Machine Learning manuscript No.
(will be inserted by the editor)

A Revisit to Support Vector Data Description

Wei-Cheng Chang · Ching-Pei Lee · Chih-Jen Lin

Received: date / Accepted: date

Abstract Support vector data description (SVDD) is a useful method for outlier detection and has been applied to a variety of applications. However, in the existing optimization procedure of SVDD, there are some issues which may lead to improper usage of SVDD. Some of the issues might already be known in practice, but the theoretical discussion, justification and correction are still lacking. Given the wide use of SVDD, these issues inspire us to carefully study SVDD in the view of convex optimization. In particular, we derive the dual problem with strong duality, prove theorems to handle theoretical insufficiency in the literature of SVDD, investigate some novel extensions of SVDD, and come up with an implementation of training SVDD with theoretical guarantee.

Keywords Density estimation · Outlier detection · Strong duality · Convex optimization

1 Introduction

Support vector data description (SVDD), proposed by Tax and Duin (2004), is a model which aims at finding spherically shaped boundary around a data set. Given a set of training data $\boldsymbol{x}_i \in \mathbf{R}^n$, $i = 1, \ldots, l$, Tax and Duin (2004) solve the following optimiza-

Most part of this work was done when C.P. Lee was in National Taiwan University.

W.C. Chang
Department of Computer Science, National Taiwan University
E-mail: b99902019@csie.ntu.edu.tw

C.P. Lee
Department of Computer Science, University of Illinois at Urbana-Champaign
E-mail: clee149@illinois.edu

C.J. Lin
Department of Computer Science, National Taiwan University
E-mail: cjlin@csie.ntu.edu.tw

tion problem.

$$\begin{aligned} \min_{R,\boldsymbol{a},\boldsymbol{\xi}} \quad & R^2 + C\sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \le R^2 + \xi_i, i = 1, \ldots, l, \\ & \xi_i \ge 0, i = 1, \ldots, l, \end{aligned} \tag{1}$$

where ϕ is a function mapping data to a higher dimensional space, and $C > 0$ is a user-specified parameter. After (1) is solved, a hyperspherical model is characterized by the center $\boldsymbol{a}$ and the radius R. A testing instance $\boldsymbol{x}$ is detected as an outlier if

$$\|\phi(\boldsymbol{x}) - \boldsymbol{a}\|^2 > R^2.$$

Because of the large number of variables in $\boldsymbol{a}$ after the data mapping, Tax and Duin (2004) considered solving the following dual problem.

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{l} \alpha_i Q_{i,i} - \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} \\ \text{subject to} \quad & \boldsymbol{e}^T \boldsymbol{\alpha} = 1, \\ & 0 \le \alpha_i \le C, i = 1, \ldots, l, \end{aligned} \tag{2}$$

where $\boldsymbol{e} = [1, \cdots, 1]^T, \boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_l]^T$, and Q is the kernel matrix such that

$$Q_{i,j} = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j), \forall 1 \le i, j \le l.$$

Problem (2) is very similar to the support vector machine (SVM) (Boser et al., 1992; Cortes and Vapnik, 1995) dual problem, and can be solved by existing optimization methods for SVM.

SVDD has been successfully applied in a wide variety of applications such as handwritten digit recognition (Tax and Duin, 2002), face recognition (Lee et al., 2006), pattern denoising (Park et al., 2007), and anomaly detection (Banerjee et al., 2007). However, there are some issues in the existing optimization procedure of SVDD. For example, Cevikalp and Triggs (2012) pointed out that the dual problem (2) is infeasible when $C < 1/l$, and Chang et al. (2007) showed that the primal problem (1) is not convex and thus one might face the problem of local optima. Furthermore, the issue of the infeasibility when $C < 1/l$ was actually faced by users of our SVDD tools based on LIBSVM (Chang and Lin, 2011). These issues motivate us to provide a thorough analysis of SVDD in the view of convex optimization.

Given the fact that some problems of SVDD are known in practice but yet the lack of theoretical research works that successfully provide satisfactory solutions and explanations to these problems, the goal of this paper is to establish a comprehensive study, by means of convex optimization theory, to fill the gap between the practical and the theoretical parts of SVDD. Therefore, in this paper, we first review some concepts in the literature of convex optimization so as to calibrate the definitions. We then rigorously derive theorems to handle the insufficiency in existing studies of SVDD. In particular, the dual problem with strong duality is derived by considering the convex reformulation of (1) in Chang et al. (2007), and theorems are proposed to make the primal-dual relation valid for any $C > 0$. We also investigate a novel extension of SVDD that replaces the loss term of (1) by the squared-hinge loss.

The remainder of this paper is organized as follows. Section 2 outlines some essential knowledge of convex optimization and reviews details of issues in the existing optimization procedure of SVDD. Section 3 presents the proposed theorems that cover rigorous derivations along with implementation details. Section 4 further discusses some extended cases, while section 5 concludes this work.

2 Issues in Existing Studies of SVDD

In this section, we first briefly review some concepts in the literature of convex optimization, and then carefully discuss issues we mentioned in Section 1.

2.1 Convex Optimization, Strong Duality, and KKT Conditions

Many machine learning models are formulated as convex optimization problems. According to Boyd and Vandenberghe (2004), a convex optimization problem is of the form

$$\begin{aligned} \min \quad & f_0(\boldsymbol{w}) \\ \text{subject to} \quad & f_i(\boldsymbol{w}) \leq 0, i = 1, \ldots, m, \\ & h_i(\boldsymbol{w}) = 0, i = 1, \ldots, p, \end{aligned} \tag{3}$$

where $f_0, \ldots, f_m$ are convex functions and $h_1, \ldots, h_p$ are affine. When the problem (3) is difficult to solve (e.g., $\boldsymbol{w}$ is high dimensional), people may seek to solve the Lagrange dual problem, which is of the form

$$\begin{aligned} \max \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \lambda_i \geq 0, i = 1, \ldots, m, \end{aligned} \tag{4}$$

where $\boldsymbol{\lambda} \in \mathbf{R}^m$ and $\boldsymbol{\nu} \in \mathbf{R}^p$ are called the Lagrange multipliers and $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is the Lagrange dual function such that

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \equiv \inf_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{w}} \left(f_0(\boldsymbol{w}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{p} \nu_i h_i(\boldsymbol{w}) \right).$$

To distinguish from the dual problem (4), (3) is often referred to as the primal problem. Note that the optimal value of (4), denoted by d^*, is only a lower bound of p^*, the optimal value of (3). In particular, we have

$$d^* \leq p^*, \tag{5}$$

which holds even if the original problem (3) is not convex. This property is known as weak duality, and the value $p^* - d^*$ is referred to as the duality gap. However, to ensure that solving the dual problem is a viable alternative of solving the primal problem, we need equality in (5) to hold. This stronger property is called strong duality.

A convex optimization problem enjoys several nice properties including:

1. Any locally optimal point is also globally optimal.
2. Strong duality holds in a large family of convex optimization problems.

Note that strong duality does not necessarily hold for all convex optimization problems. Thus, we need some additional conditions, called constraint qualifications, to ensure strong duality. We will discuss a condition that is applicable to our case in Theorem 3.

Another way to characterize primal and dual solutions is through the *Karush-Kuhn-Tucker* (KKT) optimality conditions. Many optimization algorithms can be interpreted as methods for solving KKT conditions stated as follows.

