, k$, *and* $z_{(i)}^* = 0$, $i = k+1, \ldots, m$, *where* $z_{(i)}^*$ *is the element of the optimal solution corresponding to* $L_{(i)}$ *and* $p_{(i)}$. *Using this for the objective of (10.5) and*

comparing the result with (10.7), we find that $c^* = L_{(k+1)}$. *Note that*

$$L_{(k+1)} = \begin{cases} VaR_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) & \text{if} \quad \sum_{i=1}^{k} p_{(i)} < 1-\alpha < \sum_{i=1}^{k+1} p_{(i)}, \\ VaR^{+}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) & \text{if} \quad \sum_{i=1}^{k} p_{(i)} = 1-\alpha, \end{cases}$$

where $VaR^{+}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L})$ *is* $VaR^{+}_{(\alpha,\mathbb{P})}$ *of* $\boldsymbol{L}$ *with* $\mathbb{P} = \boldsymbol{p}$, *that is,* $\min_c\{c : \sum_{i=1}^{m} p_i 1_{\{L_i \le c\}} > \alpha\}$, *where* $1_{\{cond\}}$ *is the 0–1 indicator function (i.e.,* $1_{\{\text{cond}\}} = 1$ *if cond is true, 0 otherwise). Accordingly, formula (10.7) is an expression (10.3) for the discrete distribution case.*

CVaR can also be considered to be a generalization of the average and the maximum of the underlying random variable.

- With $\alpha = 0$, CVaR is equal to the mean of the loss, that is, $\text{CVaR}_{(0,\mathbb{P})}[\tilde{L}] = \mathbb{E}_{\mathbb{P}}[\tilde{L}]$ (or $\text{CVaR}_{(0,\boldsymbol{p})}(\boldsymbol{L}) = \mathbb{E}_{\boldsymbol{p}}(\boldsymbol{L}) := \boldsymbol{p}^{\top}\boldsymbol{L}$ for the discrete distribution case).
- With α close to 1, it approximates the largest loss. Indeed, in the case of a discrete distribution, it is true that $\text{CVaR}_{(\alpha',\boldsymbol{p})}(\boldsymbol{L}) = \max\{L_1, \ldots, L_m\} = L_{(1)}$ for $\alpha' > 1 - p_{(1)}$.

As a result of the nondecreasing property with respect to α, CVaR is typically between the maximum and mean for an intermediate $\alpha \in (0, 1 - p_{(1)})$ (see Figure 10.1d).

10.1.2 Basic Properties of CVaR

The convexity of a risk functional often makes the associated risk minimization tractable and enables us to exploit duality theory. Those advantages of the convexity of CVaR are worth emphasizing again.

Let us consider a probabilistic representation. Letting B be the target level of loss and $\alpha \in (0, 1)$ be the significant level, define

$$\mathbb{P}\{\tilde{L} \ge B\} \le 1 - \alpha. \tag{10.8}$$

In general, the set of losses $\tilde{L}$ satisfying (10.8) is nonconvex. To avoid this nonconvexity, the left side of (10.8) is often approximated with a convex upper bound of the form $\mathbb{E}_{\mathbb{P}}[f(\tilde{L})]$, where f is a convex function on $\mathbb{R}$ such that $f(L) \ge 1_{\{L \ge B\}}$ for all $L \in \mathbb{R}$. Note that $\mathbb{E}_{\mathbb{P}}[f(\tilde{L})] \le 1 - \alpha$ implies $\mathbb{P}\{\tilde{L} \ge B\} \le 1 - \alpha$ since $\mathbb{P}\{\tilde{L} \ge B\} = \mathbb{E}_{\mathbb{P}}[1_{\{\tilde{L} \ge B\}}] \le \mathbb{E}_{\mathbb{P}}[f(\tilde{L})]$. The expression $\mathbb{E}_{\mathbb{P}}[f(\tilde{L})] \le 1 - \alpha$ is thus called a *conservative approximation* of (10.8). To tighten the bound, it is enough to consider a piecewise linear function $f(L) = \max\{(L - C)/(B - C), 0\}$ with some C such that $B > C$ (see Figure 10.2). Namely, (10.8) can be replaced with $\mathbb{E}_{\mathbb{P}}[\max\{(\tilde{L} - C)/(B - C), 0\}] \le 1 - \alpha$, which becomes

$$C + \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}}[\max\{\tilde{L} - C, 0\}] \le B \quad \text{for some } C.$$

Noting that this is equivalent to $\text{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] \le B$, we can see that CVaR provides a tight convex conservative approximation of the probabilistic condition (10.8).

CVaR has three properties that are useful in financial risk management; CVaR is

1. *monotonic*: $\text{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] \ge \text{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}']$ when $\tilde{L} \ge \tilde{L}'$;

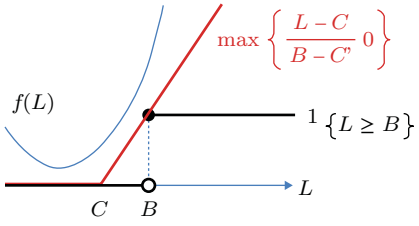

Figure 10.2 Convex functions dominating $1_{\{L \geq B\}}$.

2. *translation invariant*: $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L} + \tau] = \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] + \tau$ for all $\tau \in \mathbb{R}$; and
3. *positively homogeneous*: $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tau\tilde{L}] = \tau\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}]$ for all $\tau > 0$.

If a convex risk functional satisfies all three of these properties, it is said to be *coherent* (Artzner *et al.*, 1999).[8] Monotonicity is a useful property in machine learning contexts, whereas translation invariance and positive homogeneity exist for technical reasons rather than for intuitive reasons. However, each of these properties plays roles in tractability and in compatibility with regularization terms (Gotoh and Uryasev, 2013).

In general, monotonicity, translation invariance, and positive homogeneity can be characterized in a dual manner. To avoid unnecessary technicalities, we will assume a finite sample space $\Omega = \{\omega_1, \ldots, \omega_m\}$ (i.e., which implies that risk functionals $\mathcal{F}$ are functions on $\mathbb{R}^m$).[9] Let us define $\mathcal{F}^*$ to be the conjugate of $\mathcal{F}$, that is, $\mathcal{F}^*(\lambda) := \sup_{\boldsymbol{L}}\{\boldsymbol{L}^\top\boldsymbol{L} - \mathcal{F}(\boldsymbol{L})\}$. Furthermore, let us define dom $\mathcal{F}$ as the effective domain of $\mathcal{F}$, that is, dom $\mathcal{F} := \{\boldsymbol{L} \in \mathbb{R}^m : \mathcal{F}(\boldsymbol{L}) < +\infty\}$.

Theorem 10.1 (dual characterization of risk functional properties (Ruszczyński and Shapiro, 2006)) *Suppose that $\mathcal{F} : \mathbb{R}^m \to (-\infty, \infty)$ is an l.s.c.,*[10] *proper,*[11] *and convex function. Accordingly:*

1. *$\mathcal{F}$ is monotonic if and only if dom $\mathcal{F}^*$ is in the nonnegative orthant.*
2. *$\mathcal{F}$ is translation invariant if and only if $\forall\lambda \in$ dom $\mathcal{F}^*$, $\mathbf{1}_m^\top\lambda = 1$.*
3. *$\mathcal{F}$ is positively homogeneous if and only if $\mathcal{F}$ can be represented in the form*

$$\mathcal{F}(\boldsymbol{L}) = \sup_{\lambda}\{\boldsymbol{L}^\top\lambda : \lambda \in \mathrm{dom}\,\mathcal{F}^*\}. \tag{10.9}$$

See Ruszczyński and Shapiro (2006) for the proof.

[8] CVaR is also *law invariant* (i.e., if the distribution functions of $\tilde{L}$ and $\tilde{L}'$ are identical, $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] = \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}']$) and *co-monotonically additive* (i.e., $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L} + \tilde{L}'] = \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] + \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}']$ for any $\tilde{L}, \tilde{L}'$ satisfying $(\tilde{L}(\omega) - \tilde{L}(\omega'))(\tilde{L}'(\omega) - \tilde{L}'(\omega')) \geq 0$ for any $\omega, \omega' \in \Omega$). A coherent risk measure that has these two properties is called a *spectral (or distortion) risk measure*. See, for example, Acerbi (2002) for details.

[9] The results hold true in a more general setting. See, for example, Rockafellar and Uryasev (2013) and Ruszczyński and Shapiro (2006) for more general statements.

[10] $\mathrm{CVaR}_{(\alpha,\mathbb{P})}$ is *lower semicontinuous (l.s.c.)*, that is, $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] \leq \liminf_{k\to\infty} \mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}^k]$ for any $\tilde{L}$ and any sequence $\tilde{L}^1, \tilde{L}^2, \ldots$, converging to $\tilde{L}$.

[11] $\mathrm{CVaR}_{(\alpha,\mathbb{P})}$ is *proper*, that is, $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}] > -\infty$ for all $\tilde{L}$, and $\mathrm{CVaR}_{(\alpha,\mathbb{P})}[\tilde{L}'] < \infty$ for some $\tilde{L}'$.

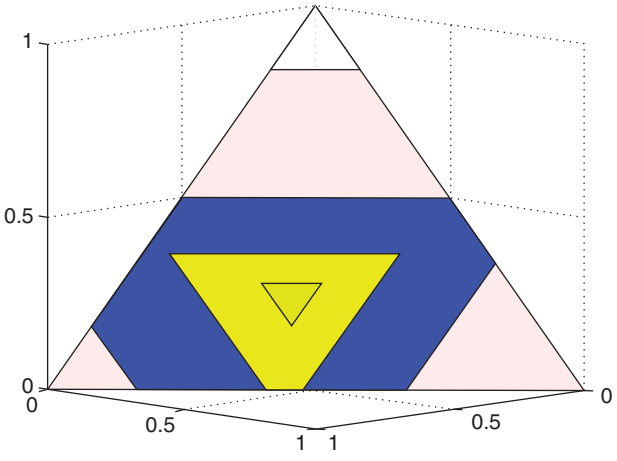

Figure 10.3 Illustration of $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}$ in a discrete distribution on $\mathbb{R}^3$ with $(p_1, p_2, p_3) = (5/12, 4/12, 3/12)$. This figure shows how $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}$ varies depending on α ($\alpha = 0.1, 0.3, 0.5, 0.7$). As α approaches 1, $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}$ approaches the unit simplex $\mathbb{\Pi}^3$. The risk envelope shrinks to the point $(p_1, p_2, p_3) = (5/12, 4/12, 3/12)$ as α decreases to 0.

In particular, from (10.9), we can see that any l.s.c. proper positively homogeneous convex risk functional can be characterized by the effective domain of its conjugate, which is referred to as the *risk envelope*. Let us denote the risk envelope of CVaR by $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}$ (i.e., $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})} := \mathrm{dom}\, \mathrm{CVaR}^*_{(\alpha,\boldsymbol{p})}$).

Noting that the dual LP (10.6) is written in the form of (10.9), the risk envelope of CVaR is

$$\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})} := \{\boldsymbol{q} \in \mathbb{R}^m : \mathbf{1}_m^\top \boldsymbol{q} = 1, \mathbf{0} \le \boldsymbol{q} \le \boldsymbol{p}/(1-\alpha)\}. \tag{10.10}$$

Figure 10.3 illustrates an example of the risk envelope of CVaR with $m = 3$. The conjugate of CVaR is given by

$$\mathrm{CVaR}^*_{(\alpha,\boldsymbol{p})}(\lambda) = \delta_{\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}}(\lambda),$$

where δ_C is the indicator function of a set C (i.e., $\delta_C(\xi) := 0$ if $\xi \in C$, and $+\infty$ otherwise). Since $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})} \subset \mathbb{\Pi}^m := \{\boldsymbol{q} \in \mathbb{R}^m : \mathbf{1}_m^\top \boldsymbol{q} = 1, \boldsymbol{q} \ge \mathbf{0}\}$, the dual LP (10.6) is symbolically represented as the worst-case expected loss over a set of probabilities, that is,

$$\mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(\boldsymbol{L}) = \max_{\boldsymbol{q}} \{\mathbb{E}_{\boldsymbol{q}}(\boldsymbol{L}) : \boldsymbol{q} \in \mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}\}. \tag{10.11}$$

Indeed, any coherent function on $\mathbb{R}^m$ can be characterized by using a non-empty closed convex set in $\mathbb{\Pi}^m$ in place of $\mathcal{Q}_{\mathrm{CVaR}(\alpha,\boldsymbol{p})}$ (see, e.g., Artzner *et al.*, 1999).

10.1.3 Minimization of CVaR

Now, let us consider the case where the loss is defined with one or more parameters, and find the parameter values that minimize CVaR.

Let the loss $\tilde{L}$ be parametrized with $\boldsymbol{\theta} \in \mathbb{R}^n$ (i.e., $\tilde{L}(\boldsymbol{\theta})$), and suppose that the probability is independent of $\boldsymbol{\theta}$. Rockafellar and Uryasev (2002) prove the following theorem.

