

Towards Convexity in Anomaly Detection: A New Formulation of SSLM with Unique Optimal Solutions

Hongying Liu

*School of Mathematical Sciences,
Beihang University
Beijing, 100191, China*

LIUHONGYING@BUAA.EDU.CN

Hao Wang*

*School of Information Science and Technology
ShanghaiTech University
Shanghai, 201210, China*

HAW309@GMAIL.COM

Haoran Chu

*School of Information Science and Technology
ShanghaiTech University
Shanghai, 201210, China*

CHUHR2023@SHANGHAITECH.EDU.CN

Yibo Wu

*School of Mathematical Sciences,
Beihang University
Beijing, 100191, China*

WYBMATH@BUAA.EDU.CN

Editor: Ji Zhu

Abstract

An unsolved issue in widely used methods such as Support Vector Data Description (SVDD) and Small Sphere and Large Margin SVM (SSLM) for anomaly detection is their nonconvexity, which hampers the analysis of optimal solutions in a manner similar to SVMs and limits their applicability in large-scale scenarios. In this paper, we introduce a novel convex SSLM formulation which has been demonstrated to revert to a convex quadratic programming problem for hyperparameter values of interest. Leveraging the convexity of our method, we derive numerous results that are unattainable with traditional nonconvex approaches. We conduct a thorough analysis of how hyperparameters influence the optimal solution, pointing out scenarios where optimal solutions can be trivially found and identifying instances of ill-posedness. Most notably, we establish connections between our method and traditional approaches, providing a clear determination of when the optimal solution is unique—a task unachievable with traditional nonconvex methods. We also derive the ν -property to elucidate the interactions between hyperparameters and the fractions of support vectors and margin errors in both positive and negative classes.

Keywords: support vector data description, small sphere and large margin, anomaly detection, ν -property, convex optimization

1. Introduction and Background

Anomaly detection, also known as outlier detection, is a crucial task in data analysis and aims to recognize whether a new observation belongs to the same distribution as existing observations. It is applicable across a wide range of domains where detecting anomalous data patterns is crucial,

including anomaly detection, quality control, fraud detection, image analysis, and remote sensing. In these applications, only large datasets of the ‘normal’ or ‘positive’ condition (also known as positive examples, labeled as $y_i = 1$) are encountered, and the goal is to classify data into this primary class while not explicitly modeling or considering other classes of ‘abnormalities’ (also known as negative examples, labeled as $y_i = -1$).

A wide array of anomaly detection methods exists, and the support vector machine (SVM) and the neural networks are the most popular tools for this purpose (Steinwart et al., 2005; Zhou et al., 2024). Several “(hyper)sphere-based” models have emerged among the various SVM-based approaches, each proving efficient and powerful in anomaly detection. These include the ball volume minimization model (Schölkopf et al., 1995), the support vector data description (SVDD) (Tax and Duin, 2004), the soft-boundary SVDD (Tax and Duin, 1999b), as well as its convexified variant (Chang et al., 2014), and the small sphere and large-margin (SSLM) approach (Wu and Ye, 2009). At the core of these methods lies the construction of a hypersphere that encapsulates normal instances while ensuring that most, if not all, of the normal data points lie within or on the boundary of the hypersphere. This often leads to quadratically constrained quadratic programming (QCQP) problems involving the radius and center point of the hypersphere. These methods, particularly SVDD and SSLM, offer numerous advantages. They exhibit robustness when dealing with high-dimensional data, can handle nonlinear relationships through kernelization, and demonstrate a strong ability to generalize well to unseen data. Furthermore, they are valued for their interpretability, as anomalies are identified based on their distance from the center of the hypersphere.

Despite their wide application and numerous advantages, these methods also come with limitations, one of which stems from their nonconvex formulation, which is the central focus of this paper. The model proposed by Schölkopf et al. (1995) assume that normal data are encapsulated within a single hypersphere, and the optimal hypersphere is simply the one with the smallest volume, resulting in a convex model. The optimal hypersphere *global* can be easily determined in this case. However, real applications rarely accommodate this assumption, and such a model may become highly sensitive when dealing with outliers, non-convex anomalies, or data with complex structures. Modern models such as SVDD and SSLM allow negative examples and the misclassification of both positive and negative examples. Although the optimal solutions of such models possess powerful interpretation and generalization abilities, they give rise to *nonconvex* QCQP problems. This is a well-known critical issue that remains unsolved, with many associated issues and limitations stemming from it.