Theorem 1 *(Boyd and Vandenberghe, 2004, Section 5.5.3) Consider any optimization problem (not necessarily convex) that both the objective and the constraint functions are differentiable. Let $\tilde{\boldsymbol{w}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ be any primal and dual optimal solutions with zero duality gap. Then we have*

$$\begin{aligned} f_i(\tilde{\boldsymbol{w}}) &\leq 0, i = 1, \ldots, m, \\ h_i(\tilde{\boldsymbol{w}}) &= 0, i = 1, \ldots, p, \\ \tilde{\lambda}_i &\geq 0, i = 1, \ldots, m, \\ \tilde{\lambda}_i f_i(\tilde{\boldsymbol{w}}) &= 0, i = 1, \ldots, m, \\ \frac{\partial L(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu})}{\partial \boldsymbol{w}} |_{(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})} &= 0. \end{aligned} \tag{6}$$

The conditions above are called the Karush-Kuhn-Tucker (KKT) conditions. In other words, KKT conditions are necessary conditions for optimality if strong duality holds.

In addition, if the optimization problem is convex, then KKT conditions are sufficient conditions for optimality.

In summary, for any optimization problem that possesses strong duality with differentiable objective and constraint functions, any pair of primal and dual optimal points must satisfy the KKT conditions.

From Theorem 1, if the problem considered is convex, then KKT conditions provide both necessary and sufficient conditions for optimality. However, if strong duality does not hold, KKT conditions are not necessary conditions for optimality. We will point out the importance of that KKT conditions being the necessary conditions for SVDD in later sections.

2.2 Convexity and Duality of (1)

Chang et al. (2007) argued that problem (1) is not convex. The reason is that the following function of $(\boldsymbol{a}, R, \xi_i)$

$$\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 - R^2 - \xi_i$$

is concave with respect to R. Therefore, problem (1) is not in the form of (3), implying that it is not a convex optimization problem. By defining

$$\bar{R} = R^2,$$

Chang et al. (2007) proposed the following convex reformulation of (1).

$$\begin{aligned} \min_{\bar{R},\boldsymbol{a},\boldsymbol{\xi}} \quad & \bar{R} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \leq \bar{R} + \xi_i, i = 1, \ldots, l, \\ & \xi_i \geq 0, i = 1, \ldots, l, \\ & \bar{R} \geq 0. \end{aligned} \tag{7}$$

A new constraint specifying the non-negativity of $\bar{R}$ is added. Because

$$\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 - \bar{R} - \xi_i = \boldsymbol{a}^T\boldsymbol{a} - 2\phi(\boldsymbol{x}_i)^T\boldsymbol{a} - \bar{R} - \xi_i + \text{constant} \tag{8}$$

is linear (and thus convex) to $\bar{R}$ as well as ξ_i, and is strictly convex with respect to $\boldsymbol{a}$, (7) is in the form of (3).

Note that the objective function of (7) is convex rather than strictly convex. Thus, (7) may possess multiple optimal solutions with the same optimal objective value. We discuss the uniqueness of optimal solutions of (7) in the following theorem.

Theorem 2 *The optimal $\boldsymbol{a}$ of* (7) *is unique. In contrast, the optimal $\bar{R}$ and $\boldsymbol{\xi}$ of* (7) *are not unique.*

The proof is in Appendix A.

The difference between (7) and (1) is that, if $(R^*, \boldsymbol{a}^*, \boldsymbol{\xi}^*)$ is a local optimal solution of (1), then $(-R^*, \boldsymbol{a}^*, \boldsymbol{\xi}^*)$ is also a local optimum, while their convex combinations might not be optimal. Thus we are not certain which local optimum is globally optimal. On the other hand, being a convex optimization problem, (7) guarantees that any convex combinations of its optimal solutions is still an optimal solution.

Under an additional assumption $\bar{R} > 0$, Chang et al. (2007) then derived the dual problem of (7) and found that it is identical to (2) derived by Tax and Duin (2004). Nonetheless, Chang et al. (2007) did not check constraint qualifications. Thus, strong duality is not guaranteed and the KKT conditions only serve as sufficient conditions of optimality. We will further discuss this problem when computing the optimal radius R in Section 3.2.

2.3 Issues in Deriving the Lagrange Dual Problem

The Lagrange dual problem of (1) is

$$\max_{\boldsymbol{\alpha}\geq\boldsymbol{0},\boldsymbol{\gamma}\geq\boldsymbol{0}} \left(\inf_{R,\boldsymbol{a},\boldsymbol{\xi}} \quad L(R, \boldsymbol{a}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \right),$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are Lagrange multipliers, and L is the Lagrangian

$$L(\boldsymbol{a}, R, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = R^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i(R^2 + \xi_i - \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2) - \sum_{i=1}^{l}\gamma_i\xi_i.$$

To obtain the infimum of L, Tax and Duin (2004) set both the partial derivatives of L with respect to R and $\boldsymbol{a}$ to be zero.

$$\frac{\partial L}{\partial R} = 0 \quad \Rightarrow \quad R(1 - \sum_{i=1}^{l} \alpha_i) = 0, \tag{9}$$

$$\frac{\partial L}{\partial \boldsymbol{a}} = 0 \quad \Rightarrow \quad \sum_{i=1}^{l} \alpha_i \phi(\boldsymbol{x}_i) - \boldsymbol{a} \sum_{i=1}^{l} \alpha_i = 0. \tag{10}$$

From (9), they stated

$$\sum_{i=1}^{l} \alpha_i = 1, \tag{11}$$

and got

$$\boldsymbol{a} = \sum_{i=1}^{l} \alpha_i \phi(\boldsymbol{x}_i)$$

by combining (10) and (11). They then obtained the dual problem (2) based on the above results.

However, if $R = 0$, (11) does not necessarily hold. Thus it is unclear if further derivations based on (11) are valid. To rule out a similar issue in deriving the dual problem of (7), Chang et al. (2007) explicitly assume $\bar{R} > 0$. Nonetheless, we will later show in Section 3.2 that this assumption may fail. Finally, as pointed out in Cevikalp and Triggs (2012), problem (2) does not have any feasible solution when $0 < C < 1/l$. In contrast, the primal problems (1) and (7) are feasible for any $C > 0$. Therefore, neither the relation between (1) and (2) nor that between (7) and (2) is entirely clear.

3 Convexity and the Dual Problem of SVDD

In this section, we carefully address all issues mentioned in Section 2. First, we consider the convex reformulation (7) and check its constraint qualifications before rigorously deriving the dual problem. Second, we propose theorems to remove the assumption $\bar{R} > 0$ and further justify the primal-dual relation for any $C > 0$.

3.1 Strong Duality and Constraint Qualifications

In order to ensure strong duality of problem (7), we check if (7) satisfies any constraint qualifications. Many types of constraint qualifications have been developed in the field of convex optimization. We consider Slater's condition here.

Theorem 3 *(Boyd and Vandenberghe, 2004, Section 5.2.3, Refined Slater's condition) For any set of functions $f_i, i = 0, \dots, m$, and any set S, define*

$$\mathbf{D} \equiv \left(\cap_{i=0}^{m} domain\,(f_i)\right),$$

and

$$relint(S) \equiv \{\boldsymbol{w} \in S \mid \exists r > 0, B_r(\boldsymbol{w}) \cap aff(S) \subset S\},$$

where $B_r(\boldsymbol{w})$ is the open ball centered at $\boldsymbol{w}$ with radius r, and aff(S) is the affine hull of S. Consider the convex optimization problem (3)*, If the first k constraint functions $f_1, \ldots f_k$ are affine, then strong duality for problem* (3) *holds if there exists a $\boldsymbol{w} \in$ relint$(\mathbf{D})$ such that*

$$\begin{aligned}
f_i(\boldsymbol{w}) &\leq 0, i = 1, \ldots, k, \\
f_i(\boldsymbol{w}) &< 0, i = k+1, \ldots, m, \\
h_i(\boldsymbol{w}) &= 0, i = 1, \ldots, p.
\end{aligned}$$

Note that it is rather simple to verify Theorem 3 for SVM since SVM involves only affine constraints while SVDD has nonlinear constraints.

We then apply this theorem to obtain the strong duality of (7).