Theorem 10.2 (CVaR minimization) *Let* $G(\boldsymbol{\theta}, c) := c + \frac{1}{1-\alpha}\mathbb{E}_{\mathbb{P}}[\max\{\tilde{L}(\boldsymbol{\theta}) - c, 0\}]$*, and let* $\Theta \subset \mathbb{R}^n$ *denote the set of admissible* $\boldsymbol{\theta}$*, and* $(\boldsymbol{\theta}^\star, c^\star) \in \arg\min_{\boldsymbol{\theta}\in\Theta, c} G(\boldsymbol{\theta}, c)$*. Then,*

1. $\min_{\boldsymbol{\theta}\in\Theta} CVaR_{(\alpha,\mathbb{P})}[\tilde{L}(\boldsymbol{\theta})] = G(\boldsymbol{\theta}^\star, c^\star) = \min_{\boldsymbol{\theta}\in\Theta, c} G(\boldsymbol{\theta}, c)$.
2. $\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta}\in\Theta} CVaR_{(\alpha,\mathbb{P})}[\tilde{L}(\boldsymbol{\theta})]$ *and* $c^\star \in [VaR_{(a,\mathbb{P})}[\tilde{L}(\boldsymbol{\theta}^\star)], VaR^{+}_{(a,\mathbb{P})}[\tilde{L}(\boldsymbol{\theta}^\star)]]$.
3. *Furthermore, if* $\tilde{L}(\boldsymbol{\theta})$ *is convex with respect to* $\boldsymbol{\theta}$*, then so are both* $CVaR_{(\alpha,\mathbb{P})}[\tilde{L}(\boldsymbol{\theta})]$ *and* $G(\boldsymbol{\theta}, c)$.
4. *If* $\tilde{L}(\boldsymbol{\theta})$ *is homogeneous with respect to* $\boldsymbol{\theta}$ *(i.e., for any* $a \in \mathbb{R}$*,* $\tilde{L}(a\boldsymbol{\theta}) = a\tilde{L}(\boldsymbol{\theta})$*), then both* $CVaR_{(\alpha,\mathbb{P})}[\tilde{L}(\boldsymbol{\theta})]$ *and* $G(\boldsymbol{\theta}, c)$ *are positively homogeneous with respect to* $(\boldsymbol{\theta}, c)$.

The first property states that the minimization of CVaR reduces to simultaneous minimization of the function $G(\boldsymbol{\theta}, c)$ in $\boldsymbol{\theta}$ and c. The second statement guarantees that the interpretation of the variable c as an approximate VaR (i.e., α-quantile of L) remains valid even in the case of CVaR minimization. The third property states that the associated CVaR minimization is a convex minimization if $\tilde{L}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. The fourth property, which is not exactly stated in Rockafellar and Uryasev (2002), is that the (positive) homogeneity of the loss propagates to that of CVaR in terms of the involved parameters. As will be discussed in Section 10.3, this property plays a role in analyzing the form of a regularized empirical CVaR minimization.

Example 10.3 (CVaR-minimizing portfolio selection) *Let* $\tilde{R}_j$ *denote a random rate of return of an investable asset j, and suppose that* $\tilde{\boldsymbol{R}} := (\tilde{R}_1, \ldots, \tilde{R}_n)$ *follows a discrete distribution satisfying*

$$(\tilde{R}_1, \ldots, \tilde{R}_n)(\omega_i) = (R_{i,1}, \ldots, R_{i,n}) \quad \text{and} \quad p_i := \mathbb{P}\{\omega = \omega_i\} > 0, i = 1, \ldots, m.$$

Let θ_j *denote the investment ratio of asset j. To make the investment self-financing, we impose a constraint* $\sum_{j=1}^n \theta_j = 1$*. In addition, to meet the investor's requirements, several constraints are imposed on* $\boldsymbol{\theta}$*. We will impose, for example, a restriction of the form* $\mathbf{0} \le \boldsymbol{\theta} \le \boldsymbol{u}$ *with upper bounds* $\boldsymbol{u} \in \mathbb{R}^n$ *and* $\boldsymbol{r}^\top\boldsymbol{\theta} \ge \tau$ *using the expected return* $\boldsymbol{r} = \mathbb{E}_p[\tilde{\boldsymbol{R}}]$ *and the minimum target return* $\tau > 0$*. The problem of determining a portfolio* $(\theta_1, \ldots, \theta_n)$ *that minimizes CVaR defined with* $\tilde{L}(\boldsymbol{\theta}) = -\tilde{\boldsymbol{R}}^\top\boldsymbol{\theta}$ *(or, equivalently,* $\boldsymbol{L}(\boldsymbol{\theta}) = -(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_m)^\top\boldsymbol{\theta}$ *with* $\boldsymbol{R}_i = (R_{i,1}, \ldots, R_{i,n})^\top$*) is an LP:*

$$\left|\begin{array}{ll} \underset{\boldsymbol{\theta}, c, \boldsymbol{z}}{\text{minimize}} & c + \frac{1}{1-\alpha}\sum_{i=1}^m p_i z_i \\ \text{subject to} & z_i \ge -\sum_{j=1}^n R_{ij}\theta_j - c, i = 1, \ldots, m, \\ & z_i \ge 0, i = 1, \ldots, m, \\ & \sum_{j=1}^n \theta_j = 1, \quad \sum_{j=1}^n r_j\theta_j \ge \tau, \quad 0 \le \theta_j \le u_j, j = 1, \ldots, n. \end{array}\right.$$

Typically, a discrete distribution is obtained from historical observations, for example, periodic asset returns (e.g., daily, weekly, or monthly) in real markets, and $p_i = 1/m$ is used unless there is particular information about $\boldsymbol{p}$.[12]

Example 10.4 (CVaR-based passive portfolio selection) *Strategies seeking to mimic market indexes such as the S&P 500 (i.e., a certain average of asset prices) are popular in portfolio management.*[13] *Let $\tilde{I}$ denote the return of an index, and assume that $(\tilde{R}_1, \ldots, \tilde{R}_n, \tilde{I})$ follows a discrete distribution satisfying $(\tilde{R}_1, \ldots, \tilde{R}_n, \tilde{I})(\omega_i) = (R_{i,1}, \ldots, R_{i,n}, I_i)$ and $p_i := \mathbb{P}\{\omega = \omega_i\} > 0, i = 1, \ldots, m$. Measure the deviation of the portfolio return $\tilde{\boldsymbol{R}}^\top \boldsymbol{\theta}$ from the benchmark return $\tilde{I}$ by using the CVaR associated with the loss $\tilde{L}(\boldsymbol{\theta}) = |\tilde{I} - \tilde{\boldsymbol{R}}^\top \boldsymbol{\theta}|$ (or, equivalently, $\boldsymbol{L}(\boldsymbol{\theta}) = (|I_1 - \boldsymbol{R}_1^\top \boldsymbol{\theta}|, \ldots, |I_m - \boldsymbol{R}_m^\top \boldsymbol{\theta}|)^\top$ with $\boldsymbol{R}_i = (R_{i,1}, \ldots, R_{i,n})^\top$). The problem of finding a portfolio mimicking the index can then be formulated as*

$$\left|\begin{array}{ll} \underset{\boldsymbol{\theta},c,\boldsymbol{z}}{\text{minimize}} & c + \frac{1}{1-\alpha} \sum_{i=1}^m p_i z_i \\ \text{subject to} & z_i \geq I_i - \sum_{j=1}^n R_{ij}\theta_j - c, i = 1, \ldots, m, \\ & z_i \geq -I_i + \sum_{j=1}^n R_{ij}\theta_j - c, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \\ & \sum_{j=1}^n \theta_j = 1, \quad \sum_{j=1}^n r_j \theta_j \geq \tau, \quad 0 \leq \theta_j \leq u_j, j = 1, \ldots, n. \end{array}\right.$$

10.2 Support Vector Machines

SVMs are one of the most successful supervised learning methods that can be applied to classification or regression. This section introduces several SVM formulations, whose relation to CVaR minimization will be discussed in the succeeding sections.

10.2.1 Classification

Suppose that we have m samples $(\boldsymbol{x}_i, y_i), i = 1, \ldots, m$, where $\boldsymbol{x}_i := (x_{i1}, \ldots, x_{in})^\top \in \mathbb{R}^n$ denotes the vector of the attributes of sample i and $y_i \in \{-1, +1\}$ denotes its binary label, $i = 1, \ldots, m$. SVM classification (or SVC, for short) finds a hyperplane, $\boldsymbol{w}^\top \boldsymbol{x} = b$, that separates the training samples as much as possible. The labels of the new (unknown) samples can be predicted on the basis of which side of the hyperplane they fall on.

By using the so-called kernel trick, SVC constructs a nonlinear classifier, a hyperplane in a high (possibly, infinite) dimensional space. Namely, it implicitly uses a mapping $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^N$ (N can be infinite) and obtains a hyperplane $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) = b$ and a decision function $d(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) - b)$, where $\text{sign}(z)$ is 1 if $z \geq 0$ and -1 otherwise.

[12] If $(\tilde{R}_1, \ldots, \tilde{R}_n)$ follows a multivariate normal distribution $N(\boldsymbol{r}, \boldsymbol{\Sigma})$, the portfolio return (i.e., $\tilde{\boldsymbol{R}}^\top \boldsymbol{\theta}$), follows a normal distribution $N(\boldsymbol{r}^\top \boldsymbol{\theta}, \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta})$. With the loss $\tilde{L}(\boldsymbol{\theta}) = -\tilde{\boldsymbol{R}}^\top \boldsymbol{\theta}$, the CVaR minimization reduces to a second-order cone program, that is, $\min_{\boldsymbol{\theta} \in \Theta} -\boldsymbol{r}^\top \boldsymbol{\theta} + C_\alpha \sqrt{\boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta}}$ with $C_\alpha := \exp\left(-\{\Psi^{-1}(\alpha)\}^2/2\right)/\{(1-\alpha)\sqrt{2\pi}\}$ (see formula (10.4)). This is equivalent to the so-called mean–variance criterion (Markowitz, 1952) for a specific trade-off parameter.

[13] This type of investment strategy is called *passive*, while those seeking to beat the market average are called *active*.

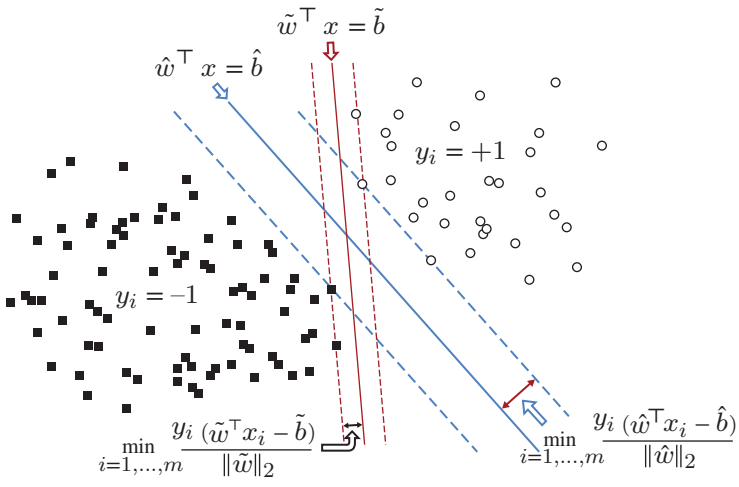

Figure 10.4 Two separating hyperplanes and their geometric margins. The dataset is said to be *linearly separable* if there exist $\boldsymbol{w} \neq \boldsymbol{0}$ and b such that $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i - b) > 0$ for all $i = 1, \ldots, m$. If the dataset is linearly separable, there are infinitely many hyperplanes separating the dataset. According to generalization theory (Vapnik, 1995), the hyperplane $\hat{\boldsymbol{w}}^\top \boldsymbol{x} = \hat{b}$ is preferable to $\tilde{\boldsymbol{w}}^\top \boldsymbol{x} = \tilde{b}$. The optimization problem (10.12) (or, equivalently, (10.13)) finds a hyperplane that separates the datasets with the largest margin.

The Vapnik–Chervonenkis theory shows that a large geometric margin classifier has a small generalization error (Vapnik, 1995). Namely, the search for a hyperplane that has the largest distance to the nearest data points decreases the upper bound of the out-of-sample error. Motivated by this theoretical result, Boser *et al.* (1992) developed an algorithm for finding a hyperplane $(\boldsymbol{w}, b)$ with the maximum geometric margin, which is formulated as

$$\underset{\boldsymbol{w},b}{\text{maximize}} \min_{i=1,\ldots,m} \frac{y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b)}{\|\boldsymbol{w}\|_2} = -\underset{\boldsymbol{w},b}{\text{minimize}} \max_{i=1,\ldots,m} \frac{-y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b)}{\|\boldsymbol{w}\|_2}, \tag{10.12}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm (or the Euclidean norm), that is, $\|\boldsymbol{w}\|_2 := \sqrt{\boldsymbol{w}^\top \boldsymbol{w}}$.