1.1 Related Work

Many researchers have acknowledged the prevailing challenge posed by the nonconvexity issue inherent in sphere-based Support Vector Machines (SVMs). Over the past decades, several efforts have been devoted to addressing this intricate problem. Their focus has been centered on developing innovative strategies and algorithms to overcome the limitations imposed by nonconvexity, in a bid to enhance the performance and efficiency of sphere-based SVMs.

Consider the training data $\mathbf{x}_1, \dots, \mathbf{x}_\ell \in \mathcal{X}$, where for each $i \in \{1, \dots, \ell\}$ (in compact notation: $i \in [1, \ell]$) the vector \mathbf{x}_i represents the input vector, $\ell \in \mathbb{N}$ is the number of observations and \mathcal{X} is the data domain and is assumed to be a subset of \mathbb{R}^N . Denote \mathbf{a} and R for the ball center and radius, respectively. Schölkopf et al. (1995) minimized the volume R^2 of the ball while requiring

all examples to be contained in this ball $\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2, i \in [1, \ell]$. A test point is classified as normal if it is enclosed by this hypersphere and abnormal otherwise.

However, this model is very sensitive to outlier data points. If there are one or several outlier data points in the training set, a very large hypersphere will be obtained. To accommodate the existence of outliers in the training data, Tax and Duin (1999a,b) proposed soft-boundary-SVDD:

$$\begin{aligned} & \underset{\mathbf{a} \in \mathbb{R}^N, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell}{\text{minimize}} && R^2 + C \sum_{i=1}^{\ell} \xi_i \\ & \text{subject to} && \|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \xi_i \geq 0, i \in [1, \ell], \end{aligned} \quad (1.1)$$

where ξ_i is the relaxed variable and the positive parameter C controls the trade-off between the volume and the empirical errors.

The nonconvexity term R^2 in the constraints of these methods can be removed by introducing a new nonnegative variable $r = R^2$. Chang et al. (2014) pointed out that the constraint $r \geq 0$ in this problem is not necessary for $C > 1/\ell$, meaning in this case the optimal solution to the relaxation problem

$$\begin{aligned} & \underset{\mathbf{a} \in \mathbb{R}^N, r \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell}{\text{minimize}} && r + C \sum_{i=1}^{\ell} \xi_i \\ & \text{subject to} && \|\mathbf{x}_i - \mathbf{a}\|^2 \leq r + \xi_i, \xi_i \geq 0, i \in [1, \ell] \end{aligned} \quad (1.2)$$

always has $r \geq 0$. If $C \leq 1/\ell$, then $\mathbf{a}_* = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathbf{x}_i$ and $r_* = 0$ is an optimal solution for (1.2). Notice that (1.2) is a convex problem and satisfies the Slater condition, ensuring strong duality. The dual problem of (1.2) is

$$\begin{aligned} & \underset{\boldsymbol{\alpha} \in \mathbb{R}^\ell}{\text{maximize}} && \sum_{i=1}^{\ell} \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \\ & \text{subject to} && \sum_{i=1}^{\ell} \alpha_i = 1, \\ & && 0 \leq \alpha_i \leq C, i \in [1, \ell]. \end{aligned} \quad (1.3)$$

Wang et al. (2011, Theorem 2) and Chang et al. (2014, Theorem 2) showed that the optimal sphere center is always unique for the given dataset. As for the optimal ball radius, Wang et al. (2011, Theorem 3) claimed that the sphere radius r_* is not the unique optimal radius *if and only if*

$$|\mathcal{F}_{r_*}|C = 1 \quad (1.4)$$

where $\mathcal{F}_{r_*} = \{i : y_i = 1, \|\mathbf{x}_i - \mathbf{a}_*\|^2 - r_* > 0\}$ is the examples that lie outside the sphere. Furthermore, Chang et al. (2014, Theorem 5) showed that r is optimal for (1.2) if and only if

$$\max_{i: \alpha_i^* < C} \|\mathbf{x}_i - \mathbf{a}_*\|^2 \leq r \leq \min_{i: \alpha_i^* > 0} \|\mathbf{x}_i - \mathbf{a}_*\|^2. \quad (1.5)$$

Suppose $\boldsymbol{\alpha}^*$ is an optimal solution for the dual problem (1.3). Wang et al. (2011, Theorem 4) claimed that the radius of the ball is unique *if and only if* there exists $i_0 \in \{1, \dots, \ell\}$ such that $\alpha_{i_0}^* \in (0, C)$.

Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a positive definite kernel with reproducing kernel Hilbert space (RKHS) H and Φ be a feature map $\mathcal{X} \rightarrow H$. Then the inner product in H can be computed by evaluating with $k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$ and denote $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. To separate the data set from the origin, Schölkopf et al. (1999, 2001) proposed one-class support vector machine (OCSVM):

$$\begin{aligned} & \underset{\mathbf{w} \in H, \boldsymbol{\xi} \in \mathbb{R}^\ell, \rho \in \mathbb{R}}{\text{minimize}} && \frac{1}{2} \|\mathbf{w}\|^2 - \rho + \frac{1}{\nu \ell} \sum_{i=1}^{\ell} \xi_i \\ & \text{subject to} && \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle \geq \rho - \xi_i, \xi_i \geq 0, i \in [1, \ell], \end{aligned} \quad (1.6)$$

where $\nu \in (0, 1]$ is a given parameter. Note that although (1.2) is an equivalent convex formulation of (1.1), it remains the quadratic constraints.

If negative examples (objects which should be rejected) are available, they can be incorporated into the training set to improve the description. For convenience, we assume that target objects are labeled $y_i = 1$ for $i \in [1, m]$ and outlier objects are labeled $y_i = -1$ for $i \in [m + 1, \ell]$. Tax and Duin (2001, 2004) extended the SVDD to learn a tight hyperball that encloses the majority of the normal data points while excluding the anomalies:

$$\begin{aligned} & \underset{\mathbf{a} \in H, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell}{\text{minimize}} && R^2 + C \sum_{i=1}^m \xi_i + D \sum_{i=m+1}^{\ell} \xi_i \\ & \text{subject to} && \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \leq R^2 + \xi_i, \xi_i \geq 0, i \in [1, m] \\ & && \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \geq R^2 - \xi_i, \xi_i \geq 0, i \in [m + 1, \ell], \end{aligned} \quad (1.7)$$

where positive parameters C and D together control the trade-off between the volume and the empirical errors. These constraints enforce that the negative examples are located outside the hyperball, effectively pushing the boundary away from the anomalies. The addition of negative examples allows the model to better capture the boundary between normal and abnormal instances, potentially leading to improved performance in scenarios where anomalies are more diverse or prevalent.

Wu and Ye (2009) proposed SSLM approach, which aims to define a hyperball that tightly encloses the majority of normal data points while maintaining a large margin from the decision boundary to separate normal data from outliers. The idea is to create a compact representation of positive examples while maximizing the separation between the positive and negative instances to achieve robust anomaly detection performance:

$$\begin{aligned} & \underset{\mathbf{a} \in H, R, \rho \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell}{\text{minimize}} && R^2 - \bar{\nu} \rho^2 + \frac{1}{\bar{\nu}_1 m} \sum_{i=1}^m \xi_i + \frac{1}{\bar{\nu}_2 n} \sum_{i=m+1}^{\ell} \xi_i \\ & \text{subject to} && \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \leq R^2 + \xi_i, \xi_i \geq 0, i \in [1, m] \\ & && \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \geq R^2 + \rho^2 - \xi_i, \xi_i \geq 0, i \in [m + 1, \ell], \end{aligned} \quad (1.8)$$

where $n = \ell - m$, and ρ^2 is the margin between the surface of the hyperball and the outlier training data, $\boldsymbol{\xi} \in \mathbb{R}^\ell$ is the vector of slack variables, and the positive parameter $\bar{\nu}$ controls the trade-off between the volume and margin, positive $\bar{\nu}_1$ and $\bar{\nu}_2$ control the trade-off between the volume and the errors. (1.8) can result in a closed and tight boundary around the normal data. Wu and Ye (2009) given the dual problem of (1.8):

$$\begin{aligned} & \underset{\boldsymbol{\alpha} \in \mathbb{R}^\ell}{\text{maximize}} && - \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y_i y_j k_{ij} \alpha_i \alpha_j + \sum_{i=1}^{\ell} y_i k_{ii} \alpha_i \\ & \text{subject to} && \sum_{i=1}^{\ell} y_i \alpha_i = 1, \quad \sum_{i=1}^{\ell} \alpha_i = 2\bar{\nu} + 1, \\ & && 0 \leq \alpha_i \leq \frac{1}{\bar{\nu}_1 m}, i \in [1, m]; \quad 0 \leq \alpha_i \leq \frac{1}{\bar{\nu}_2 n}, i \in [m + 1, \ell]. \end{aligned} \quad (1.9)$$