Corollary 1 *For the convex optimization problem* (7) *with any data $\boldsymbol{x}_i, i = 1, \ldots, l$, strong duality holds.*

Proof Note that there is no equality constraints in (7), and the domain of $f_0, \ldots, f_m$ are all the same Euclidean space. Hence every point in that space lies in the relint of this space. We let $\boldsymbol{a} = \mathbf{0}, \bar{R} = 1$, and

$$\xi_i = \|\phi(\boldsymbol{x}_i)\|^2 + 1.$$

Then clearly $(\bar{R}, \boldsymbol{a}, \boldsymbol{\xi})$ is a feasible solution and

$$\begin{aligned}
\bar{R} &= 1 > 0, \\
\xi_i &\geq 1 > 0, \\
\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 - \bar{R} - \xi_i &= \|\phi(\boldsymbol{x}_i)\|^2 - 1 - \|\phi(\boldsymbol{x}_i)\|^2 - 1 = -2 < 0.
\end{aligned}$$

Thus $(\bar{R}, \boldsymbol{a}, \boldsymbol{\xi})$ satisfies Theorem 3 and strong duality for (7) holds.

In some studies of SVDD such as Chang et al. (2007) and Wang et al. (2011), they derived the dual problem by applying the KKT optimality conditions without examining strong duality. However, as discussed earlier, to have that KKT conditions are necessary and sufficient for optimality to convert a dual optimal solution to a primal optimal solution, strong duality and hence constraint qualifications are still needed.

3.2 The Dual Problem of (7)

Recall that a difficulty of deriving (11) from (9) is that R may be zero. In a similar derivation for the dual problem of (7), Chang et al. (2007) assume that $\bar{R} > 0$, but this assumption may not hold. We illustrate the infeasibility of this assumption and handle the difficulty of deriving the dual problem by the following theorem.

Theorem 4 *Consider problem* (7)*.*

1. *For any $C > 1/l$, the constraint $\bar{R} \geq 0$ in* (7) *is not necessary. That is, without this constraint, any optimal solution still satisfies $\bar{R} \geq 0$.*
2. *For any $0 < C < 1/l$, $\bar{R} = 0$ is uniquely optimal. If $C = 1/l$, then at least one optimal solution has $\bar{R} = 0$.*

The proof is in Appendix B. With Theorem 4, we can clearly see that $\bar{R}$ may be zero rather than positive, which implies that the assumption of $\bar{R} > 0$ in Chang et al. (2007) may fail whenever $C \leq 1/l$. Moreover, the first case in Theorem 4 still cannot guarantee $\bar{R} > 0$. We therefore conclude that the derivations in Tax and Duin (2004); Chang et al. (2007) and Wang et al. (2011) are not always valid.

According to Theorem 4, we now derive the dual problem by considering $C > 1/l$ and $C \leq 1/l$ separately.

Case 1: $C > 1/l$.
The Lagrangian of (7) without the constraint $\bar{R} \geq 0$ is

$$\begin{aligned} L(\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) =& \bar{R} + C\sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i(\bar{R} + \xi_i - \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2) - \sum_{i=1}^{l} \gamma_i \xi_i \\ =& \bar{R}\left(1 - \sum_{i=1}^{l} \alpha_i\right) + \sum_{i=1}^{l} \xi_i \left(C - \alpha_i - \gamma_i\right) + \sum_{i=1}^{l} \alpha_i \left(\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2\right), \end{aligned} \tag{12}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are Lagrange multipliers. The Lagrange dual problem is

$$\max_{\boldsymbol{\alpha} \geq 0, \boldsymbol{\gamma} \geq \mathbf{0}} \left(\inf_{\bar{R}, \boldsymbol{a}, \boldsymbol{\xi}} \quad L(\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \right). \tag{13}$$

Clearly, if $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ satisfies

$$1 - \boldsymbol{e}^T \boldsymbol{\alpha} \neq 0,$$

or

$$C - \alpha_i - \gamma_i \neq 0 \text{ for some } i,$$

then

$$\inf_{\bar{R}, \boldsymbol{a}, \boldsymbol{\xi}} L(\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = -\infty.$$

Such $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ should not be considered because of the maximization over $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ in (13). This leads to the following constraints in the dual problem.

$$1 - \boldsymbol{e}^T \boldsymbol{\alpha} = 0, \tag{14}$$

$$C - \alpha_i - \gamma_i = 0, i = 1, \ldots, l. \tag{15}$$

Substituting (14) and (15) into (13), and taking $\gamma_i \geq 0, \forall i$ into account, the dual problem (13) is reduced to

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \left(\inf_{\boldsymbol{a}} \sum_{i=1}^{l} \alpha_i \left(\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2\right) \right) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, i = 1, \ldots, l, \\ & \boldsymbol{e}^T \boldsymbol{\alpha} = 1. \end{aligned} \tag{16}$$

Because

$$\sum_{i=1}^{l} \alpha_i \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2$$

is strictly convex with respect to the unbounded variable $\boldsymbol{a}$, the infimum occurs at the point that the partial derivative with respect to $\boldsymbol{a}$ is zero.

$$\boldsymbol{a}\sum_{i=1}^{l}\alpha_i = \sum_{i=1}^{l}\alpha_i\phi(\boldsymbol{x}_i). \tag{17}$$

By the constraint (14), (17) is equivalent to

$$\boldsymbol{a} = \frac{\sum_{i=1}^{l}\alpha_i\phi(\boldsymbol{x}_i)}{\boldsymbol{e}^T\boldsymbol{\alpha}} = \sum_{i=1}^{l}\alpha_i\phi(\boldsymbol{x}_i). \tag{18}$$

We then obtain the following dual problem for $C > 1/l$.

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{l}\alpha_i Q_{i,i} - \boldsymbol{\alpha}^T Q\boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \le \alpha_i \le C, i = 1, \ldots, l, \\ & \boldsymbol{e}^T\boldsymbol{\alpha} = 1, \end{aligned} \tag{19}$$

which is the same as (2).

Note that if we do not apply Theorem 4 to remove the constraint $\bar{R} \ge 0$, the Lagrangian has an additional term $-\beta\bar{R}$, where β is the corresponding Lagrange multiplier. Then the constraint (14) becomes

$$1 - \boldsymbol{e}^T\boldsymbol{\alpha} - \beta = 0 \quad \text{and} \quad \beta \ge 0.$$

The situation becomes complicated because we must check if $\boldsymbol{e}^T\boldsymbol{\alpha} > 0$ or not before dividing $\boldsymbol{e}^T\boldsymbol{\alpha}$ from both sides of (17).

We discuss how to obtain the primal optimal solution after solving the dual problem. Clearly, the optimal $\boldsymbol{a}$ can be obtained by (18). Tax and Duin (2004) find $\bar{R}$ by the following setting of using an optimal α_i with $0 < \alpha_i < C$. By Theorem 1 and the constraint qualifications verified in Section 3.1, KKT optimality conditions are now both necessary and sufficient conditions. Therefore, primal and dual optimal solutions satisfy the following complementary slackness conditions.

$$\gamma_i\xi_i = 0 \text{ and } \alpha_i\left(\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 - \bar{R} - \xi_i\right) = 0, i = 1, \ldots, l. \tag{20}$$

Consider (20) along with (15). If there exists an index i such that $0 < \alpha_i < C$, then we have

$$\xi_i = 0 \quad \text{and} \quad \bar{R} = \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2. \tag{21}$$

Notice that none of Tax and Duin (2004), Chang et al. (2007) and Wang et al. (2011) verified strong duality, so KKT may not be necessary conditions. That is, they did not ensure that $\boldsymbol{\alpha}$ of (2) satisfies (20), and therefore their derivation of the optimal $\bar{R}$ by (21) is not rigorous.