If the data samples are *linearly separable*, that is, there exists a hyperplane that separates the samples $\boldsymbol{x}_i$ such that $y_i = +1$ from those such that $y_i = -1$, as in Figure 10.4, the fractional optimization (10.12) can be rewritten as the following quadratic program:

$$\begin{vmatrix} \underset{\boldsymbol{w},b}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} & y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b) \geq 1,\ i = 1, \cdots, m. \end{vmatrix} \tag{10.13}$$

This is called *hard-margin SVC*. Note that (10.13) is valid only when the training samples are linearly separable.

10.2.1.1 *C*-Support Vector Classification

Cortes and Vapnik (1995) extend the SVC algorithm to linearly nonseparable cases and trade off the margin size with the data separation error. More precisely, by introducing slack variables

$z_1, \ldots, z_m$ and adding their sum to the objective function, the hard-margin SVC formulation can be modified into

$$f_{\text{CSVC}} := \begin{vmatrix} \underset{\boldsymbol{w},b,\boldsymbol{z}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^m z_i, \\ \text{subject to} & z_i \geq -y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b) + 1, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \end{vmatrix} \tag{10.14}$$

where $C > 0$ is a user-defined parameter. Formulation (10.14) is often viewed as a correction that adds the so-called *hinge loss* $\sum_{i=1}^m \max\{L_i + 1, 0\}$ with $L_i = -y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b)$, as a surrogate of the *0–1 loss* $\sum_{i=1}^m 1_{\{L_i \geq 0\}}$, which would otherwise involve nonconvexity.[14] Formulation (10.14) is usually referred to as *C-SVC*,[15] and it has been shown to work very well in various real-world applications (see, e.g., Schölkopf and Smola, 2002).

The (Lagrangian) dual formulation of (10.14) is derived as

$$f_{\text{CSVC}} = \begin{vmatrix} \underset{\lambda}{\text{maximize}} & \sum_{i=1}^m \lambda_i - \frac{1}{2}\sum_{h=1}^m \sum_{i=1}^m \lambda_i \lambda_h y_i y_h k(\boldsymbol{x}_i, \boldsymbol{x}_h) \\ \text{subject to} & \sum_{i=1}^m y_i \lambda_i = 0, \quad 0 \leq \lambda_i \leq C, i = 1, \ldots, m. \end{vmatrix} \tag{10.15}$$

Here, $k(\boldsymbol{x}_i, \boldsymbol{x}_h) = \boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_h)$ is a *kernel function* defined directly on the inputs of $\boldsymbol{x}_i$ and $\boldsymbol{x}_h$. The use of a kernel function is preferable to that of the explicit mapping $\boldsymbol{\phi}(\cdot)$, because we can treat a highly nonlinear mapping without bothering about how large the dimension N of the mapped space should be.

Moreover, by using the optimality condition, we can recover a dual solution from a primal solution.[16] With an optimal solution $(\boldsymbol{w}^\star, b^\star, \boldsymbol{z}^\star, \lambda^\star)$, the decision function is given by

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

10.2.1.2 ν-Support Vector Classification

ν-SVC is another formulation of soft-margin SVC (Schölkopf *et al.*, 2000),

$$f_{\nu\text{-SVC}} := \begin{vmatrix} \underset{\boldsymbol{w},b,\rho,\boldsymbol{z}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^m z_i, \\ \text{subject to} & z_i \geq -y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b) + \rho, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \end{vmatrix} \tag{10.16}$$

where $\nu \in (0, 1]$ is a user-defined parameter. The dual of (10.16) is described as

$$f_{\nu\text{-SVC}} = \begin{vmatrix} \underset{\lambda}{\text{maximize}} & -\frac{1}{2}\sum_{h=1}^m \sum_{i=1}^m \lambda_i \lambda_h y_i y_h k(\boldsymbol{x}_i, \boldsymbol{x}_h) \\ \text{subject to} & \sum_{i=1}^m \lambda_i = 1, \sum_{i=1}^m y_i \lambda_i = 0, \\ & 0 \leq \lambda_i \leq \frac{1}{\nu m}, i = 1, \ldots, m. \end{vmatrix} \tag{10.17}$$

[14] More precisely, hinge loss is viewed as a special case of the convex upper bound, $\max\{(L - C)/(B - C), 0\}$ with $B = 0$ and $C = -1$, of the 0–1 loss, as shown in Figure 10.2.

[15] To make a contrast with (10.13), formulations of this type are sometimes referred to as *soft-margin SVCs*.

[16] Strong duality also holds (i.e., the optimal values of (10.14) and (10.15) approach the same f_{CSVC}).

Formulation (10.17) indicates that the optimal value of (10.16) as well as (10.17) is non-increasing with respect to ν. Moreover, the optimal value is nonpositive (or unbounded) because $(\boldsymbol{w}, b, \rho, \boldsymbol{z}) = \mathbf{0}$ is feasible for (10.16).

Note that (10.17) is not necessarily well defined for any ν between 0 and 1 (Chang and Lin, 2001; Crisp and Burges, 2000). Let m_+ (resp. m_-) denote the number of samples with positive (resp. negative) labels. When ν is larger than $\nu_{\max} := 2\min\{m_+, m_-\}/m$, we can show that the primal ν-SVC (10.16) is unbounded and the dual ν-SVC (10.17) becomes infeasible. On the other hand, when ν is smaller than some threshold $\nu_{\min}$, ν-SVC produces a trivial solution satisfying $(\boldsymbol{w}, b) = \mathbf{0}$ (Chang and Lin, 2001). The lower threshold $\nu_{\min}$ is defined as the smallest upper bound of ν with which the optimal value of ν-SVC becomes zero.[17]

Schölkopf *et al.*, (2000) show that the relation between ν-SVC and C-SVC is as follows.

Theorem 10.3 (Schölkopf *et al.* 2000) *Suppose that (10.16) has an optimal solution $(\boldsymbol{w}, b, \rho, \boldsymbol{z})$ with $\rho > 0$. Then (10.14) with $C = 1/(\rho m)$ provides the same decision function as (10.16) does.*

Crisp and Burges (2000) show that an optimal solution of ν-SVC (10.16) satisfies $\rho \geq 0$. In the above sense, ν-SVC and C-SVC are equivalent except for the case of $\rho = 0$.

10.2.1.3 Extended ν-Support Vector Classification

Recall that for $\nu \in (0, \nu_{\min})$, ν-SVC produces a trivial solution satisfying $(\boldsymbol{w}, b) = \mathbf{0}$. To prevent this, Perez-Cruz *et al.* (2003) require the norm of $\boldsymbol{w}$ to be unity:

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},b,\rho,\boldsymbol{z}}{\text{minimize}} & -\rho + \frac{1}{\nu m}\sum_{i=1}^{m} z_i, \\ \text{subject to} & z_i \geq -y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b) + \rho, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \\ & \|\boldsymbol{w}\|_2 = 1. \end{array}\right. \tag{10.18}$$

As a result of this modification, a nontrivial solution can be obtained even for $\nu \in (0, \nu_{\min})$. This modified formulation is called *extended ν-SVC* (*Eν-SVC*).

Problem (10.18) is nonconvex because of the equality-norm constraint $\|\boldsymbol{w}\|_2 = 1$.[18] For $\nu \in [\nu_{\min}, \nu_{\max}]$, E$\nu$-SVC has the same optimal solutions as ν-SVC does and can be reduced to ν-SVC. Perez-Cruz *et al.* (2003) experimentally show that the out-of-sample performance of Eν-SVC with $\nu \in (0, \nu_{\min}]$ is often better than that with $\nu \in (\nu_{\min}, \nu_{\max}]$.

10.2.1.4 One-class ν-Support Vector Classifications

Next, we will consider a problem that has been referred to as *outlier/novelty detection*, *high-density region estimation*, or *domain description*.[19]

[17] ν-SVC with $\nu = \nu_{\min}$ may result in a nontrivial solution, whereas ν-SVC with $\nu \in (0, \nu_{\min})$ always results in the trivial solution. The computation of $\nu_{\min}$ will be discussed in Section 10.4.

[18] Perez-Cruz *et al.* (2003) propose an iterative algorithm for computing a solution. It goes as follows. First, for some $\bar{\boldsymbol{w}}$ satisfying $\|\bar{\boldsymbol{w}}\|_2^2 = 1$, define an LP by replacing $\|\boldsymbol{w}\|_2^2 = 1$ by $\bar{\boldsymbol{w}}^\top \boldsymbol{w} = 1$, and solve it. Then, use the obtained solution $\hat{\boldsymbol{w}}$ to update $\bar{\boldsymbol{w}}$, and repeat this procedure until convergence.

[19] This class of problems is sometimes referred to as one-class classification or, more broadly, unsupervised learning.

Let $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \subset \mathbb{R}^n$ be a given dataset. The one-class problem is to define (possible) outliers in X. An outlier detection model known as a *one-class ν-support vector machine* (Schölkopf and Smola, 2002) is formulated as

$$\begin{cases} \underset{\boldsymbol{w},\rho,z}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^m z_i, \\ \text{subject to} & z_i \geq -\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) + \rho, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \end{cases} \tag{10.19}$$

where $\nu \in (0, 1]$ is a user-defined parameter. With an optimal solution $(\boldsymbol{w}^\star, \rho^\star, z^\star)$, a sample $\boldsymbol{x}$ satisfying $\boldsymbol{w}^{\star\top}\boldsymbol{\phi}(\boldsymbol{x}) < \rho^\star$ is regarded as an "outlier." Or, equivalently, we can define a high-density region to be the set $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{w}^{\star\top}\boldsymbol{\phi}(\boldsymbol{x}) \geq \rho^\star\}$. We will see in Section 10.3.1 that the (10.19) formulation can be interpreted on the basis of CVaR.

Support vector domain description (*SVDD*) (Tax and Duin, 1999) is a variant of the one-class problem. It detects (possible) outliers on the basis of the quadratically constrained optimization problem,

$$\begin{cases} \underset{\boldsymbol{\gamma},R,z}{\text{minimize}} & R^2 + C\sum_{i=1}^m z_i, \\ \text{subject to} & z_i \geq \|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\gamma}\|_2^2 - R^2, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \end{cases} \tag{10.20}$$

where $C > 0$ is a user-defined parameter. SVDD defines outliers as points $\boldsymbol{x}_i$ satisfying $\|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\gamma}^\star\|_2 > R^\star$ by using an optimal solution $(\boldsymbol{\gamma}^\star, R^\star, z^\star)$ of (10.20). Since the high-density region of $\boldsymbol{x}$ is defined as $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\gamma}^\star\|_2 \leq R^\star\}$, this type of problem is called *high-density region estimation* or *domain description*. The high-density region is compact even when $\boldsymbol{\phi}$ is a linear mapping, whereas one-class ν-SVM (10.19) is not compact then.

10.2.2 Regression

Following its success in classification, SVC was extended so that it could handle real-valued outputs (Drucker *et al.*, 1997), (Schölkopf *et al.*, 2000). The *support vector regression* (*SVR*) method performs well in regression analysis and is a popular data analysis tool in machine learning and signal processing.

Let us consider the regression problem of obtaining a model $y = \boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}) + b$ using m training samples, $(\boldsymbol{x}_i, y_i), i = 1, \cdots, m$, where $\boldsymbol{x}_i \in \mathbb{R}^n$ is an input and $y_i \in \mathbb{R}$ is the corresponding output value.

10.2.2.1 ϵ-Support Vector Regression

In the ϵ-SVR framework (Drucker *et al.*, 1997), the model, or equivalently, $(\boldsymbol{w}, b)$, is determined so that the following regularized empirical risk functional is minimized:

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^m \max\{|y_i - \boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_i) - b| - \varepsilon, 0\},$$

where C and ε are positive constants. Among the two parameters, $C > 0$ is a regularization constant that controls the trade-off between the goodness-of-fit and the complexity of the model. The parameter ε controls the sensitivity to the residuals (i.e., $|y_i - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b|$). A potential weakness of the ε-SVR formulation is that the choice of ε is not intuitive.[20]

10.2.2.2 ν-Support Vector Regression

Another formulation of SVR, called *ν-SVR*, was proposed by Schölkopf *et al.* (2000); it uses a parameter $\nu \in (0, 1)$, instead of ε in ε-SVR. The optimization formulation of ν-SVR is

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},b,c,\boldsymbol{z}}{\text{minimize}} & \frac{H}{2}\|\boldsymbol{w}\|_2^2 + c + \frac{1}{\nu m}\sum_{i=1}^m z_i, \\ \text{subject to} & z_i \geq y_i - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b - c, i = 1, \ldots, m, \\ & z_i \geq -y_i + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) + b - c, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \end{array}\right. \tag{10.21}$$

where $H > 0$ is a user-defined constant. By setting $\nu = 1$ and restricting $c = \epsilon$, the (10.21) formulation reduces to the ϵ-SVR formulation.