The term R^2 in the constraints of (1.7) and (1.8) associated with the positive examples can be removed by introducing new nonnegative variable $r = R^2$. Similarly, the term ρ^2 in the objective of (1.8) can be removed by introducing new nonnegative variable $t = \rho^2$. While the nonconvexity term $\|\mathbf{a}\|^2$ appearing in the constraint with negative examples in (1.7) and (1.8) is inherent, which results in a nonconvex QCQP problems. It is worthwhile to ask whether a convex quadratic program (QP) formulation for (1.2), (1.7) and (1.8), as the well-known support vector machines (SVM)

$$\begin{aligned} & \underset{\mathbf{w} \in H, \boldsymbol{\xi} \in \mathbb{R}^\ell, \rho \in \mathbb{R}}{\text{minimize}} && \frac{\nu}{2} \|\mathbf{w}\|^2 + \frac{1}{\ell} \sum_{i=1}^m \xi_i + \frac{b}{\ell} \sum_{i=m+1}^{\ell} \xi_i \\ & \text{subject to} && y_i (\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, i \in [1, \ell]. \end{aligned} \quad (1.10)$$

It is worth noting that the norm in (1.6), (1.10), (1.7) and (1.8) are induced by the inner product in RKHS H for $w, a \in H$. It reduced to the Euclidean norm if $\Phi(x) = x$.

1.2 Motivation

Nonconvexity is a prevalent issue in hypersphere-based SVM models, particularly in well-known methods such as SVDD and SSLM. This poses challenges in model property analysis, hyperparameter selection, and algorithm design.

Model Property Analysis. The nonconvexity of hypersphere-based SVMs significantly complicates the analysis of their solutions, especially when compared to hyperplane-based SVMs (Vapnik, 1995; Cucker and Zhou, 2007; Steinwart and Christmann, 2008). Fundamental questions, such as whether a unique solution exists, remain unresolved. A key issue is the lack of strong duality, which is essential for working effectively in kernel space. Nonconvexity can result in an incorrect Lagrangian dual problem, maintaining a positive duality gap between the primal and dual solutions. Moreover, solving nonconvex problems to global optimality is often computationally difficult, if not impossible, and the solutions obtained by solvers are typically only local optima, limiting their usefulness for analysis.

Hyperparameter Selection. One of the most pressing issues with nonconvex models is the challenge of hyperparameter tuning. Parameters such as the sphere size, margin, and kernel function directly impact performance and generalization ability. The difficulty arises from the inability to guarantee “global” minima for each hyperparameter, as local minima make it hard to assess whether a chosen value is truly optimal. Additionally, balancing the size of the hypersphere (which represents normal data) against the margin separating outliers can be problematic, as local minima increase the sensitivity to outliers, resulting in suboptimal anomaly detection performance.

Algorithm Design. Algorithm design for hypersphere-based SVMs has been underexplored, particularly when compared to the more established hyperplane-based SVMs (1.10). For instance, Chapelle (2007) demonstrated that the primal problem of SVMs can be efficiently solved, and Shalev-Shwartz et al. (2011) proposed the Pegasos algorithm, which uses stochastic subgradient descent to solve the primal problem of SVMs. We believe that resolving the nonconvexity issue in hypersphere-based SVMs would pave the way for the development of similarly powerful algorithms.

Approximation Methods and Efficiency. Numerous studies have proposed approximation methods to address the nonconvexity of SVDD and SSLM. For example, Chen et al. (2015) approximated SVDD using a Gaussian radial basis function kernel, facilitating easier solution acquisition. Wang and Xu (2023) replaced the hinge loss in SSLM with a ramp loss, converting the nonconvex problem into a difference-of-convex form, which was then solved using the concave-convex procedure. Guo et al. (2020) incorporated incremental learning into SSLM to improve training efficiency.