As pointed out by Wang et al. (2011), however, it is possible that all α_i values are bounded. In this circumstance, the method of Tax and Duin (2004) failed and Wang et al. (2011) spent considerable efforts to show that the optimal $\bar{R}$ is not unique but can be any value in an interval. By a simple proof, we easily obtain $\bar{R}$ in the following theorem, regardless of whether some $0 < \alpha_i < C$ exist or not.

Theorem 5 *When $C > 1/l$, given the optimal $\boldsymbol{a}$ of* (7) *and an optimal $\boldsymbol{\alpha}$ of* (19)*, a feasible $\bar{R}$ is optimal for* (7) *if and only if*

$$\max_{i:\alpha_i<C} \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \leq \bar{R} \leq \min_{i:\alpha_i>0} \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2. \tag{22}$$

The proof is in Appendix C. If there exists an index i with $0 < \alpha_i < C$, (22) is reduced to (21) and the optimal $\bar{R}$ is unique. Otherwise, if every α_i is bounded (i.e., its value is 0 or C), then (22) indicates that any $\bar{R}$ in an interval is optimal. Interestingly, (22) is similar to the inequality for the bias term b in SVM problems; see, for example, Chang and Lin (2011). For the practical implementation, we may use the following setting adopted from LIBSVM (Chang and Lin, 2011) to calculate $\bar{R}$.

1. If some indices satisfy $0 < \alpha_i < C$, then we calculate the average of $\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2$ over all such i. The reason is that each single $\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2$ may be inaccurate because of numerical errors.
2. If all α_i are bounded, then we choose $\bar{R}$ to be the middle point of the interval in (22).

Finally, it is straightforward from the primal function (7) that the optimal $\boldsymbol{\xi}$ can be computed by

$$\xi_i = \max\left(\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 - \bar{R}, 0\right), i = 1, \ldots, l. \tag{23}$$

Another interesting property is that when C is large, all models of SVDD are identical. This result was known by judging from the dual constraints in (19). Here we complement this fact by providing a theorem on the primal problem.

Theorem 6 *For any $C > 1$, problem* (7) *is equivalent to the following problem.*

$$\begin{aligned} \min_{\bar{R},\boldsymbol{a}} \quad & \bar{R} \\ \textit{subject to} \quad & \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \leq \bar{R}. \end{aligned} \tag{24}$$

The proof is in Appendix D. Note that the dual problem of (24) is (19) without the constraint $\alpha_i \leq C$ for all i. The relation between (7) and (24) is similar to that between soft- and hard-margin SVMs, where the latter uses neither C nor $\boldsymbol{\xi}$ because of assuming that the training instances are separable. For SVM, it is known that if the training instances are separable, there is a $\bar{C}$ such that for all $C > \bar{C}$, the solution is the same as that of the problem without the loss term; see, for example, Lin (2001). This $\bar{C}$ is problem dependent, but for SVDD, we have shown that $\bar{C}$ is one.

Case 2: $C \leq 1/l$.

By Theorem 4, we note that in this case, the task is reduced to finding the optimal $\boldsymbol{a}$, and any test point that is not identical to this $\boldsymbol{a}$ is categorized as an outlier. To solve the optimization problem, using Theorem 4, we first remove the variable $\bar{R}$ from problem (7). Because the minimum must occur when

$$\xi_i = \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \geq 0,$$

problem (7) can be reduced to

$$\min_{\boldsymbol{a}} \quad \sum_{i=1}^{l} \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2. \tag{25}$$

This problem is strictly convex to $\boldsymbol{a}$, so setting the gradient to be zero leads to

$$\boldsymbol{a} = \frac{\sum_{i=1}^{l} \phi(\boldsymbol{x}_i)}{l}. \tag{26}$$

Therefore, when $C \leq 1/l$, the optimal solution is independent of C. Further, the optimization problem has a closed-form solution (26) and we thus do not need to consider the dual problem.

Note that as explained in Lin (2001), our derivation here also work when $\boldsymbol{w}$ is of infinite dimension. This may happen when, for example, an RBF kernel is used.

3.3 Implementation Issues

The dual problem (19) is very similar to the SVM dual problem. They both have a quadratic objective function involving the kernel matrix, one linear constraint, and l bounded constraints. Therefore, existing optimization methods such as decomposition methods (Platt, 1998; Joachims, 1998; Fan et al., 2005) for SVM dual problems can be easily applied to our problem. We also note that (19) is related to the dual problem of one-class SVM (Schölkopf et al., 2001), which is another method for outlier detection.

In the prediction stage, for any test instance $\boldsymbol{x}$, we must check the value

$$\|\phi(\boldsymbol{x}) - \boldsymbol{a}\|^2 - \bar{R}.$$

If it is positive, then $\boldsymbol{x}$ is considered as an outlier. If a kernel is used and $C > 1/l$, then from (18), the calculation is conducted by

$$\|\phi(\boldsymbol{x}) - \boldsymbol{a}\|^2 - \bar{R} = K(\boldsymbol{x}, \boldsymbol{x}) - 2 \sum_{i:\alpha_i > 0} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \bar{R},$$

where $K(\cdot, \cdot)$ is the kernel function such that

$$K(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^T \phi(\boldsymbol{y}), \forall \boldsymbol{x}, \boldsymbol{y}.$$

The $\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \bar{R}$ term is expensive to calculate, but it is independent from the test instances. A trick is to store this constant after solving the dual problem.

4 Extensions

In this section, we discuss some extensions of SVDD.

4.1 L2-Loss SVDD

Tax and Duin (2004) consider L1 loss (hinge loss) in the formulation of SVDD. In SVM, L2 loss (squared-hinge loss) is a common alternative to L1 loss. Surprisingly, however, we have not seen any paper studying details of L2-loss SVDD. From the experience of SVM, the performance of L2-loss SVM may outweigh L1 loss SVM in some circumstances. This motivates us to conduct a thorough investigation. We will show that the derivation of L2-loss SVDD is not trivial and has some subtle differences from the L1-loss case.

The optimization problem of L2-loss SVDD is

$$\begin{aligned} \min_{\bar{R},\boldsymbol{a},\boldsymbol{\xi}} \quad & \bar{R} + C\sum_{i=1}^{l} \xi_i^2 \\ \text{subject to} \quad & \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \le \bar{R} + \xi_i, i = 1, \ldots, l, \\ & \bar{R} \ge 0. \end{aligned} \tag{27}$$

Note that the constraint $\xi_i \ge 0, \forall i$ appeared in (7) is not necessary for L2-loss SVDD, because if at an optimum, $\xi_i < 0$ for some i, we can then replace ξ_i with 0 so that

$$\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \le \bar{R} + \xi_i < \bar{R} + 0.$$

The constraints are still satisfied, but the objective value is smaller. This contradicts the assumption that ξ_i is optimal. Similar to the L1-loss case, because of using $\bar{R}$ rather than R^2, (27) is a convex optimization problem. Furthermore, Slater's condition holds, and thus so does the strong duality. Because the loss term $C\sum_{i=1}^{l} \xi_i^2$ is now strictly convex, we are able to prove the uniqueness of the optimum.

Theorem 7 *The optimal solution of* (27) *for any* $C > 0$ *is unique.*

The proof is in Appendix E.

Similar to the L1-loss case, with the help of Theorem 7, we discuss how to solve (27) in two cases according to the following theorem.