10.3 ν-SVMs as CVaR Minimizations

In this section, we reformulate several SVMs in terms of CVaR minimization. We classify the CVaR minimizations into two cases: Case 1, where the loss $\boldsymbol{L}(\boldsymbol{\theta})$ is homogeneous with respect to the involving parameters $\boldsymbol{\theta}$ (i.e., for any $a \in \mathbb{R}$, $\boldsymbol{L}(a\boldsymbol{\theta}) = a\boldsymbol{L}(\boldsymbol{\theta})$); and Case 2, where the loss $\boldsymbol{L}(\boldsymbol{\theta})$ is not homogeneous.

10.3.1 ν-SVMs as CVaR Minimizations with Homogeneous Loss

We can formulate various machine-learning methods by using different types of loss. Let us begin with a binary classification problem defined with the positively homogeneous loss.

10.3.1.1 ν-SVC as a CVaR Minimization

Using the notation of CVaR with a linear loss $L_i(\boldsymbol{w}, b) = -y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b)$ and $p_i = 1/m$, $i = 1, \ldots, m$, the quadratic program (10.16) can be symbolically rewritten as

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\cdot\text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)), \tag{10.22}$$

with $C = \nu$, where

$$\boldsymbol{Y} := \text{diag}(\boldsymbol{y}) := \begin{pmatrix} y_1 & & \\ & \ddots & \\ & & y_m \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad \boldsymbol{X} := \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^\top \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_m)^\top \end{pmatrix} \in \mathbb{R}^{m\times N}.$$

[20] The function $\max\{|y_i - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b| - \varepsilon, 0\}$ is called *Vapnik's ε-insensitive loss function*.

Note that (10.22) is a regularized empirical risk minimization in which CVaR is used as the empirical risk (see Figure 10.5). Note also that the empirical risk part in (10.22) is homogeneous in $(\boldsymbol{w}, b)$. To deal with the trade-off between the regularization term and the empirical CVaR, we may be able to perform another form of optimization,

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 \quad \text{subject to} \quad \text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)) \leq -D, \tag{10.23}$$

and[21]

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)) \quad \text{subject to} \quad \|\boldsymbol{w}\|_2 \leq E, \tag{10.24}$$

where C, D, and E are positive parameters for reconciling the trade-off. Under a mild assumption, the above three regularized empirical risk minimizations are equivalent for any positive parameters C, D, and E (see Tsyurmasto *et al.*, 2013).[22] Accordingly, C, D, and E can be restricted to 1. For example, (10.24) can be restricted to

$$\begin{cases} \underset{\boldsymbol{w},b}{\text{minimize}} & \text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)) \\ \text{subject to} & \|\boldsymbol{w}\|_2 \leq 1. \end{cases} \tag{10.25}$$

It is worth emphasizing that the equivalence of (10.22), (10.23), and (10.24) (or (10.25)) relies on the homogeneity of the empirical CVaR with respect to $(\boldsymbol{w}, b)$. Conversely, such an equivalence also holds true for any positively homogeneous risk functionals $\mathcal{F}(\cdot)$ in combination with a homogeneous loss $\boldsymbol{L}(\boldsymbol{\theta})$, that is, $\mathcal{F}(\boldsymbol{L}(\boldsymbol{\theta}))$. On the other hand, with the hinge loss employed in C-SVC (10.14), a risk-constrained variant like (10.23) is infeasible for any $D > 0$, and a parallel equivalence is no longer valid.

10.3.1.2 Eν-SVC as the Geometric Margin-based CVaR Minimization

Eν-SVC (10.18) can be symbolically rewritten as

$$\begin{cases} \underset{\boldsymbol{w},b}{\text{minimize}} & \text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)) \\ \text{subject to} & \|\boldsymbol{w}\|_2 = 1. \end{cases} \tag{10.26}$$

Comparing (10.25) and (10.26), we can see that ν-SVC (10.25) is a convex relaxation of Eν-SVC (10.26).

Note that when ν is in the range $(\nu_{\min}, \nu_{\max})$, the optimal value of (10.25) (i.e., the minimum CVaR) is negative, and $\|\boldsymbol{w}\|_2 = 1$ is attained at optimality because of the homogeneity of the objective function. In other words, when $\nu > \nu_{\min}$, the equality constraint $\|\boldsymbol{w}\|_2 = 1$ in

[21] Based on the discussion at the end of Section 10.1, the CVaR-constraint in (10.23) can be regarded as a convex conservative approximation of the chance constraint $\mathbb{P}\{\tilde{L}(\boldsymbol{w}, b) \geq -D\} \leq \nu$.

[22] This equivalence holds in the sense that these models provide the same optimal decision functions $d(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) - b)$ for any C, D, and E. Schölkopf *et al.* (2000) use the optimality condition to show the independence of the resulting classifiers of the parameter C. On the other hand, Tsyurmasto *et al.* (2013) show equivalence only on the basis of functional properties of CVaR such as positive homogeneity and continuity.

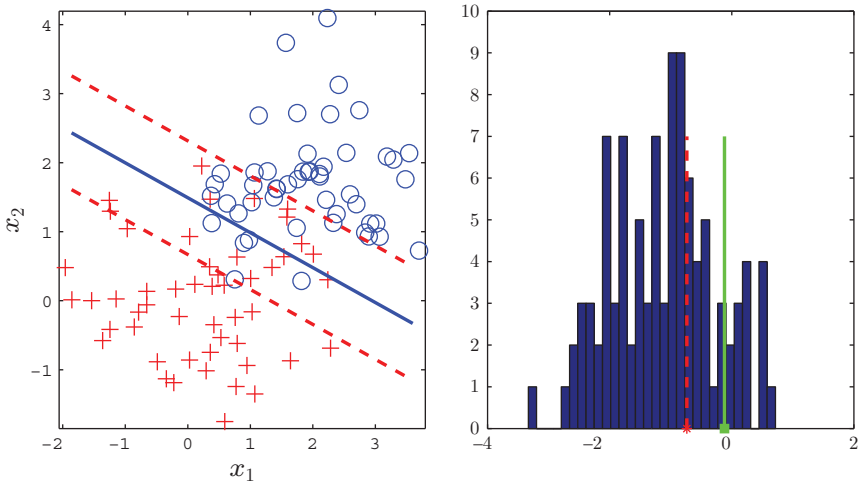

Figure 10.5 ν-SVC as a CVaR minimization. The figure on the left shows an optimal separating hyperplane $w_1^\star x_1 + w_2^\star x_2 = b^\star$ given by ν-SVC ($\nu = 0.3$). The one on the right is a histogram of the optimal distribution of the negative margin, $-y_i(w_1^\star x_{1i} + w_2^\star x_{2i} - b^\star), i = 1, \ldots, 100$. The locations of the minimized CVaR (solid line) and the corresponding VaR (broken line) are indicated in the histogram.

(10.26) can be relaxed to $\|\boldsymbol{w}\|_2 \leq 1$ without changing the optimal solution. On the other hand, when $\nu < \nu_{\min}$, ν-SVC (10.25) results in a trivial solution satisfying $(\boldsymbol{w}, b) = \boldsymbol{0}$.[23] Therefore, to obtain a solution to Eν-SVM (10.26) for $\nu < \nu_{\min}$, a nonconvex optimization method needs to be applied (Gotoh and Takeda, 2005; Perez-Cruz *et al.*, 2003; Takeda and Sugiyama, 2008). Figure 10.6 illustrates the relation between the sign of the optimal value of Eν-SVC (10.26) and ν.

Note that (10.18) can be equivalently rewritten as

$$\underset{\boldsymbol{w},b,c}{\text{minimize}} \quad c + \frac{1}{\nu m}\sum_{i=1}^{m} \max\left\{\frac{-y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b)}{\|\boldsymbol{w}\|_2} - c, 0\right\}, \tag{10.27}$$

(Takeda and Sugiyama, 2008). Namely, Eν-SVC (10.18) can be described as another CVaR minimization problem,

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \text{CVaR}_{(1-\nu,\boldsymbol{1}_m/m)}\left(-\frac{\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}_m b)}{\|\boldsymbol{w}\|_2}\right),$$

by adopting the negative geometric margin as the loss, that is, $L_i(\boldsymbol{w}, b) = -y_i(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b)/\|\boldsymbol{w}\|_2, i = 1, \ldots, m$. Since CVaR includes the maximum loss as a special limiting case (see Section 10.1.1), formulation (10.27) is a generalization of the maximum margin formulation (10.12). Figure 10.7 summarizes the relations among the four CVaR minimizations.

[23] In this case, $\|\boldsymbol{w}\|_2 = 1$ of Eν-SVC (10.26) can be relaxed to $\|\boldsymbol{w}\|^2 \geq 1$, but the resulting optimization problem is still nonconvex.

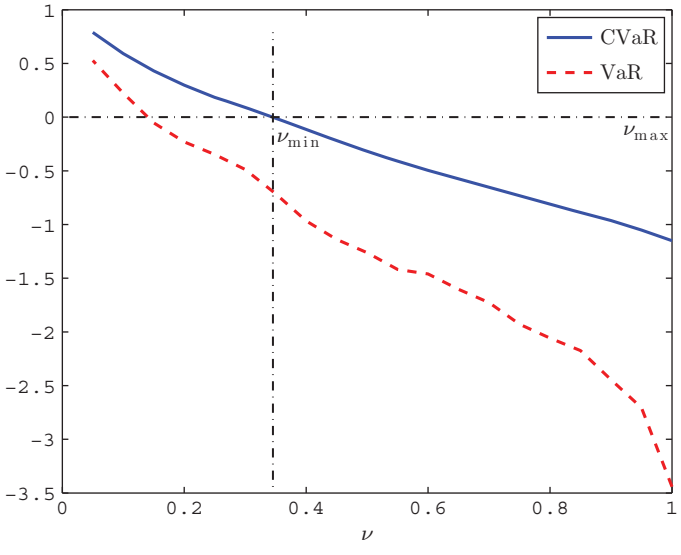

Figure 10.6 Minimized CVaR and corresponding VaR with respect to ν. CVaR indicates the optimal value of Eν-SVC (10.26) for binary classification. $\nu_{\min}$ is the value of ν at which the optimal value becomes zero. For $\nu > \nu_{\min}$, Eν-SVC (10.26) reduces to ν-SVC (10.25). For $\nu < \nu_{\min}$, ν-SVC (10.25) results in a trivial solution, while Eν-SVC (10.26) still attains a nontrivial solution with the positive optimal value.

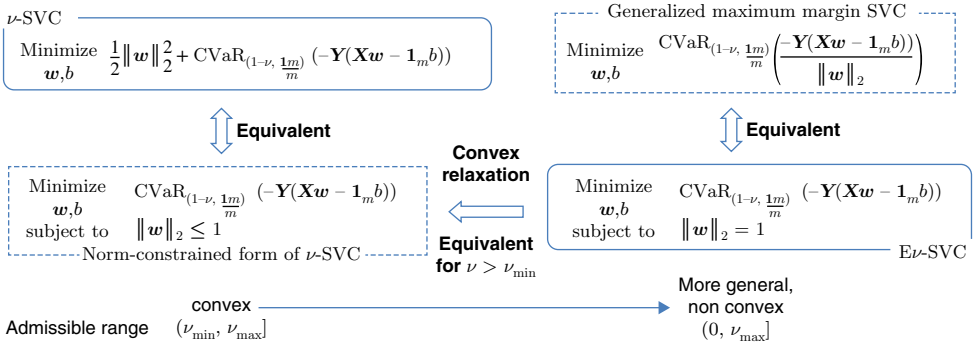

Figure 10.7 Relations among four classification formulations. The two formulations on the left are equivalent to the standard ν-SVC (10.16), while those on the right are equivalent to Eν-SVC (10.18). By resolving the nonconvexity issues that arise from the equality constraint, Eν-SVC provides a classifier that cannot be attained by ν-SVC.

10.3.1.3 One-class ν-SVC as a CVaR Minimization

With $L_i(\boldsymbol{w}) = -\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i)$ and regularization term $\frac{1}{2}\|\boldsymbol{w}\|_2^2$, we obtain a CVaR minimization,

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(-\boldsymbol{X}\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{w}\|_2^2, \tag{10.28}$$

which is equivalent to one-class ν-SVC (10.19).[24] Namely, outliers found by (10.19) are viewed as points $\boldsymbol{x}_i$ having values $-\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i)$ greater than the 100ν-percentile under the CVaR minimizer $\boldsymbol{w}$.

10.3.2 ν-SVMs as CVaR Minimizations with Nonhomogeneous Loss

Next, we consider cases where the loss is not homogeneous with respect to the involved parameters.