Bilevel Optimization and the Challenge of Nonconvexity. Bilevel optimization, which has recently attracted significant attention as a powerful tool for hyperparameter selection and meta-learning, provides a natural framework where the upper-level problem optimizes hyperparameters while the lower-level problem corresponds to model training. However, nonconvexity—whether in the upper- or lower-level—poses severe challenges. A nonconvex upper-level objective admits multiple local minima, complicating the search for globally optimal hyperparameters. More critically, nonconvexity in the lower-level problem undermines the very foundation of bilevel methods: the solution set of the inner problem can be non-unique, discontinuous, or non-differentiable, ren-

dering implicit differentiation and hypergradient techniques inapplicable (Zhang et al., 2024). In such cases, the overall bilevel formulation is not only highly nonconvex but often computationally intractable (NP-hard), extremely sensitive to initialization, and prone to instability under standard gradient-based algorithms (Liu et al., 2022). These issues highlight that ensuring convexity—at least at the lower level—is essential for making bilevel optimization practically reliable and theoretically well-grounded. This challenge directly motivates our convex reformulation, which guarantees tractability and stability while preserving the modeling power of the bilevel framework.

Nonconvexity also hampers many techniques designed to enhance computational efficiency during training. SSLM, in particular, involves multiple variables that cannot be explicitly represented as a linear combination of training samples. As a result, researchers (Cao et al., 2020; Wang and Xu, 2022; Yuan and Xu, 2021) have focused on developing screening rules to identify and eliminate non-support vectors, reducing the size of the nonconvex problem. The first such rule, proposed by Cao et al. (2020), integrates the ν -property, KKT conditions, and variational inequalities specific to SSLM.

This ongoing research demonstrates that addressing the nonconvexity in hypersphere-based SVMs can lead to significant advances in both theoretical analysis and practical algorithm development.

1.3 Contribution

The motivations discussed above prompted us to propose a convex optimization formulation to obtain the *global* optimal solution for the SSLM with given hyperparameters. This allows us to analyze our method in a manner similar to SVMs, addressing a longstanding unresolved issue in hypersphere-based SVM models like SSLM and SVDD.

- *Convex Reformulation:* We introduce a new kernelized SSLM model that is convex for any set of hyperparameters. Specifically, we prove that for prescribed values of these hyperparameters, our model reduces to either a convex quadratic programming (QP) or linear programming (LP) problem. This ensures that the *global* optimal solution can be found efficiently using existing large-scale QP/LP solvers.
- *Hyperparameter and Solution Interaction:* We analyze the relationship between hyperparameters and the resulting optimal solutions, identifying “inappropriate” hyperparameter values that lead to ill-posed problems (where the optimal margin and objective value tend to infinity). Additionally, we highlight “trivial” hyperparameter choices for which the optimal solution can be explicitly determined. This analysis provides valuable guidance for hyperparameter selection.
- *Uniqueness and Dual Characterization:* We demonstrate the uniqueness of the optimal solution for relevant hyperparameter values and characterize these solutions using dual variables. This type of analysis is rarely possible in nonconvex models, highlighting a key advantage of our convex approach.
- *ν -Property:* We derive the ν -property of our model, describing the relationship between hyperparameters and the fractions of support vectors and margin errors in both the positive and negative classes. This property allows for flexible control of the SVM model’s complexity,

which is particularly useful in situations such as class imbalance (common in anomaly detection) or when direct control over the number of support vectors is desired.

- *Connections to Traditional Methods:* Our method establishes clear connections with traditional approaches like SVDD and SSLM, making it possible to apply the analyses and results presented here to those models. This is a significant advancement towards overcoming the limitations posed by nonconvexity in traditional hypersphere-based SVMs.
- *Numerical Performance:* Through numerical experiments, we demonstrate that our convex method consistently outperforms traditional SSLM. While the nonconvexity of SSLM can lead to convergence to non-global minima when using local optimization algorithms, our convex approach reliably converges to the global minima, ensuring both efficiency and accuracy.

1.4 Outline

In Section 2, we introduce the convex SSLM model, and analyze the influence of hyperparameters on the optimal model. In Section 3, the connection of our proposed model to traditional SVDD and SSLM models is established. Section 4 presents the derivation of the Lagrangian dual problem for the proposed model, proves the uniqueness of optimal solutions for relevant hyperparameter values, and characterizes the optimal solution using dual variables, highlighting the ν -property of our convex QP formulation. In Section 5, we provide experimental results that demonstrate the advantages of our method over traditional SSLM. Finally, Section 6 concludes the paper.