Theorem 8 *Let* $(\boldsymbol{a}^*, \boldsymbol{\xi}^*)$ *be the optimal solution of the following problem,*[1]

$$\begin{aligned} \min_{\boldsymbol{a},\boldsymbol{\xi}} \quad & \sum_{i=1}^{l} \xi_i^2 \\ \textit{subject to} \quad & \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \le \xi_i. \end{aligned} \tag{28}$$

If $\sum_{i=1}^{l} \xi_i^* > 0$*, we define*

$$C^* = \frac{1}{2\sum_{i=1}^{l} \xi_i^*}, \tag{29}$$

and have:

1. *For any* $C > C^*$*, the optimal* $\bar{R}$ *satisfies* $\bar{R} > 0$*. Furthermore, the constraint* $\bar{R} \ge 0$ *in* (27) *is not necessary.*
2. *For any* $0 < C \le C^*$*,* $\bar{R} = 0$ *is optimal.*

The proof is in Appendix F. The case $\sum_{i=1}^{l} \xi_i^* = 0$ not covered in Theorem 8 happens only when $\phi(\boldsymbol{x}_1) = \ldots = \phi(\boldsymbol{x}_l)$. In this case, the optimal solution of (28) is $\boldsymbol{a}^* = \phi(\boldsymbol{x}_i)$ and $\xi_i^* = 0, \forall i = 1, \ldots, l$. We can easily rule out this situation beforehand. Clearly, C^* plays the same role as $1/l$ in Theorem 4 for L1-loss SVDD. The main difference is that C^* is problem dependent.

Following Theorem 8, we discuss the two situations $C > C^*$ and $C \le C^*$ in detail.

[1] The uniqueness of $(\boldsymbol{a}^*, \boldsymbol{\xi}^*)$ can be derived by the same method to prove Theorem 7.

Case 1: $C > C^*$.
The Lagrangian of (27) without the constraint $\bar{R} \geq 0$ is

$$L(\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \bar{R} + C \sum_{i=1}^{l} \xi_i^2 - \sum_{i=1}^{l} \alpha_i \left(\bar{R} + \xi_i - \| \phi(\boldsymbol{x}_i) - \boldsymbol{a} \|^2 \right), \tag{30}$$

where $\boldsymbol{\alpha}$ is the Lagrange multiplier. The Lagrange dual problem is

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \left(\inf_{\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}} \quad L(\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}, \boldsymbol{\alpha}) \right). \tag{31}$$

Clearly, if

$$1 - \boldsymbol{e}^T \boldsymbol{\alpha} \neq 0,$$

then

$$\inf_{\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}} L(\boldsymbol{a}, \bar{R}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = -\infty.$$

Thus we have the following constraint in the dual problem.

$$1 - \boldsymbol{e}^T \boldsymbol{\alpha} = 0. \tag{32}$$

In addition, L is strictly convex to $\xi_i, \forall i$, so we have

$$\frac{\partial L}{\partial \xi_i} = 0 \quad \Rightarrow \quad \xi_i = \frac{\alpha_i}{2C}, i = 1, \ldots, l. \tag{33}$$

Substituting (32) and (33) into (31), the dual problem (31) is reduced to

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \left(\inf_{\boldsymbol{a}} \sum_{i=1}^{l} \alpha_i \| \phi(\boldsymbol{x}_i) - \boldsymbol{a} \|^2 - \sum_{i=1}^{l} \frac{\alpha_i^2}{4C} \right) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \infty, i = 1, \ldots, l, \\ & \boldsymbol{e}^T \boldsymbol{\alpha} = 1. \end{aligned} \tag{34}$$

Similar to the derivation from (16) to (17), the infimum occurs when

$$\boldsymbol{a} = \sum_{i=1}^{l} \alpha_i \phi(\boldsymbol{x}_i). \tag{35}$$

Finally, the dual problem is

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{l} \alpha_i Q_{i,i} - \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \sum_{i=1}^{l} \frac{\alpha_i^2}{4C} \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \infty, i = 1, \ldots, l, \\ & \boldsymbol{e}^T \boldsymbol{\alpha} = 1, \end{aligned} \tag{36}$$

which is very similar to (19). One minor difference is that similar to the dual problem of L2-loss SVM, (36) has an additional $\sum_{i=1}^{l} (\alpha_i^2 / 4C)$ term, so the optimization problem is strongly convex. Despite the situation that $\boldsymbol{\alpha}$ is unbounded above, note that the equality constraint along with the non-negative constraints in (36) implicitly set an upper bound for $\boldsymbol{\alpha}$.

After the dual problem (36) is solved, we use (35) to compute the optimal $\boldsymbol{a}$. The computation of the optimal $\boldsymbol{\xi}$ is shown in (33). Combining the KKT condition

$$\alpha_i(\bar{R} + \xi_i - \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2) = 0$$

with (33), we obtain the optimal $\bar{R}$.

$$\bar{R} = \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 - \frac{\alpha_i}{2C}, \text{ for any } i \text{ such that } \alpha_i > 0.$$

We are always able to find an $\alpha_i > 0$ to compute the optimal $\bar{R}$ directly because of the constraints $0 \leq \alpha_i \leq \infty, \forall i = 1 \ldots, l$ and $\boldsymbol{e}^T\boldsymbol{\alpha} = 1$ in (36). In contrast, for the L1-loss case, because of $0 \leq \alpha_i \leq C$, we may not be able to find an $\alpha_i \in (0, C)$. Note that we also follow the computation of $\bar{R}$ in L1-loss SVDD to average the results of all $\alpha_i > 0$.

Case 2: $C \leq C^*$.
We first note that after setting $\bar{R}$ to be zero, (27) degenerates to (28) that is independent of C. Thus, if we already know the value of C^* by solving (28), then clearly we already know the optimal solution in this case. The remaining issue is how to solve (28). Note that this problem is in a more complicated form than (25), and thus does not have a closed-form solution.

We consider the KKT conditions and obtain that at the optimal $(\boldsymbol{a}, \boldsymbol{\xi})$, there exists $\boldsymbol{\alpha}$ such that

$$\begin{aligned}\alpha_i &= 2\xi_i \geq 0,\\ \sum_{i=1}^{l} \alpha_i(\boldsymbol{a} - \phi(\boldsymbol{x}_i)) &= \mathbf{0}.\end{aligned}$$

Therefore, as long as $\boldsymbol{\alpha} \neq \mathbf{0}$, by normalizing $\boldsymbol{\alpha}$ we have

$$\boldsymbol{a} = \sum_{i=1}^{l} \beta_i \phi(\boldsymbol{x}_i), \text{ where } \beta_i \geq 0, \sum_{i=1}^{l} \beta_i = 1.$$

The only case that $\boldsymbol{\alpha} = \mathbf{0}$ can happen is $\phi(\boldsymbol{x}_1) = \ldots = \phi(\boldsymbol{x}_l)$ and this can be easily ruled out before we try to solve the problem. Now we can turn to solve the following problem.

$$\begin{aligned}\min_{\boldsymbol{\beta}} \quad & \sum_{i=1}^{l} \|\phi(\boldsymbol{x}_i) - \sum_{i=1}^{l} \beta_i \phi(\boldsymbol{x}_i)\|^4\\ \text{subject to} \quad & 0 \leq \beta_i, i = 1, \ldots, l, \\ & \boldsymbol{e}^T\boldsymbol{\beta} = 1.\end{aligned} \tag{37}$$

Because constraints in (37) are the same as those in (2) and (36), decomposition methods can be modified to solve (37).

4.2 Smallest Circle Encompassing the Data

The radius of the smallest circle encompassing all training instances is useful for evaluating an upper bound of leave-one-out error for SVMs (Vapnik and Chapelle, 2000; Chung et al., 2003). It can be computed by a simplified form of (7) without considering $\boldsymbol{\xi}$.

$$\begin{aligned} \min_{\bar{R}, \boldsymbol{a}} \quad & \bar{R} \\ \text{subject to} \quad & \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \leq \bar{R}. \end{aligned}$$

Note that this is identical to (24). Past works have derived the dual problem of (24). As expected, it is (19) without the constraint $\alpha_i \leq C, \forall i$. A practical issue is that when applying an optimization procedure for (19) to solve the dual problem here, replacing C with ∞ may cause numerical issues. We address this issue by applying Theorem 6. That is, to solve (24), all we have to do is to solve (7) with any $C > 1$.