10.3.2.1 CVaR-based Regression and ν-SVR

Assuming a regression model $y = b + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$, we can employ a loss of the form $L_i(b, \boldsymbol{w}) = \epsilon(y_i - \{b + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i)\})$, where $\epsilon : \mathbb{R} \to [0, +\infty]$. Note that this is no longer homogeneous in $(b, \boldsymbol{w})$. By using this loss and $\boldsymbol{p}$, we can readily attain a regression version of CVaR minimization:

$$\underset{b,\boldsymbol{w}}{\text{minimize}} \quad \text{CVaR}_{(1-\nu,\boldsymbol{p})}(\epsilon(y_1 - \{b + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_1)\}), \ldots, \epsilon(y_m - \{b + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_m)\})). \tag{10.29}$$

In particular, when $\nu = 1$, (10.29) includes a number of popular regression formulations. For example:

- With $\boldsymbol{p} = \mathbf{1}_m/m$ and $\epsilon(z) = z^2$, (10.29) is equivalent to *ordinary least squares* (*OLS*).
- With arbitrary $\boldsymbol{p} \in \Pi^m$ and $\epsilon(z) = z^2$, it is equivalent to a weighted least square with a weight vector $\boldsymbol{p}$.
- With $\boldsymbol{p} = \mathbf{1}_m/m$ and $\epsilon(z) = az$ for $z \geq 0$ and $\epsilon(z) = -(1-a)z$ for $z < 0$ with some $a \in (0, 1)$, (10.29) is equivalent to *quantile regression* (Koenker and Bassett, 1978).

By adding a certain regularizer, we can attain more generalized formulations:

$$\underset{b,\boldsymbol{w}}{\text{minimize}} \quad \text{“Objective of (10.29)”} + Cg(\|\boldsymbol{w}\|), \tag{10.30}$$

where $C \geq 0$ is a constant, $g : [0, \infty) \to (-\infty, +\infty]$ a nondecreasing function, and $\|\cdot\|$ a norm. For example:

- With $\nu = 1, \boldsymbol{p} = \mathbf{1}_m/m$, $\epsilon(z) = z^2$, and $g(\|\boldsymbol{w}\|) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, it is equivalent to *ridge regression* (Hoerl and Kennard, 1970).

[24] Since the loss $L_i(\boldsymbol{w}) = -\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i)$ is positively homogeneous with respect to $\boldsymbol{w}$, we can show that (10.19) is equivalent to $\min\{\frac{1}{2}\|\boldsymbol{w}\|_2^2 : \text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w})) \leq -1\}$ and $\min\{\text{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w})) : \|\boldsymbol{w}\|_2 \leq 1\}$ in the sense that all of them provide the same decision function, as in two-class ν-SVC.

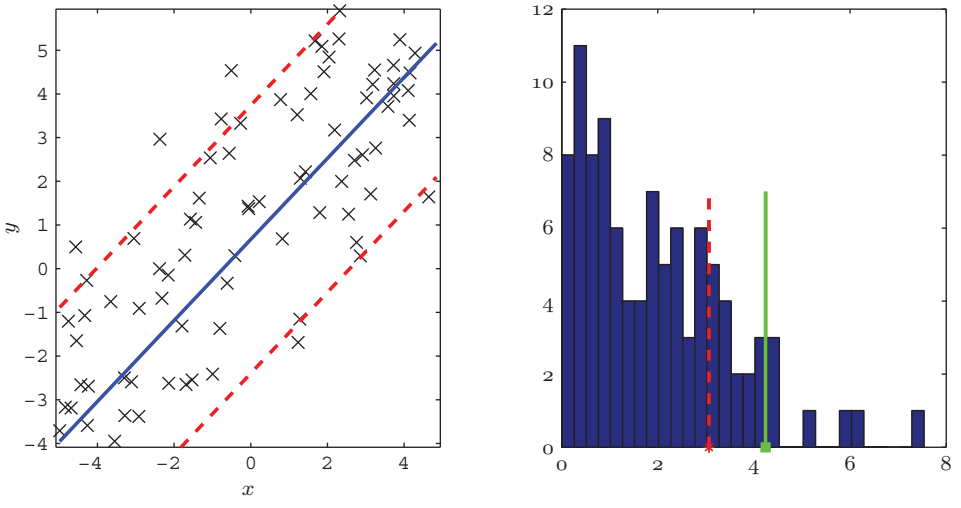

Figure 10.8 ν-SVR as a CVaR minimization. The left figure shows the regression model $y = w^\star x + b^\star$ given by ν-SVR ($\nu = 0.2$). The right one shows the histogram of the optimal distribution of the residual $|y_i - w^\star x_i - b^\star|, i = 1, \ldots, 100$. The locations of the minimized CVaR (green solid line) and the corresponding VaR (red dashed line) are indicated in the histogram.

- With $\nu = 1$, $\boldsymbol{p} = \mathbf{1}_m/m$, $\epsilon(z) = z^2$, and $g(\|\boldsymbol{w}\|) = \|\boldsymbol{w}\|_1$, it is equivalent to *lasso regression* (Tibshirani, 1996).

See Chapter 11 written by Uryasev for the definitions and a general look at ridge regression and lasso.

If $\nu < 1$, (10.30) is related to ν-SVR. Indeed, with $\boldsymbol{p} = \mathbf{1}_m/m$, $\epsilon(z) = |z|$, and $g(\|\boldsymbol{w}\|) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, it is equivalent to ν-SVR (10.21). Figure 10.8 shows the results of ν-SVR and the distribution of the residual.

Different from formulations (10.16) and (10.19) for the classification problem, formulation (10.21) depends on a trade-off parameter H in addition to ν. Indeed, the value of H changes the decision function $d(x)$ of (10.21). (10.21) is also different from the norm-constrained formulation:

$$\left|\begin{array}{ll} \underset{\boldsymbol{w},b,c,\boldsymbol{z}}{\text{minimize}} & c + \frac{1}{\nu m}\sum_{i=1}^m z_i, \\ \text{subject to} & z_i \geq y_i - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) - b - c, i = 1, \ldots, m, \\ & z_i \geq -y_i + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) + b - c, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m, \\ & \|\boldsymbol{w}\|_2 \leq E, \end{array}\right. \tag{10.31}$$

unless E and H are appropriately set. This dependence on the parameters is due to the lack of positive homogeneity of the loss $\boldsymbol{L}(b, \boldsymbol{w})$. On the other hand, the CVaR minimization (10.29) has an optimal solution for any $\nu \in (0, 1)$, unlike ν-SVC.

10.3.2.2 Domain Description Problems as CVaR Minimizations

If we use a loss of the form $L_i(\boldsymbol{\gamma}) = g(\|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\gamma}\|)$ with a nondecreasing convex function g defined over $(0, \infty)$ and a norm $\|\cdot\|$, we arrive at another CVaR minimization,

$$\underset{\boldsymbol{\gamma}}{\text{minimize}} \quad \text{CVaR}_{(1-\nu,\boldsymbol{p})}(g(\|\boldsymbol{\phi}(\boldsymbol{x}_1) - \boldsymbol{\gamma}\|), \ldots, g(\|\boldsymbol{\phi}(\boldsymbol{x}_m) - \boldsymbol{\gamma}\|)),$$

which can be explicitly rewritten as a convex optimization problem,

$$\begin{cases} \underset{\boldsymbol{\gamma},c,\boldsymbol{z}}{\text{minimize}} & c + \frac{1}{\nu}\sum_{i=1}^m p_i z_i, \\ \text{subject to} & z_i \geq g(\|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\gamma}\|) - c, i = 1, \ldots, m, \\ & z_i \geq 0, i = 1, \ldots, m. \end{cases} \tag{10.32}$$

In particular, when we employ $g(\|\cdot\|) = \|\cdot\|_2^2$, $C = 1/\nu$, and $\boldsymbol{p} = \mathbf{1}_m/m$, it is equivalent to SVDD (10.20). Namely, SVDD minimizes the CVaR of the distribution of the squared Euclidean distance $\|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\gamma}\|^2$ from a center $\boldsymbol{\gamma}$.

Formulation (10.32) can be considered to be a generalized version of the so-called *minimum enclosing ball*. Indeed, let us suppose that g is an increasing function. When $\nu < \min_i p_i$, formulation (10.32) becomes the optimization for obtaining a minimum ball enclosing the set of m points $\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_m)$:

$$\begin{cases} \underset{\boldsymbol{\gamma},r}{\text{minimize}} & r, \\ \text{subject to} & r \geq \|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\gamma}\|, i = 1, \ldots, m. \end{cases}$$

On the other hand, when $\nu = 1$, (10.32) becomes

$$\underset{\boldsymbol{\gamma}}{\text{minimize}} \sum_{i=1}^m p_i g(\|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\gamma}\|),$$

which is known to characterize various centers of points $\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_m)$ depending on the norm $\|\cdot\|$ employed.[25]

10.3.3 Refining the ν-Property

So far, we have shown that ν-SVMs can be viewed as CVaR minimizations, each being associated with a certain loss function. This fact enables us to look at SVMs on the basis of the distribution of the loss, $(L_1, \ldots, L_m)$, as described in Section 10.1.

10.3.3.1 ν-property

An advantage of ν-SVM over C-SVM is that ν can be interpreted on the basis of the so-called *ν-property*.

[25] For example, with $g(\|\cdot\|) = \|\cdot\|_2^2$, the optimal $\boldsymbol{\gamma}$ is equal to the weighted average of $\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_m)$; with $g(\|\cdot\|) = \|\cdot\|_1$, it is equal to the median center of $\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_m)$; with $g(\|\cdot\|) = \|\cdot\|_2$, it is equal to the geometric median (or one-median) of $\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_m)$. When $\boldsymbol{\phi}(\boldsymbol{x}_i) \in \mathbb{R}^2$ or $\mathbb{R}^3$, the optimal γ is sometimes called the Fermat–Weber point.

The ν-property is usually defined using the Karush–Kuhn–Tucker (KKT) condition (see, e.g., Vanderbei (2014) for the KKT condition). More precisely, let us consider ν-SVC (10.16). Given an optimal solution $(\boldsymbol{w}^\star, b^\star, c^\star, \boldsymbol{z}^\star)$ to (10.16) and $\boldsymbol{\lambda}^\star$ to (10.17), let us denote the set of samples that contribute to the margin error and the set of *support vectors (SVs)* by

$$\mathrm{Err} := \{i \in I : z_i^\star > 0\}, \quad \mathrm{SV} := \{i \in I : \lambda_i^\star > 0\},$$

where $I := \{1, \dots, m\}$. Accordingly, the *margin error* and number of SVs can be expressed as |Err| and |SV|. Note that Err $\subset$ SV and

$$|\mathrm{SV}| - |\mathrm{Err}| = |\{i \in I : -y_i(\boldsymbol{x}_i^\top \boldsymbol{w}^\star - b^\star) = c^\star\}|. \tag{10.33}$$

Proposition 10.1 ***(ν-property (Schölkopf et al., 2000))*** Any KKT solution to ν-SVC (10.16) or (10.17) satisfies

$$\frac{|\mathrm{Err}|}{m} \le \nu \le \frac{|\mathrm{SV}|}{m}.$$

This proposition says that ν is an upper bound of the fraction of margin errors and a lower bound of the fraction of SVs.[26]

Because of (10.33), we can see that the number of SVs is bounded above by a number depending on ν, as well. Indeed, we have

$$m\nu \le |\mathrm{SV}| \le m\nu + |\{i \in I : -y_i(\boldsymbol{x}_i^\top \boldsymbol{w}^\star - b^\star) = c^\star\}|.$$

10.3.3.2 Quantile-based ν-property

The ν-property described above depends on the KKT condition of the optimization problems (10.16) and (10.17). However, the interpretation of ν is independently obtained from the definition of the quantile (i.e., VaR).

Let us refine the margin errors and support vectors with the notion of VaR. Denoting

$$\mathrm{Err}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta}) := \{i \in I : L_i(\boldsymbol{\theta}) > \mathrm{VaR}_{(1-\nu,\boldsymbol{p})}(\boldsymbol{L}(\boldsymbol{\theta}))\},$$

we can define the *quantile-based margin error* as $|\mathrm{Err}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})|$. Furthermore, we can denote the set of *quantile-based support vectors* by

$$\mathrm{SV}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta}) := \{i \in I : L_i(\boldsymbol{\theta}) \ge \mathrm{VaR}_{(1-\nu,\boldsymbol{p})}(\boldsymbol{L}(\boldsymbol{\theta}))\}.$$

Note that the difference between $\mathrm{Err}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})$ and $\mathrm{SV}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})$ is only in the equality in the above definitions. Indeed, we have $|\mathrm{SV}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})| - |\mathrm{Err}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})| = |\{i \in I : L_i(\boldsymbol{\theta}) = \mathrm{VaR}_{(1-\nu,\boldsymbol{p})}(\boldsymbol{L}(\boldsymbol{\theta}))\}|$, and the following proposition is straightforward.