2. The Convex SSLM

In this section, we present our *convex* SSLM (CSSLM) on the data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ with $y_i = 1, i \in [1, m]$ and $y_i = -1, i \in [m+1, \ell]$ with $n = \ell - m$. The m positive examples are required to be contained in the small hypersphere with radius $r \geq 0$,

$$\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \leq r, \quad y_i = 1.$$

Let $t \geq 0$ be the margin between the two spheres. The n negative examples are excluded from the hypersphere

$$\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \geq r + t, \quad y_i = -1.$$

The empirical risk is $(\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - r)_+$ if $y_i = 1$ and $b(r + t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2)_+$ if $y_i = -1$ with $(z)_+ := \max\{z, 0\}$. By minimizing the small sphere and the empirical risk, we arrive at the regularized empirical risk minimization formulation of CSSLM

$$\inf_{\mathbf{a} \in H, r \geq 0, t \geq 0} g(\mathbf{a}, r, t) \quad (\mathcal{F})$$

with

$$g(\mathbf{a}, r, t) = \frac{1}{2}(\nu r - \mu t) + \frac{1}{2\ell} \sum_{i=1}^m (\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - r)_+ + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (r + t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2)_+.$$

Here prescribed $\nu > 0$ controls the trade-off between the volume and the loss, $\mu \geq 0$ controls the trade-off between the volume and the margin, and $b > 0$ controls the trade-off between the loss by positive examples and that by negative examples.

Before proceeding to propose the CSSLM based on (\mathcal{F}) , we first point out in the following (See Appendix A for the proof) that the objective of (\mathcal{F}) is unbounded below if the value of μ is set too large.

Lemma 1 *If $\mu > \min\{m/\ell, bn/\ell\}$, then (\mathcal{F}) is unbounded below.*

The parameter $\mu \leq \min\{m/\ell, bn/\ell\}$ is necessary to ensure that the objective remains bounded. The margin term $-\mu t/2$ decreases the objective as t grows, and the hinge losses from positive and negative examples act as a counterbalance. Positive examples can offset at most their dataset proportion m/ℓ , while weighted negatives contribute at most bn/ℓ . If μ exceeds the smaller of these, the margin term dominates, causing $t \rightarrow \infty$ and the objective to be unbounded. In the rest of this paper, we assume ν and μ are selected such that

$$\mu \leq \min\left\{\frac{m}{\ell}, \frac{bn}{\ell}\right\}. \quad (2.1)$$

Obviously, (\mathcal{F}) is a nonconvex model. We do not suggest directly solving a *nonconvex* model like (\mathcal{F}) . We show that for all values satisfying (2.1), the *global* optimal solution of (\mathcal{F}) can be found by solving a *convex* problem.

First, suppose the regularization strength is smaller than the ratio of positive examples in the dataset. Since the average slope of the hinge-loss from positive samples is m/ℓ , this balance ensures that the loss can counteract the regularization, making the problem a convex quadratic program. This condition is given as

$$\nu + \mu < m/\ell. \quad (2.2)$$

Then the global optimal solution of (\mathcal{F}) can be found by solving the problem

$$\inf_{\mathbf{a} \in H, r \in \mathbb{R}, t \geq 0} g(\mathbf{a}, r, t). \quad (2.3)$$

Compared with (\mathcal{F}) , the absence of constraint $r \geq 0$ represents a significant difference, which enables (2.3) to be convex in the absence of negative example.

Theorem 2 *If (2.2) holds, then the global optimal solution (\mathbf{a}_*, r_*, t_*) of (2.3) satisfies $r_* \geq 0$ and therefore is global optimal for (\mathcal{F}) . Conversely, every global minimizer $(\bar{\mathbf{a}}, \bar{r}, \bar{t})$ of (\mathcal{F}) is also optimal for (2.3).*

Proof Suppose (\mathbf{a}_*, r_*, t_*) is the optimal solution of (2.3). Assume by contradiction that $r_* < 0$. To derive a contradiction, we now investigate the value of g in two cases based on the value of $r_* + t_*$. For this purpose, define

$$\hat{\mathbf{a}} := \arg \min_{\mathbf{a} \in H} \frac{1}{m} \sum_{i=1}^m \|\Phi(x_i) - \mathbf{a}\|^2 = \frac{1}{m} \sum_{i=1}^m \Phi(x_i). \quad (2.4)$$