5 Conclusions

In this paper, we point out insufficiencies in the existing literature of SVDD. We then conduct a thorough investigation, rigorously derive the dual problem of SVDD with strong duality to ensure the optimality of the original primal problem, discuss additional properties, and study some extensions of SVDD. Based on this work, we have updated the extension of LIBSVM for SVDD at LIBSVM Tools.[2]

A Proof of Theorem 2

If $\boldsymbol{a}$ is not unique, then there are two optimal solutions $(\bar{R}_1, \boldsymbol{a}_1, \boldsymbol{\xi}_1)$ and $(\bar{R}_2, \boldsymbol{a}_2, \boldsymbol{\xi}_2)$ such that

$$\boldsymbol{a}_1 \neq \boldsymbol{a}_2 \tag{38}$$

and

$$\bar{R}_1 + C \sum_{i=1}^{l} (\boldsymbol{\xi}_1)_i = \bar{R}_2 + C \sum_{i=1}^{l} (\boldsymbol{\xi}_2)_i. \tag{39}$$

We begin with showing that the optimal objective value is positive. Otherwise, constraints in (7) implies that

$$\bar{R}_1 = \bar{R}_2 = 0 \quad \text{and} \quad \boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = \mathbf{0}.$$

Then

$$0 = \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}_1\| = \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}_2\|, \forall i = 1, \ldots, l$$

implies

$$\boldsymbol{a}_1 = \boldsymbol{a}_2,$$

a contradiction to (38).

Because $\|\phi(\boldsymbol{x}) - \boldsymbol{a}\|^2$ is strictly convex with respect to $\boldsymbol{a}$ and $\boldsymbol{a}_1 \neq \boldsymbol{a}_2$ from (38), there exists $\theta \in (0,1)$ such that for all $i = 1, \ldots, l$,

$$\begin{aligned} \|\phi(\boldsymbol{x}_i) - (\theta \boldsymbol{a}_1 + (1-\theta)\boldsymbol{a}_2)\|^2 &< \theta \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}_1\|^2 + (1-\theta)\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}_2\|^2 \\ &\leq \theta(\bar{R}_1 + (\boldsymbol{\xi}_1)_i) + (1-\theta)(\bar{R}_2 + (\boldsymbol{\xi}_2)_i). \end{aligned} \tag{40}$$

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#libsvm_for_svdd_and_finding_the_smallest_sphere_containing_all_data.

From (40), there exists $\varDelta \in (0,1)$ such that

$$\varDelta \left(\theta \begin{bmatrix} \bar{R}_1 \\ \boldsymbol{\xi}_1 \end{bmatrix} + (1-\theta) \begin{bmatrix} \bar{R}_2 \\ \boldsymbol{\xi}_2 \end{bmatrix} \right)$$

satisfies that $\forall i = 1, \ldots, l$,

$$\|\phi(\boldsymbol{x}_i) - (\theta \boldsymbol{a}_1 + (1-\theta)\boldsymbol{a}_2)\|^2 \leq \varDelta \left(\theta \left(\bar{R}_1 + (\boldsymbol{\xi}_1)_i \right) + (1-\theta) \left(\bar{R}_2 + (\boldsymbol{\xi}_2)_i \right) \right).$$

Therefore,

$$\theta \begin{bmatrix} \varDelta \bar{R}_1 \\ \boldsymbol{a}_1 \\ \varDelta \boldsymbol{\xi}_1 \end{bmatrix} + (1-\theta) \begin{bmatrix} \varDelta \bar{R}_2 \\ \boldsymbol{a}_2 \\ \varDelta \boldsymbol{\xi}_2 \end{bmatrix}$$

is a feasible points for (7). However, with (39), the new objective value is

$$\varDelta\theta(\bar{R}_1 + C\sum_{i=1}^{l}(\boldsymbol{\xi}_1)_i) + \varDelta(1-\theta)(\bar{R}_2 + C\sum_{i=1}^{l}(\boldsymbol{\xi}_2)_i) = \varDelta(\bar{R}_1 + C\sum_{i=1}^{l}(\boldsymbol{\xi}_1)_i).$$

Because we have shown that the optimal objective value is positive, this new value is strictly smaller than the optimal objective value we have. Thus, there is a contradiction. Therefore, our assumption in (38) is incorrect and the optimal $\boldsymbol{a}$ is unique.

We give an example to show that the optimal $\bar{R}$ and $\boldsymbol{\xi}$ are not unique. Consider problem (7) with $C = 1/2$ and the input instances are $\boldsymbol{x}_1 = 1, \boldsymbol{x}_2 = -1, \boldsymbol{x}_3 = 2, \boldsymbol{x}_4 = -2$. We claim that the following two points $(\bar{R}_1, \boldsymbol{a}, \boldsymbol{\xi}_1)$ and $(\bar{R}_2, \boldsymbol{a}, \boldsymbol{\xi}_2)$ where

$$\begin{aligned} \bar{R}_1 &= 1, \quad \boldsymbol{a} = \boldsymbol{0}, \quad \boldsymbol{\xi}_1 = [0,0,3,3]^T, \\ \bar{R}_2 &= 4, \quad \boldsymbol{a} = \boldsymbol{0}, \quad \boldsymbol{\xi}_2 = [0,0,0,0]^T, \end{aligned} \tag{41}$$

are both optimal. Clearly, they are both feasible. Furthermore, we have

$$\bar{R} + C\sum_{i=1}^{4}\xi_i \geq \bar{R} + \frac{1}{2}(\xi_3 + \xi_4) \geq \frac{1}{2}\min_{\boldsymbol{a}} \quad (\boldsymbol{a}-2)^2 + (\boldsymbol{a}+2)^2 = 4.$$

Thus the optimal objective value is at least 4. Because points in (41) give the objective value 4, $(\bar{R}_1, \boldsymbol{a}, \boldsymbol{\xi}_1)$ and $(\bar{R}_2, \boldsymbol{a}, \boldsymbol{\xi}_2)$ are both optimal.

B Proof of Theorem 4

For any $C > 1/l$, consider problem (7) without the constraint $\bar{R} \geq 0$. Assume it has an optimal $(\bar{R}, \boldsymbol{a}, \boldsymbol{\xi})$ with $\bar{R} < 0$. We consider a new point $(0, \boldsymbol{a}, \boldsymbol{\xi} + \bar{R}\boldsymbol{e})$, where $\boldsymbol{e}$ is the vector of ones. This point is feasible because

$$0 \leq \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \leq \bar{R} + \xi_i = 0 + (\xi_i + \bar{R})$$

and therefore

$$\xi_i + \bar{R} \geq 0.$$

Because $C > 1/l$ and $\bar{R} < 0$, the new objective function satisfies

$$0 + C\sum_{i=1}^{l}(\xi_i + \bar{R}) = C\sum_{i=1}^{l}\xi_i + lC\bar{R} < C\sum_{i=1}^{l}\xi_i + \bar{R}, \tag{42}$$

a contradiction to the assumption that $(\bar{R}, \boldsymbol{a}, \xi)$ is optimal. Therefore, even if the $\bar{R} \geq 0$ constraint is not explicitly stated in problem (7), it is still satisfied by any optimal solution.

For any $C \leq 1/l$, assume $(\bar{R}, \boldsymbol{a}, \boldsymbol{\xi})$ is an optimum of (7) with $\bar{R} > 0$. We consider a new point $(0, \boldsymbol{a}, \boldsymbol{\xi} + \bar{R}\boldsymbol{e})$. This point is feasible because

$$0 \leq \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \leq \bar{R} + \xi_i = 0 + (\xi_i + \bar{R})$$

and

$$\xi_i + \bar{R} \geq 0.$$

Because $C \leq 1/l$ and $\bar{R} > 0$, the new objective function satisfies

$$0 + C\sum_{i=1}^{l}(\xi_i + \bar{R}) = C\sum_{i=1}^{l}\xi_i + lC\bar{R} \leq C\sum_{i=1}^{l}\xi_i + \bar{R}. \tag{43}$$

Along with the constraint $\bar{R} \geq 0$, the new point with $\bar{R} = 0$ is optimal when $C \leq 1/l$. Furthermore, when $C < 1/l$, (43) becomes a strict inequality. This contradicts the assumption that $(\bar{R}, \boldsymbol{a}, \xi)$ is optimal for (7), so the optimal $\bar{R}$ must be zero for $C < 1/l$.