Proposition 10.2 The following holds for any $\boldsymbol{\theta}$:

$$\frac{|\mathrm{Err}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})|}{m} \le \nu \le \frac{|\mathrm{SV}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})|}{m}.$$

[26] Note that Eν-SVC also has this property because this proposition only relies on the KKT condition.

Similarly to the standard notion of SVs, the following holds:

$$m\nu \le |\mathrm{SV}_{(\nu,\boldsymbol{p})}(\boldsymbol{\theta})| \le m\nu + |\{i \in I : L_i(\boldsymbol{\theta}) = \mathrm{VaR}_{(1-\nu,\boldsymbol{p})}(\boldsymbol{L}(\boldsymbol{\theta}))\}|. \tag{10.34}$$

Note that the inequalities in Proposition 10.2 are valid for any $\boldsymbol{\theta}$, whereas the ordinary ν-property of Proposition 10.1 is shown for a KKT solution (i.e., an optimal solution). By separately defining the risk functional and optimization as in Sections 10.1.1 and 10.1.3, we can introduce the ν-property independently of the optimality condition. Accordingly, the above relation (10.34) suggests that the number of SVs can be reduced by making ν small.

10.3.3.3 Generalization Bounds for ν-SVC and Eν-SVC

The goal of learning methods is to obtain a classifier or a regressor that has a small generalization error. As mentioned in Section 10.2.1, the maximum margin hyperplane of hard-margin SVC minimizes an upper bound of the generalization error. This is considered as the reason why a high generalization performance can be obtained by hard-margin SVC for linearly separable datasets. Here, we give generalization error bounds based on the CVaR risk measure for ν-SVC and Eν-SVC and show that minimizing CVaR leads to a lower generalization bound, which will explain why a high generalization performance can be obtained by ν-SVC and Eν-SVC for linearly nonseparable datasets.

A classifier $d(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) - b)$ is learned on a training set. Here, we assume that such training samples are drawn from an unknown independent and identically distributed (i.i.d.) probability distribution $P(\boldsymbol{x}, y)$ on $\mathbb{R}^n \times \{\pm 1\}$. The goal of the classification task is to obtain a classifier d (precisely, $(\boldsymbol{w}, b)$ of d) that minimizes the generalization error defined as

$$\mathrm{GE}[d] := \int 1_{\{d(\boldsymbol{x}) \neq y\}} \, \mathrm{d}P(\boldsymbol{x}, y),$$

which corresponds to the misclassification rate for unseen test samples, but unfortunately, GE$[d]$ cannot be computed since P is unknown. A bound on the generalization error is derived, as discussed further here, and used for theoretical analysis of the learning model.

We begin with the case of $\nu \in (\nu_{\min}, \nu_{\max})$, where E$\nu$-SVC is equivalent to ν-SVC.

Theorem 10.4 (Takeda and Sugiyama, 2008) *Let $\nu \in (\nu_{\min}, \nu_{\max}]$ and $\boldsymbol{L}(\boldsymbol{w}, b) = -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)$. Suppose that $P(\boldsymbol{x}, y)$ has support in a ball of radius R around the origin. Then, for all $(\boldsymbol{w}, b)$ such that $\|\boldsymbol{w}\|_2 \le 1$ and $CVaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b)) < 0$, there exists a positive constant c such that the following bound holds with a probability of at least $1 - \delta$:*

$$\begin{aligned} GE[d] &\le \frac{|Err_{(\nu,\mathbf{1}_m/m)}(\boldsymbol{w}, b)|}{m} + \Gamma_c(VaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))) \\ &\le \nu + \Gamma_c(VaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))), \end{aligned} \tag{10.35}$$

where

$$\Gamma_c(\gamma) := \sqrt{\frac{2}{m}\left(\frac{4c^2(R^2+1)^2}{\gamma^2}\log_2(2m) - 1 + \log\frac{2}{\delta}\right)}.$$

The generalization error bound in (10.35) is furthermore upper-bounded as

$$GE[d] \leq \nu + \Gamma_c(CVaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))).$$

The function $\Gamma(\gamma)$ decreases as $|\gamma|$ increases. Note also that for $\nu \in (\nu_{\min}, \nu_{\max}]$, ν-SVC and, equivalently, Eν-SVC attain a negative minimum CVaR, that is, $\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}^\star, b^\star)) < 0$ (see Section 10.3.1). Accordingly, Theorem 10.4 implies that the minimum generalization bound regarding $(\boldsymbol{w}, b)$ is attained at an optimal solution of ν-SVC (10.25), which minimizes $\mathrm{CVaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))$ subject to $\|\boldsymbol{w}\|_2 \leq 1$. That is, it is expected that the classifier of ν-SVC has a small generalization (i.e., out-of-sample) error.

Next, we consider the case of $\nu \in (0, \nu_{\min}]$, for which ν-SVC results in a trivial solution satisfying $\boldsymbol{w} = \mathbf{0}$, but E$\nu$-SVC leads to a reasonable solution. The discussion below depends on the sign of VaR (see Figure 10.6, where the range $(0, \nu_{\min})$ is divided into two subranges corresponding to a negative VaR or a nonnegative VaR).

Theorem 10.5 (Takeda and Sugiyama, 2008) *Let $\nu \in (0, \nu_{\min})$, and let $(\boldsymbol{w}, b)$ satisfy $\|\boldsymbol{w}\|_2 = 1$.*

- *Additionally, if $(\boldsymbol{w}, b)$ satisfies $VaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b)) < 0$, there exists a positive constant c such that the following bound holds with a probability of at least $1 - \delta$:*

$$GE[d] \leq \nu + \Gamma_c(VaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))).$$

- *On the other hand, if $(\boldsymbol{w}, b)$ satisfies $VaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b)) > 0$, there exists a positive constant c such that the following bound holds with a probability of at least $1 - \delta$:*

$$GE[d] \geq \nu - \Gamma_c(VaR_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))).$$

This theorem implies that the upper bound or lower bound of GE$[d]$ can be lowered if $\mathrm{VaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))$ is reduced; indeed, minimizing $\mathrm{VaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))$ with respect to $(\boldsymbol{w}, b)$ subject to $\mathrm{VaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b)) < 0$ minimizes the upper bound, while minimizing $\mathrm{VaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b))$ subject to $\mathrm{VaR}_{(1-\nu,\mathbf{1}_m/m)}(\boldsymbol{L}(\boldsymbol{w}, b)) > 0$ minimizes the lower bound. Recalling that VaR is upper-bounded by CVaR and that Eν-SVC (10.26) is a CVaR minimization subject to $\|\boldsymbol{w}\|_2 = 1$, Eν-SVC is expected to reduce the generalization error through minimization of the upper bound or lower bound.

10.4 Duality

In this section, we present the dual problems of the CVaR-minimizing formulations in the Section 10.3. As mentioned in Section 10.2, dual representations expand the range of algorithms and enrich the theory of SVM.

10.4.1 Binary Classification

As explained in Section 10.1.3, optimization problems (10.22) with $C = 1$ and (10.25) (i.e., (10.24) with $E = 1$), are equivalent formulations of ν-SVC because of the positive

homogeneity of CVaR and the loss $\boldsymbol{L}(\boldsymbol{w}, b) = -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)$. Correspondingly, the dual formulations of the CVaR-based representations (10.22) with $C = 1$ and (10.25) can be derived as

$$\begin{cases} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\frac{1}{2}\|\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}\|_2^2 \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0, \boldsymbol{\lambda} \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})}, \end{cases} \tag{10.36}$$

and

$$\begin{cases} \underset{\boldsymbol{\lambda}}{\text{maximize}} & -\|\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}\|_2 \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0, \boldsymbol{\lambda} \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})}, \end{cases} \tag{10.37}$$

with $\boldsymbol{p} = \mathbf{1}_m/m$.[27]

By using a kernel function $k(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{\xi})$, we can readily obtain a kernelized nonlinear classification. Indeed, letting $\boldsymbol{K} := \boldsymbol{X}\boldsymbol{X}^\top$ and replacing the objective functions of (10.36) and (10.37) with $-\frac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\boldsymbol{\lambda}$ and $-\sqrt{\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\boldsymbol{\lambda}}$, respectively, each of them becomes a kernelized formulation.

10.4.2 Geometric Interpretation of ν-SVM

Observe that $\sum_{i=1}^m y_i\lambda_i = 0$ can be rewritten as $\sum_{i\in I_+} \lambda_i = \sum_{i\in I_-} \lambda_i$ and that

$$\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\boldsymbol{\lambda} = \|\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{\lambda}\|_2^2 = (\sum_{i=1}^m y_i \boldsymbol{\phi}(\boldsymbol{x}_i)\lambda_i)^2 = (\sum_{i\in I_+} \boldsymbol{\phi}(\boldsymbol{x}_i)\lambda_i - \sum_{i\in I_-} \boldsymbol{\phi}(\boldsymbol{x}_i)\lambda_i)^2,$$

where $I_+ := \{i \in \{1, \dots, m\} : y_i = +1\}$, and $I_- := \{1, \dots, m\} \backslash I_+$. With a change of variables $\boldsymbol{\mu} := \boldsymbol{\lambda}/2$, the duals (10.36) and (10.37) can be rewritten as

$$-\begin{cases} \underset{\boldsymbol{\xi}^+,\boldsymbol{\xi}^-}{\text{minimize}} & \frac{1}{8}\|\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-\|_2^2 \\ \text{subject to} & \boldsymbol{\xi}^+ \in D_+^\nu, \boldsymbol{\xi}^- \in D_-^\nu, \end{cases} \quad \text{and} \quad -\begin{cases} \underset{\boldsymbol{\xi}^+,\boldsymbol{\xi}^-}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-\|_2 \\ \text{subject to} & \boldsymbol{\xi}^+ \in D_+^\nu, \boldsymbol{\xi}^- \in D_-^\nu, \end{cases}$$

where

$$D_\bullet^\nu := \{\boldsymbol{\xi} \in \mathbb{R}^N : \boldsymbol{\xi} = \sum_{i\in I_\bullet} \boldsymbol{\phi}(\boldsymbol{x}_i)\mu_i, \boldsymbol{\mu} \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,2\boldsymbol{p}^\bullet)}\}, \quad \bullet \in \{+, -\},$$

where $\boldsymbol{p}^+ \in \mathbb{R}^{m_+}$ and $\boldsymbol{p}^- \in \mathbb{R}^{m_-}$ denote vectors whose elements come from $\boldsymbol{p}$ corresponding to $i \in I_+$ and $i \in I_-$, respectively.[28] Accordingly, the dual problems can be interpreted as ones of finding the nearest two points each belonging to D_+^ν and D_-^ν (see Figure 10.9). These sets, D_+^ν and D_-^ν, are referred to as *reduced convex hulls* or *soft convex hulls* in Bennett and Bredensteiner (2000) and Crisp and Burges (2000). Indeed, for $\nu < \nu_{\min}$, they are equivalent to the convex hulls of $\{\boldsymbol{\phi}(\boldsymbol{x}_i) : i \in I_+\}$ and $\{\boldsymbol{\phi}(\boldsymbol{x}_i) : i \in I_-\}$, respectively; the size of set $D_\bullet^\nu$ monotonically decreases in ν, that is, $D_\bullet^{\nu_1} \subset D_\bullet^{\nu_2}$ for $\nu_1 > \nu_2$; for $\nu \geq \nu_{\max}$, they shrink to their centers $\sum_{i\in I_\bullet} p_i^\bullet \boldsymbol{\phi}(\boldsymbol{x}_i)$.

[27] Obviously, the difference between (10.36) and (10.37) is only in the objective, and in practice, (10.36) is easier to solve since a quadratic program is more stably solvable than a second-order cone program.

[28] $m_+ = |I_+|$ and $m_- = |I_-|$.

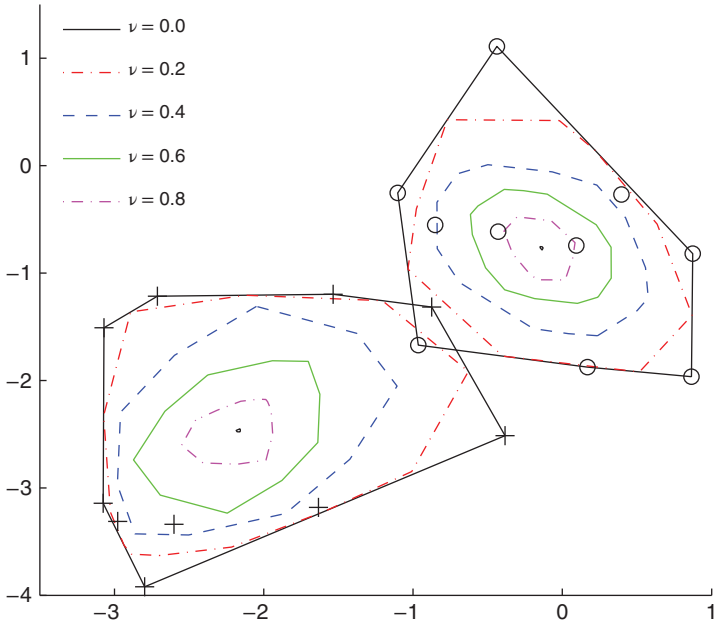

Figure 10.9 Two-dimensional examples of reduced convex hulls. Here, '+' and '∘' represent the data samples. As ν increases, the size of each reduced convex hull shrinks. The reduced convex hull is a single point for $\nu = 1$, whereas it is equal to the convex hull for ν sufficiently close to 0. For linearly inseparable datasets, the corresponding convex hulls (or the reduced convex hulls for a small ν) intersect, and the primal formulation (10.25) results in a trivial solution satisfying $(\boldsymbol{w}, b) = \mathbf{0}$.