If $r_* + t_* \leq 0$, we have

$$\begin{aligned}
 g(\mathbf{a}_*, r_*, t_*) &= \frac{1}{2}(\nu + \mu - \frac{m}{\ell})r_* - \frac{\mu}{2}(r_* + t_*) + \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \\
 &\geq \frac{1}{2}(\nu + \mu - \frac{m}{\ell})r_* + \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \\
 &> \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \\
 &\geq \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \hat{\mathbf{a}}\|^2 \\
 &= g(\hat{\mathbf{a}}, 0, 0),
 \end{aligned}$$

where the first inequality follows from $r_* + t_* \leq 0$ and $\mu \geq 0$, and the strict inequality is by (2.2) and $r_* < 0$, and the last inequality is due to the fact that $\hat{\mathbf{a}}$ is the minimizer of $\sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2$ because (2.4). Obviously, $(\hat{\mathbf{a}}, 0, 0)$ is feasible for (2.3). This contradicts the optimality of (\mathbf{a}_*, r_*, t_*) .

If $r_* + t_* > 0$, we have

$$\begin{aligned}
 &g(\mathbf{a}_*, r_*, t_*) \\
 &= \frac{1}{2}(\nu + \mu - \frac{m}{\ell})r_* - \frac{\mu}{2}(r_* + t_*) + \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (r_* + t_* - \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|)_+ \\
 &> -\frac{\mu}{2}(r_* + t_*) + \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (r_* + t_* - \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|)_+ \\
 &= g(\mathbf{a}_*, 0, r_* + t_*),
 \end{aligned}$$

where the inequality follows by (2.2) and $r_* < 0$. Notice that $(\mathbf{a}_*, 0, r_* + t_*)$ is also feasible for (2.3) by $r_* + t_* > 0$. This contradicts the optimality of (\mathbf{a}_*, r_*, t_*) . Overall, we have shown that $r_* \geq 0$, so that (\mathbf{a}_*, r_*, t_*) is also feasible for (\mathcal{F}) . Hence, it is globally optimal for (\mathcal{F}) .

Conversely, let $(\bar{\mathbf{a}}, \bar{r}, \bar{t})$ be globally optimal for (\mathcal{F}) . It follows that $g(\bar{\mathbf{a}}, \bar{r}, \bar{t}) \leq g(\mathbf{a}_*, r_*, t_*)$ because (\mathbf{a}_*, r_*, t_*) is feasible for (\mathcal{F}) ; moreover, $(\bar{\mathbf{a}}, \bar{r}, \bar{t})$ is also feasible for (2.3). Hence, $(\bar{\mathbf{a}}, \bar{r}, \bar{t})$ is also optimal for (2.3). \blacksquare

Now we give the convex QP formulation of (2.3). We introduce the intermediate variable

$$s = \|\mathbf{a}\|^2 - r, \quad (2.5)$$

instead of using r as the variable, so that the empirical risk becomes $(e_i)_+$ if $y_i = 1$ and $b(t - e_i)_+$ if $y_i = -1$ with $e_i = (k_{ii} + s)/2 - \langle \mathbf{a}, \Phi(\mathbf{x}_i) \rangle$. Now (2.3) reverts to

$$\inf_{\mathbf{a} \in H, s \in \mathbb{R}, t \geq 0} f(\mathbf{a}, s, t) := \frac{\nu}{2}(\|\mathbf{a}\|^2 - s) - \frac{\mu}{2}t + \frac{1}{\ell} \sum_{i=1}^m (e_i)_+ + \frac{b}{\ell} \sum_{i=m+1}^{\ell} (t - e_i)_+. \quad (2.6)$$