C Proof of Theorem 5

From the KKT conditions (20) and (15), at an optimum we have for all i,

$$\begin{aligned} &\bar{R} \geq \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2, \text{ if } \alpha_i < C, \\ &\bar{R} \leq \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2, \text{ if } \alpha_i > 0. \end{aligned}$$

The inequality (22) immediately follows. Note that because Slater's condition is guaranteed, KKT conditions are necessary and sufficient for optimality (Theorem 1). Thus both the if and the only if directions are true.

D Proof of Theorem 6

From (14), (15) and the constraint $\alpha_i \geq 0, \forall i$, if $C > 1$, then $\gamma_i > 0$ and the KKT optimality condition $\gamma_i\xi_i = 0$ in (20) implies that $\xi_i = 0$. Therefore, the $C\sum_{i=1}^{l}\xi_i$ term can be removed from the objective function of (7). The $\bar{R} \geq 0$ constraint is not needed because without $\boldsymbol{\xi}$, $\bar{R} \geq \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2$ has implicitly guaranteed the non-negativity of $\bar{R}$. Therefore, if $C > 1$, problems (7) and (24) are equivalent.

E Proof of Theorem 7

Because of using the strictly convex loss term $C\sum_{i=1}^{l}\xi_i^2$, the optimal $\boldsymbol{\xi}$ is unique. Then

$$\bar{R} = \text{Primal optimal value} - C\sum_{i=1}^{l}\xi_i^2$$

is unique because a convex programming problem has a unique optimal objective value.

To prove the uniqueness of the optimal $\boldsymbol{a}$, we follow a similar argument in the proof of Theorem 2. If $\boldsymbol{a}$ is not unique, there are two optimal solutions $(\bar{R}, \boldsymbol{a}_1, \boldsymbol{\xi})$ and $(\bar{R}, \boldsymbol{a}_2, \boldsymbol{\xi})$ such that

$$\boldsymbol{a}_1 \neq \boldsymbol{a}_2. \tag{44}$$

Similar to the uniqueness proof for the L1-loss case, the optimal objective value is positive.

Because $\|\phi(\boldsymbol{x}) - \boldsymbol{a}\|^2$ is strictly convex with respect to $\boldsymbol{a}$ and $\boldsymbol{a}_1 \neq \boldsymbol{a}_2$ from (44), there exists $\theta \in (0,1)$ such that for all $i = 1, \ldots, l$,

$$\begin{aligned} \|\phi(\boldsymbol{x})_i - (\theta\boldsymbol{a}_1 + (1-\theta)\boldsymbol{a}_2)\|^2 &< \theta\|\phi(\boldsymbol{x})_i - \boldsymbol{a}_1\|^2 + (1-\theta)\|\phi(\boldsymbol{x})_i - \boldsymbol{a}_2\|^2 \\ &\leq \theta(\bar{R} + \xi_i) + (1-\theta)(\bar{R} + \xi_i) \\ &= \bar{R} + \xi_i. \end{aligned} \tag{45}$$

From (45), there exists $\Delta \in (0,1)$ such that for all $i = 1, \ldots, l$,

$$\|\phi(\boldsymbol{x})_i - (\theta\boldsymbol{a}_1 + (1-\theta)\boldsymbol{a}_2)\|^2 \leq \Delta^2(\bar{R} + \xi_i) \leq \Delta^2\bar{R} + \Delta\xi_i.$$

Therefore,

$$\begin{bmatrix} \Delta^2 \bar{R} \\ \theta \boldsymbol{a}_1 + (1-\theta)\boldsymbol{a}_2 \\ \Delta \boldsymbol{\xi} \end{bmatrix}$$

is a feasible point of (27). However, the new objective value is

$$(\Delta^2 \bar{R}) + C \sum_{i=1}^{l} (\Delta \xi_i)^2 = \Delta^2 (\bar{R} + C \sum_{i=1}^{l} \xi_i).$$

Because we have shown that the optimal objective value is positive, this new value is strictly smaller than the optimal objective value we have. Thus, there is a contradiction. Therefore, our assumption in (44) is incorrect and the optimal $\boldsymbol{a}$ is unique.

F Proof of Theorem 8

First, we prove that if $C > C^*$, then the optimal $\bar{R} > 0$. If this result is wrong, there exists $C > C^*$ such that the optimal $\bar{R} = 0$. Then clearly the optimal solution is $(0, \boldsymbol{a}^*, \boldsymbol{\xi}^*)$. For any $\epsilon > 0$, $(\epsilon, \boldsymbol{a}^*, \boldsymbol{\xi}^* - \epsilon \boldsymbol{e})$ is a feasible solution of (27) because $\epsilon \geq 0$ and

$$\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}^*\| \leq \xi_i^* = \epsilon + \xi_i^* - \epsilon, i = 1, \ldots, l.$$

The objective value of (27) at this point is

$$\epsilon + C \sum_{i=1}^{l} (\xi_i^* - \epsilon)^2.$$

By Theorem 7 that the optimal solution is unique and $(0, \boldsymbol{a}^*, \boldsymbol{\xi}^*) \neq (\epsilon, \boldsymbol{a}^*, \boldsymbol{\xi}^* - \epsilon \boldsymbol{e})$, we have

$$C \sum_{i=1}^{l} (\xi_i^*)^2 < C \sum_{i=1}^{l} (\xi_i^* - \epsilon)^2 + \epsilon = C \sum_{i=1}^{l} (\xi_i^*)^2 - 2C\epsilon \sum_{i=1}^{l} \xi_i^* + Cl\epsilon^2 + \epsilon, \forall \epsilon > 0.$$

Divide both sides by ϵ and let $\epsilon \to 0$. We obtain

$$-2C \sum_{i=1}^{l} \xi_i^* + 1 \geq 0.$$

This shows that if the optimal $\bar{R}$ of (27) is zero, then with the assumption $\sum_{i=1}^{l} \xi_i^* > 0$,

$$C \leq \frac{1}{2 \sum_{i=1}^{l} \xi_i^*} = C^*,$$

a violation to the condition $C > C^*$. Therefore, the optimal $\bar{R}$ must satisfy $\bar{R} > 0$.

We now further prove that the constraint $\bar{R} \geq 0$ in (27) is not necessary. Consider (27) without the $\bar{R} \geq 0$ constraint.

$$\begin{aligned} \min_{\bar{R}, \boldsymbol{a}, \boldsymbol{\xi}} \quad & \bar{R} + C \sum_{i=1}^{l} \xi_i^2 \\ \text{subject to} \quad & \|\phi(\boldsymbol{x}_i) - \boldsymbol{a}\|^2 \leq \bar{R} + \xi_i, i = 1, \ldots, l. \end{aligned} \tag{46}$$