10.4.3 Geometric Interpretation of the Range of ν for ν-SVC

Crisp and Burges (2000) pointed out that $\nu_{\min}$ is the largest ν such that two reduced convex hulls D_+^ν and D_-^ν intersect. Namely, $\nu_{\min}$ is the value such that $D_+^{\nu_{\min}}$ and $D_-^{\nu_{\min}}$ touch externally. Indeed, $\nu_{\min}$ can be computed by solving the following optimization problem:

$$\frac{1}{\nu_{\min}} := \left| \begin{array}{ll} \underset{\eta, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-}{\text{minimize}} & \eta \\ \text{subject to} & \displaystyle\sum_{i \in I_+} \boldsymbol{\phi}(\boldsymbol{x}_i)\mu_i^+ = \sum_{j \in I_-} \boldsymbol{\phi}(\boldsymbol{x}_j)\mu_j^-, \\ & \boldsymbol{\mu}^+ \in \mathcal{Q}_{\mathrm{CVaR}(1-\frac{1}{\eta}, \frac{2}{m}\mathbf{1}_{m_+})}, \quad \boldsymbol{\mu}^- \in \mathcal{Q}_{\mathrm{CVaR}(1-\frac{1}{\eta}, \frac{2}{m}\mathbf{1}_{m_-})}. \end{array} \right.$$

Note that this problem reduces to an LP.

On the other hand, if ν is smaller than $\nu_{\min}$, D_+^ν and D_-^ν intersect, and thus, ν-SVC attains zero optimal value. This is the geometric interpretation of the trivial solution mentioned in Section 10.2.1.

However, to make D_+^ν and D_-^ν non-empty, we need to choose ν satisfying $\nu \leq \frac{2m_+}{m}$ and $\nu \leq \frac{2m_-}{m}$. Consequently, $\nu_{\max}$ is defined as $\min\{\frac{2m_+}{m}, \frac{2m_-}{m}\}$.

10.4.4 Regression

The dual problems for regression (10.21) can be derived in a similar way as the SVC formulations. Using the notation $|\lambda| := (|\lambda_1|, \dots, |\lambda_m|)^\top$, the (kernelized) dual formulation of (10.21) can be symbolically represented by

$$\begin{vmatrix} \underset{\lambda}{\text{maximize}} & -\frac{1}{2H}\lambda^\top \boldsymbol{K}\lambda + \boldsymbol{y}^\top\lambda \\ \text{subject to} & \mathbf{1}_m^\top\lambda = 0, \quad |\lambda| \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}, \end{vmatrix} \tag{10.38}$$

where $\boldsymbol{K} = (k(\boldsymbol{x}_i, \boldsymbol{x}_h))_{i,h}$ with a kernel function $k(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\phi}(\boldsymbol{\xi})$. On the other hand, the dual of the norm-constrained version (10.31) can be symbolically represented by

$$\begin{vmatrix} \underset{\lambda}{\text{maximize}} & \boldsymbol{y}^\top\lambda - E\sqrt{\lambda^\top\boldsymbol{K}\lambda} \\ \text{subject to} & \mathbf{1}_m^\top\lambda = 0, \quad |\lambda| \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}. \end{vmatrix} \tag{10.39}$$

10.4.5 One-class Classification and SVDD

The (kernelized) dual formulations of the CVaR-based one-class classification (10.28) are given by

$$\begin{vmatrix} \underset{\lambda}{\text{maximize}} & -\lambda^\top\boldsymbol{K}\lambda \\ \text{subject to} & \lambda \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\mathbf{1}_m/m)}. \end{vmatrix} \tag{10.40}$$

On the other hand, the dual of the CVaR-based SVDD (10.32) is derived as

$$\begin{vmatrix} \underset{\lambda}{\text{maximize}} & \sum_{i=1}^m \sqrt{k(\boldsymbol{x}_i, \boldsymbol{x}_i)}\lambda_i - \sum_{i=1}^m \sum_{h=1}^m k(\boldsymbol{x}_i, \boldsymbol{x}_h)\lambda_i\lambda_h \\ \text{subject to} & \lambda \in \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})}. \end{vmatrix}$$

10.5 Extensions to Robust Optimization Modelings

The assumptions of machine-learning theory do not always fit real situations.[29] Some modifications can be made to bridge the gap between theory and practice. Among them is the *robust optimization* modeling (Ben-Tal *et al.*, 2009) option. In this section, we show that two kinds of robust modeling of the CVaR minimization for binary classification are tractable.

10.5.1 Distributionally Robust Formulation

The uniform distribution $\boldsymbol{p} = \mathbf{1}_m/m$ is reasonable as long as the dataset is i.i.d. sampled. However, if this does not hold true, the choice $\boldsymbol{p} = \mathbf{1}_m/m$ may not be the best. Let us tackle such

[29] For example, the i.i.d. assumption is often violated. In such a situation, it may be better to choose a non-uniform reference probability.

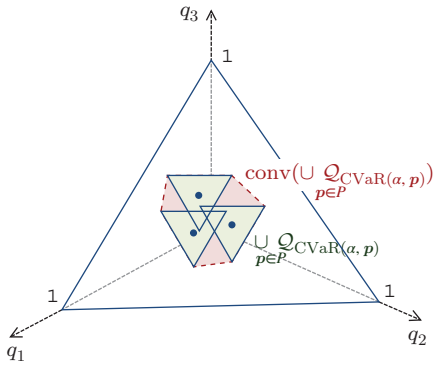

Figure 10.10 Convex hull of the union of risk envelopes ($m = 3$).

uncertainties in the choice of the reference probability $\boldsymbol{p}$ with a min-max strategy. Let $P \subset \mathbb{R}^m$ be the set of possible $\boldsymbol{p}$, and let us call it the *uncertainty set*. For simplicity, we will assume that P is closed and bounded. For some P, we replace $\mathrm{CVaR}_{(1-\nu,\boldsymbol{p})}(\boldsymbol{L})$ with the distributionally worst-case CVaR, defined by $\max_{\boldsymbol{p}\in P} \mathrm{CVaR}_{(1-\nu,\boldsymbol{p})}(\boldsymbol{L})$.

Recalling (10.11), the worst-case CVaR is represented by

$$\begin{aligned}\max_{\boldsymbol{p}\in P, \boldsymbol{q}\in\mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})}} \boldsymbol{q}^\top \boldsymbol{L} &= \max_{\boldsymbol{q}} \{\boldsymbol{q}^\top \boldsymbol{L} : \boldsymbol{q} \in \mathop{\cup}_{\boldsymbol{p}\in P} \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})}\} \\ &= \max_{\boldsymbol{q}} \{\boldsymbol{q}^\top \boldsymbol{L} : \boldsymbol{q} \in \mathrm{conv}(\mathop{\cup}_{\boldsymbol{p}\in P} \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})}\}.\end{aligned}$$

The second equality holds because of the linearity of the function $\boldsymbol{q}^\top \boldsymbol{L}$ (see Figure 10.10 for an illustration of a convex hull of the union of risk envelopes). Note that the above expression shows that the distributionally worst-case CVaR is coherent as long as $\cup_{\boldsymbol{p}\in P} \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})} \subset \mathbb{\Pi}^m$. (Recall the dual representation of the coherent risk measure described in Section 10.1.2.) Consequently, the distributionally robust ν-SVC can be written as

$$f_P^\star := \begin{vmatrix} \underset{\boldsymbol{w},b}{\text{minimize}} & \max_{\boldsymbol{q}} \{ -\boldsymbol{q}^\top \boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}_m b) : \boldsymbol{q} \in \mathrm{conv}(\mathop{\cup}_{\boldsymbol{p}\in P} \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})})\} \\ \text{subject to} & \|\boldsymbol{w}\| \le 1. \end{vmatrix} \tag{10.41}$$

Similar to the case of (10.37), the (kernelized) dual form of (10.41) is derived as[30]

$$(f_P^\star)^2 = \begin{vmatrix} \underset{\lambda}{\text{maximize}} & -\lambda^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\lambda \\ \text{subject to} & \boldsymbol{y}^\top \lambda = 0, \lambda \in \mathrm{conv}(\mathop{\cup}_{\boldsymbol{p}\in P} \mathcal{Q}_{\mathrm{CVaR}(1-\nu,\boldsymbol{p})}). \end{vmatrix} \tag{10.42}$$

For some P, problem (10.42) becomes a tractable convex optimization. Here are special cases of the examples given in Gotoh and Uryasev (2013).

[30] The objective function of (10.42) is squared: $-\lambda^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\lambda = -\lambda^\top \boldsymbol{Y}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{Y}\lambda = -\|\boldsymbol{X}^\top \boldsymbol{Y}\lambda\|_2^2$.

Example 10.5 (Finite-scenario uncertainty) *If we employ an uncertainty set defined by* $P = \{\boldsymbol{p}^1, \ldots, \boldsymbol{p}^J\}$*, with* J *candidates* $\boldsymbol{p}^1, \ldots, \boldsymbol{p}^J \in \mathbb{\Pi}^m$*, (10.42) can be rewritten as*

$$\begin{vmatrix} \underset{\lambda,\boldsymbol{\pi},\boldsymbol{\tau}}{\text{maximize}} & -\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0, \mathbf{1}_m^\top \boldsymbol{\lambda} = 1, \mathbf{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \boldsymbol{\pi} = \sum_{k=1}^J \tau_k \boldsymbol{p}_k, \mathbf{1}_J^\top \boldsymbol{\tau} = 1, \boldsymbol{\tau} \ge \mathbf{0}. \end{vmatrix}$$

This formulation was first presented in Wang (2012), which extends the robust formulation to a multiclass classification setting.

Example 10.6 (Distance-based uncertainty) *If we use an uncertainty set defined by* $P = \{\boldsymbol{q} \in \mathbb{\Pi}^m : \boldsymbol{q} = \boldsymbol{p} + \boldsymbol{A}\boldsymbol{\zeta}, \|\boldsymbol{\zeta}\| \le 1\}$*, with* $\boldsymbol{p} \in \mathbb{\Pi}^m$*, the symmetric positive definite matrix* $\boldsymbol{A} \in \mathbb{R}^{m \times m}$*, and* $\|\cdot\|$ *a norm in* $\mathbb{R}^m$*, (10.42) can be rewritten as*

$$\begin{vmatrix} \underset{\lambda,\boldsymbol{\pi},\boldsymbol{\zeta}}{\text{maximize}} & -\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0, \mathbf{1}_m^\top \boldsymbol{\lambda} = 1, \mathbf{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \boldsymbol{\pi} = \boldsymbol{p} + \boldsymbol{A}\boldsymbol{\zeta}, \mathbf{1}_m^\top \boldsymbol{A}\boldsymbol{\zeta} = 0, \|\boldsymbol{\zeta}\| \le 1. \end{vmatrix}$$

Example 10.7 (Entropy-based uncertainty) *If we use an uncertainty set defined by* $P = \left\{\boldsymbol{q} \in \mathbb{\Pi}^m : \sum_{i=1}^m q_i \ln(q_i/p_i) \le t\right\}$*, with* $t > 0$ *and* $\boldsymbol{p} \in \mathbb{\Pi}^m$ *such that* $p_i > 0$*, (10.42) can be rewritten as*

$$\begin{vmatrix} \underset{\lambda,\boldsymbol{\pi}}{\text{maximize}} & -\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}\boldsymbol{\lambda} \\ \text{subject to} & \boldsymbol{y}^\top \boldsymbol{\lambda} = 0, \mathbf{1}_m^\top \boldsymbol{\lambda} = 1, \mathbf{0} \le \boldsymbol{\lambda} \le \boldsymbol{\pi}/\nu, \\ & \sum_{i=1}^m \pi_i \ln(\pi_i/p_i) \le t, \mathbf{1}_m^\top \boldsymbol{\pi} = 1. \end{vmatrix}$$

These convex optimization problems can be solved using off-the-shelf nonlinear programming solver software packages.

10.5.2 Measurement-wise Robust Formulation

Aside from the distributionally robust formulation in Section 10.5.1, another form of uncertainty can be introduced to CVaR defined with a linear loss such as $\boldsymbol{L}(\boldsymbol{w}, b) = -\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w} - \mathbf{1}_m b)$ and $\boldsymbol{L}(\boldsymbol{w}) = -\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}$.

Recall that classifiers obtained by SVCs depend on the support vectors, which typically make up a small subset of samples. Accordingly, they are likely to be susceptible to the measurement error of a sample, that is, $\boldsymbol{\phi}(\boldsymbol{x}_i) - \Delta\boldsymbol{\phi}(\boldsymbol{x}_i)$. To mitigate the effect of such a perturbation, we may be able to use the min-max strategy, similarly to the case of distributionally robust optimization.

Let us define the set of perturbations as

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

where $\|\cdot\|$ is some norm in $\mathbb{R}^n$. Given $\mathcal{S}$, the worst-case CVaR is defined as

$$\mathrm{WCVaR}^{\mathcal{S}}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\mathbf{1}_m b)) := \max_{\boldsymbol{\Delta}\in\mathcal{S}} \mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}\{(\boldsymbol{X}-\boldsymbol{\Delta})\boldsymbol{w}-\mathbf{1}_m b\}).$$

This worst-case CVaR enables us to consider variants of ν-SVC. For example, CVaR minimization (10.25) can be modified into

$$\begin{cases} \underset{\boldsymbol{w},b}{\text{minimize}} & \mathrm{WCVaR}^{\mathcal{S}}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\mathbf{1}_m b)) \\ \text{subject to} & \|\boldsymbol{w}\|_2 \le 1, \end{cases} \tag{10.43}$$

whereas the usual ν-SVC (10.22) can be modified into

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \mathrm{WCVaR}^{\mathcal{S}}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\mathbf{1}_m b)). \tag{10.44}$$

Note that (10.43) and (10.44) are formulated as min-max optimizations. However, we can represent a robust optimization in a tractable manner by using the following formula.[31]

Proposition 10.3 (Gotoh and Uryasev, 2013) For any $(\boldsymbol{w}, b)$, we have

$$\mathrm{WCVaR}^{\mathcal{S}}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\mathbf{1}_m b)) = C\|\boldsymbol{w}\|^\circ + \mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\mathbf{1}_m b)), \tag{10.45}$$

where $\|\cdot\|^\circ$ is the dual norm of $\|\cdot\|$, that is, $\|\boldsymbol{x}\|^\circ := \max\{\boldsymbol{x}^\top \boldsymbol{z} : \|\boldsymbol{z}\| \le 1\}$.

On the basis of formula (10.45), the robust ν-SVC formulations (10.43) and (10.44) can be rewritten as

$$\begin{cases} \underset{\boldsymbol{w},b}{\text{minimize}} & C\|\boldsymbol{w}\|^\circ + \mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\mathbf{1}_m b)) \\ \text{subject to} & \|\boldsymbol{w}\|_2 \le 1, \end{cases}$$

and

$$\underset{\boldsymbol{w},b}{\text{minimize}} \quad \tfrac{1}{2}\|\boldsymbol{w}\|_2^2 + C\|\boldsymbol{w}\|^\circ + \mathrm{CVaR}_{(\alpha,\boldsymbol{p})}(-\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{w}-\mathbf{1}_m b)). \tag{10.46}$$

Note that in defining $\mathcal{S}$, we do not have to limit the norm $\|\cdot\|$ to being the ℓ_2-norm. Indeed, if $\|\cdot\| = \|\cdot\|_\infty$ is used, formulation (10.46) virtually has a regularization term of the form $\frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\|\boldsymbol{w}\|_1$, since $\|\cdot\|_\infty^\circ = \|\cdot\|_1$. In this sense, the min-max strategy yields a justification of the modified regularization term used in the *elastic net* (Zou and Hastie, 2005).[32]

10.6 Literature Review

In this final section, we briefly summarize the related literature and further extensions to the results described in this chapter.

[31] A similar result for C-SVC is derived in Xu *et al*. (2009), in which a smaller perturbation set is used in place of $\mathcal{S}$.
[32] Proposition 10.3 relies on the monotonicity and the translation invariance of CVaR, but we cannot say that the elastic net is underpinned by the same reasoning. Note that the risk functional of the elastic net satisfies neither of these properties. However, we can say that DrSVM of Wang *et al*. (2006), which is a C-SVC with an elastic net-type regularizer, is underpinned by the reasoning in Xu *et al*. (2009).

10.6.1 CVaR as a Risk Measure

The term CVaR was originally introduced by Rockafellar and Uryasev (2000), and *tail VaR* and *expected shortfall* are sometimes used for signifying the same notion. Rockafellar and Uryasev (2000) show that its minimization can be represented by an LP and the optimal $c^\star$ in the minimization formula yields an approximate VaR. Rockafellar and Uryasev (2002) extensively analyze the basic properties of CVaR and its minimization for general distributions.

On the other hand, the empirical CVaR minimization requires a large number of scenarios to make an accurate estimation. Indeed, Lim *et al.* (2011) demonstrate by simulation that CVaR minimization leads to poor out-of-sample performance.

10.6.2 From CVaR Minimization to SVM

Gotoh and Takeda (2005) were the first to describe the connection between CVaR minimization and ν-SVM (actually, Eν-SVC), while Takeda and Sugiyama (2008) were the first to point out that the model of Gotoh and Takeda (2005) is equivalent to Eν-SVC.

Regarding the robustification of CVaR minimization, Zhu and Fukushima (2009) proposed a distributionally robust portfolio selection. They consider finite scenarios and norm-based uncertainty and formulate convex optimization problems without the dual representation of CVaR. Wang (2012) uses a similar robust optimization modeling to formulate a distributionally robust multiclass ν-SVC. Gotoh *et al.* (2014) and Gotoh and Uryasev (2013) extend this robust modeling to the cases of coherent and convex functionals. Indeed, they extend the examples in Section 10.5.2 to non-CVaR functionals.

As for the measurement-wise robust ν-SVC, Gotoh and Uryasev (2013) show that any monotonic and translation-invariant functional results in a regularized empirical functional, as in Proposition 10.3.

10.6.3 From SVM to CVaR Minimization

On the other hand, the notions used in SVMs can be used in portfolio selection. Indeed, the two examples of portfolio optimization at the end of Section 10.1.3 are constrained versions of one-class ν-SVC and ν-SVR. Gotoh and Takeda (2011) present a regularized CVaR minimization by placing a norm constraint on the portfolio vector. They use regularization in the same way as in DeMiguel *et al.* (2009), where the variance-minimizing portfolio is coupled with a norm constraint. They develop generalization bounds for a portfolio selection based on the norm-constrained VaR or CVaR minimization. Gotoh and Takeda (2012) also develop generalization error bounds for portfolio selection and devise a fractional optimization problem on the basis of the empirical VaR and CVaR. El Karoui *et al.* (2012) use the asymptotic variance of the empirical CVaR as a regularizer.

10.6.4 Beyond CVaR

CVaR is the most popular *coherent measure of risk* (Artzner *et al.* 1999). The coherence of CVaR was proven by Pflug (2000) as well as by Rockafellar and Uryasev (2002). It is easy to see how such a generalized class of risk measures can be incorporated in SVMs. Gotoh

et al. (2014) apply a coherent measure of risk to SVMs in place of the CVaR of the geometric margin. By making a straightforward extension of the maximum margin SVC, they can use a negative geometric margin as the loss and deal with nonconvex optimization. On the other hand, Gotoh and Uryasev (2013) study SVC formulations on the basis of convex functionals and discuss what properties of the risk measure affect the SVC formulation. Tsyurmasto *et al.* (2013) focus on the positive homogeneity of risk measures.

A generalized risk functional of the form $\mathcal{F}[\tilde{L}] = \inf_c \{c + \mathbb{E}_{\mathbb{P}}[v(\tilde{L} - c)]\}$, where v is a nondecreasing convex function on $\mathbb{R}$, was first studied by Ben-Tal and Teboulle (1986). This functional is named *optimized certainty equivalent (OCE)*. Note that OCE can be viewed as a generalization of CVaR, since CVaR is equivalent to OCE when $v(z) = \max\{z, 0\}/(1 - \alpha)$. Kanamori *et al.* (2013) (unintentionally) apply OCE to SVC as an extension of ν-SVC. Gotoh and Uryasev (2013) point out that the use of OCE-based SVC is related to the Csiszár f-divergence (Csiszár, 1967).

Rockafellar and Uryasev (2013) provide a systematic view of risk functionals, named the *risk quadrangle*, within which a wide class of convex functionals used in risk management, statistics, economic theory, and so on are shown to be related to each other. Indeed, CVaR is associated with quantile regression (Koenker and Bassett, 1978) within the quantile-based quadrangle.

References

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

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

Bennett, K.P. and Bredensteiner, E.J. (2000). Duality and geometry in SVM classifiers. *Proceedings of International Conference on Machine Learning*, 57–64.

Ben-Tal, A. and El-Ghaoui, L. and Nemirovski, A. (2009). *Robust optimization*. Princeton: Princeton University Press.

Ben-Tal, A. and Teboulle, M. (1986). Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, 32, 1445–1466.

Boser B.E., Guyon I.M. and Vapnik V.N. (1992). A training algorithm for optimal margin classifiers. *COLT*, 144–152.

Chang, C.C. and Lin, C.J. (2001). Training ν support vector classifiers: theory and algorithms. *Neural Computation*, 13 (9), 2119–2147.

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

Crisp, D.J. and Burges, C.J.C. (2000). A geometric interpretation of ν-SVM classifiers. In *Advances in Neural Information Processing Systems 12* (ed. Solla, S.A., Leen, T.K. and Müller, K.R.). Cambridge, MA: MIT Press, 244–250.

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

DeMiguel, V., Garlappi, L., Nogales, F.J. and Uppal, R. (2009). A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Management Science*, 55 (5) 798–812.

Drucker, H., Burges C.J.C., Kaufman L., Smola A. and Vapnik V. (1997). Support vector regression machines. In *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press, 155–161.

El Karoui, N., Lim, A.E.B. and Vahn, G.Y. (2012). Performance-based regularization in mean-CVaR portfolio optimization. arXiv:1111.2091v2.

Gotoh, J. and Takeda, A. (2012). Minimizing loss probability bounds for portfolio selection. *European Journal of Operational Research*, 217 (2), 371–380.

Gotoh, J. and Takeda, A. (2011). On the role of norm constraints in portfolio selection. *Computational Management Science*, 8 (4), 323–353.

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

Gotoh, J., Takeda, A. and Yamamoto, R. (2014). Interaction between financial risk measures and machine learning methods. *Computational Management Science*, 11 (4), 365–402.

Gotoh, J. and Uryasev, S. (2013). *Support vector machines based on convex risk functionals and general norms*. Research report #2013-3. Gainesville, FL: University of Florida, Department of Industrial and Systems Engineering. www.ise.ufl.edu/uryasev/publications

Hoerl, E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

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

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Lim, A.E.B., Shanthikumar, J.G. and Vahn, G.Y. (2011). Conditional value-at-risk in portfolio optimization: coherent but fragile. *Operations Research Letters*, 39 (3), 163–171.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7, 77–91.

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

Pflug, G.C. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic constrained optimization: methodology and applications* (ed. Uryasev, S.). Berlin: Springer, 278–287.

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

Rockafellar, R.T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26 (7), 1443–1471.

Rockafellar, R.T. and Uryasev, S. (2013). The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18 (1–2), 33–53.

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

Schölkopf, B. and Smola, A.J. (2002). *Learning with kernels – support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

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

Takeda, A. and Sugiyama, M. (2008). ν-Support vector machine as conditional value-at-risk minimization. In *Proceedings of 25th Annual International Conference on Machine Learning*. New York: ACM Press, 1056–1063.

Tax, D.M.J. and Duin, R.P.W. (1999). Support vector domain description. *Pattern Recognition Letters*, 20, 1191–1199.

Tibshirani, R. (1996). Optimal reinsertion: regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58 (1), 267–288.

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

Uryasev, S. (2016). Regression models in risk management. Chapter 11 of Ali N. Akansu, Sanjeev R. Kulkarni and Dmitry Malioutov ed. Financial Signal Processing and Machine Learning. Wiley.

Vanderbei, R.J. (2014). *Linear programming: foundations and extensions*, 4th ed. Berlin: Springer.

Vapnik, V.N. (1995). *The nature of statistical learning theory*. Berlin: Springer.

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

Wang, L., Zhu, J. and Zou, H. (2006). The doubly regularized support vector machine. *Statistica Sinica*, 16 (2), 589.

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

Zabarankin, M. and Uryasev, S. (2013). *Statistical decision problems: selected concepts and portfolio safeguard case studies*. Berlin: Springer.

Zhu, S. and Fukushima, M. (2009). Worst-case conditional value-at-risk with application to robust portfolio management. *Operations Research*, 57 (5), 1155–1168.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67 (2), 301–320.