Assume it has an optimal solution $(\hat{R}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{\xi}})$ with $\hat{R} < 0$. Let $(\tilde{R}, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\xi}})$ be the unique optimal solution of (27). Clearly,

$$\hat{R} + C \sum_{i=1}^{l} \hat{\xi}_i^2 \leq \tilde{R} + C \sum_{i=1}^{l} \tilde{\xi}_i^2 \tag{47}$$

because $(\tilde{R}, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\xi}})$ is feasible for (46). Because the optimal $\bar{R}$ satisfies $\bar{R} > 0$ whenever $C > C^*$, consider a new point

$$(R_\theta, \boldsymbol{a}_\theta, \boldsymbol{\xi}_\theta) = \theta(\hat{R}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{\xi}}) + (1-\theta)(\tilde{R}, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\xi}})$$

such that $R_\theta = 0$ for some $\theta \in (0, 1)$. By convexity of $\| \cdot \|^2$, $(R_\theta, \boldsymbol{a}_\theta, \boldsymbol{\xi}_\theta)$ is a feasible point of (27). However, the new objective value satisfies

$$R_\theta + C\sum_{i=1}^{l}(\xi_\theta)_i^2 \leq \theta(\hat{R} + C\sum_{i=1}^{l}\hat{\xi}_i^2) + (1-\theta)(\tilde{R} + C\sum_{i=1}^{l}\tilde{\xi}_i^2) \tag{48}$$

$$\leq \tilde{R} + C\sum_{i=1}^{l}\tilde{\xi}_i^2, \tag{49}$$

where inequality (48) is from convexity and (49) is from (47). That is, $(R_\theta, \boldsymbol{a}_\theta, \boldsymbol{\xi}_\theta)$ is optimal for (27). With Theorem 7, this result contradicts the previously proven property that the optimal $\bar{R}$ is larger than zero when $C > C^*$. Thus the optimal $\bar{R}$ for (46) always satisfies $\bar{R} \geq 0$. We have thus proven the first statement of Theorem 8.

Next we prove that if $C < C^*$, then at optimum, $\bar{R} = 0$. If this result is wrong, there exists $C < C^*$ such that the optimal $(\hat{R}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{\xi}})$ has $\hat{R} > 0$. The point $(0, \boldsymbol{a}^*, \boldsymbol{\xi}^*)$ is not optimal at the current C because the optimal solution is unique according to Theorem 7. Thus

$$C\sum_{i=1}^{l}(\xi_i^*)^2 > \hat{R} + C\sum_{i=1}^{l}\hat{\xi}_i^2. \tag{50}$$

For any $\theta \in (0, 1]$, define

$$(R_\theta, \boldsymbol{a}_\theta, \boldsymbol{\xi}_\theta) \equiv \theta(\hat{R}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{\xi}}) + (1-\theta)(0, \boldsymbol{a}^*, \boldsymbol{\xi}^*).$$

By convexity of the constraints, it is a feasible point. We get

$$C\sum_{i=1}^{l}(\xi_i^*)^2 > \theta(\hat{R} + C\sum_{i=1}^{l}\hat{\xi}_i^2) + (1-\theta)(C\sum_{i=1}^{l}(\xi_i^*)^2) \geq R_\theta + C\sum_{i=1}^{l}(\xi_\theta)_i^2, \tag{51}$$

where the first inequality is from (50) and the second one is from convexity of square functions. We can easily see that $(0, \boldsymbol{a}_\theta, \boldsymbol{\xi}_\theta + \boldsymbol{e}R_\theta)$ is also feasible, and

$$C\sum_{i=1}^{l}((\xi_\theta)_i + R_\theta)^2 \geq C\sum_{i=1}^{l}(\|\phi(\boldsymbol{x}_i) - \boldsymbol{a}_\theta\|)^2 \geq C\sum_{i=1}^{l}(\xi_i^*)^2. \tag{52}$$

Combining (51) and (52), because $R_\theta > 0$, we get

$$C(2\sum_{i=1}^{l}(\xi_\theta)_i + lR_\theta) > 1. \tag{53}$$

We note that both R_θ and $\sum_{i=1}^{l}(\xi_\theta)_i$ are continuous functions of θ, and

$$\lim_{\theta \to 0} R_\theta = 0, \quad \lim_{\theta \to 0}\sum_{i=1}^{l}(\xi_\theta)_i = \sum_{i=1}^{l}\xi_i^*.$$

Therefore, with the assumption $\sum_{i=1}^{l}\xi_i^* > 0$, (53) gives

$$C \geq \lim_{\theta \to 0}\frac{1}{2\sum_{i=1}^{l}(\xi_\theta)_i + lR_\theta} = \frac{1}{2\sum_{i=1}^{l}\xi_i^*} = C^*,$$

a violation to the condition $C < C^*$. We have thus proven that if $C < C^*$, the optimal $\bar{R} = 0$.

Finally, we check the situation when $C = C^*$. We have shown that $(0, \boldsymbol{a}^*, \boldsymbol{\xi}^*)$ is the unique optimum of (27) for all $C \in (0, C^*)$. Assume that the optimal solution at $C = C^*$ is $(\hat{R}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{\xi}})$. Then we have that for any $C \in (0, C^*)$,

$$\begin{aligned} \hat{R} + C^* \sum_{i=1}^{l} \hat{\xi}_i^2 &\leq 0 + C^* \sum_{i=1}^{l} (\xi_i^*)^2, \\ 0 + C \sum_{i=1}^{l} (\xi_i^*)^2 &\leq \hat{R} + C \sum_{i=1}^{l} \hat{\xi}_i^2. \end{aligned}$$

Let $C \to C^*$, we have

$$C^* \sum_{i=1}^{l} (\xi_i^*)^2 \leq \hat{R} + C^* \sum_{i=1}^{l} \hat{\xi}_i^2 \leq C^* \sum_{i=1}^{l} (\xi_i^*)^2.$$

Thus

$$C^* \sum_{i=1}^{l} (\xi_i^*)^2 = \hat{R} + \sum_{i=1}^{l} \hat{\xi}_i^2.$$

That is, $(0, \boldsymbol{a}^*, \boldsymbol{\xi}^*)$ is an optimal solution of (27) at $C = C^*$. By Theorem 7, the optimal solution of (27) is unique at any C and thus at C^*. We then have $\hat{R} = 0$.

References

Amit Banerjee, Philippe Burlina, and Reuven Meth. Fast hyperspectral anomaly detection via SVDD. In *Proceedings of IEEE International Conference on Image Processing*. IEEE, 2007.

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

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

Hakan Cevikalp and Bill Triggs. Efficient object detection using cascades of nearest convex model classifiers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3138–3145, 2012.

Chien-Chung Chang, Hsi-Chen Tsai, and Yuh-Jye Lee. A minimum enclosing balls labeling method for support vector clustering. Technical report, National Taiwan University of Science and Technology, 2007. URL `http://dmlab8.csie.ntust.edu.tw/downloads/papers/SVC_MEB.pdf`.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Kai-Min Chung, Wei-Chun Kao, Chia-Liang Sun, Li-Lun Wang, and Chih-Jen Lin. Radius margin bounds for support vector machines with the RBF kernel. *Neural Computation*, 15: 2643–2681, 2003.

Corina Cortes and Vladimir Vapnik. Support-vector network. *Machine Learning*, 20:273–297, 1995.

Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using second order information for training SVM. *Journal of Machine Learning Research*, 6:1889–1918, 2005. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/quadworkset.pdf`.

Thorsten Joachims. Making large-scale SVM learning practical. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 169–184, Cambridge, MA, 1998. MIT Press.

Sang-Woong Lee, Jooyoung Park, and Seong-Whan Lee. Low resolution face recognition based on support vector data description. *Pattern Recognition*, 39(9):1809–1812, 2006.

Chih-Jen Lin. Formulations of support vector machines: a note from an optimization point of view. *Neural Computation*, 13(2):307–317, 2001.

Jooyoung Park, Daesung Kang, Jongho Kim, James T. Kwok, and Ivor W. Tsang. SVDD-based pattern denoising. *Neural Computation*, 19(7):1919–1938, 2007.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

David M. J. Tax and Robert P. W. Duin. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, 2:155–173, 2002.

David M. J. Tax and Robert P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

Vladimir Vapnik and Olivier Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036, 2000. URL `citeseer.nj.nec.com/vapnik99bounds.html`.

Xiaoming Wang, Fu-lai Chung, and Shitong Wang. Theoretical analysis for solution of support vector data description. *Neural Networks*, 24(4):360–369, 2011.