The radius r of the sphere appears as a latent variable in this model, and the small sphere is attained by minimizing $\|\mathbf{a}\|^2 - s$ instead of r . By introducing the auxiliary variable ξ_i , (2.6) is indeed a

convex QP model¹:

$$\begin{aligned}
 & \underset{\mathbf{a} \in H, s, t \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell}{\text{minimize}} & h(\mathbf{a}, s, t, \boldsymbol{\xi}) & := \frac{\nu}{2}(\|\mathbf{a}\|^2 - s) - \frac{\ell\mu}{2}t + \sum_{i=1}^m \xi_i + b \sum_{i=m+1}^{\ell} \xi_i \\
 & \text{subject to} & & y_i \left[\frac{k_{ii}}{2} - \langle \Phi(\mathbf{x}_i), \mathbf{a} \rangle + \frac{s}{2} \right] - \xi_i \leq 0, \quad \xi_i \geq 0, \quad i \in [1, m] \\
 & & & y_i \left[\frac{k_{ii}}{2} - \langle \Phi(\mathbf{x}_i), \mathbf{a} \rangle + \frac{s}{2} \right] + \frac{t}{2} - \xi_i \leq 0, \quad \xi_i \geq 0, \quad i \in [m+1, \ell] \\
 & & & t \geq 0.
 \end{aligned} \tag{P}$$

It is often more convenient for SVM methods to train in the kernel space without knowing the feature mapping Φ explicitly, then transfer (P) from the kernel space to \mathbb{R}^ℓ . Then we have the convex QP problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^\ell, s \in \mathbb{R}, t \geq 0} \frac{\nu}{2}(\boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta} - s) - \frac{\mu}{2}t + \frac{1}{\ell} \sum_{i=1}^m \left(\frac{k_{ii} + s}{2} - \sum_{j=1}^{\ell} \beta_j k_{ij} \right)_+ + \frac{b}{\ell} \sum_{i=m+1}^{\ell} \left(\frac{t}{2} - \left(\frac{k_{ii} + s}{2} - \sum_{j=1}^{\ell} \beta_j k_{ij} \right) \right)_+. \tag{K}$$

Problem (K) is suitable to incorporate with different kernel techniques. The sphere center in (K) now also becomes a hidden variable. Nevertheless, it is trivial to recover the optimal solution to (K) by the following representer theorem.

Theorem 3 *Let H be a RKHS with positive definite symmetric kernel $k(\cdot, \cdot)$. Suppose (2.2) holds. Then (\mathbf{a}_*, r_*, t_*) is optimal for (2.3) if and only if $(\boldsymbol{\beta}^*, s_*, t_*)$ is optimal for (K) with $r_* = \|\mathbf{a}_*\|^2 - s_*$ and*

$$\mathbf{a}_* = \sum_{i=1}^{\ell} \beta_i^* \Phi(\mathbf{x}_i).$$

Proof Define

$$H_\ell := \text{span}\{\Phi(\mathbf{x}_1), \dots, \Phi(\mathbf{x}_\ell)\} = \left\{ \mathbf{a} \in H : \mathbf{a}(\cdot) = \sum_{i=1}^{\ell} \beta_i k(\mathbf{x}_i, \cdot), \boldsymbol{\beta} \in \mathbb{R}^\ell \right\}.$$

We first show that if (\mathbf{a}_*, r_*, t_*) is optimal for (2.3), then $\mathbf{a}_* \in H_\ell$. By contradiction, suppose that $\mathbf{a}_* \notin H_\ell$. Since H_ℓ is a linear subspace of H , we can decompose $\mathbf{a}_* \in H$ as $\mathbf{a}_* = \bar{\mathbf{a}}_* + \mathbf{a}_*^\perp$ with $\bar{\mathbf{a}}_* \in H_\ell$ and $\mathbf{0} \neq \mathbf{a}_*^\perp \in H_\ell^\perp$ where H_ℓ^\perp is the orthogonal complements of H_ℓ . Moreover, since $\Phi(\mathbf{x}_i) \in H_\ell$, the Pythagorean theorem then gives

$$\|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 = \|\Phi(\mathbf{x}_i) - \bar{\mathbf{a}}_*\|^2 + \|\mathbf{a}_*^\perp\|^2 \text{ for } i \in [1, \ell].$$

Elementary computations show that

$$g(\mathbf{a}_*, r_*, t_*) = g(\bar{\mathbf{a}}_*, r_* - \|\mathbf{a}_*^\perp\|^2, t_*) + \frac{\nu}{2} \|\mathbf{a}_*^\perp\|^2.$$

Thus

$$g(\mathbf{a}_*, r_*, t_*) > g(\bar{\mathbf{a}}_*, r_* - \|\mathbf{a}_*^\perp\|^2, t_*)$$

because $\mathbf{a}_*^\perp \neq \mathbf{0}$. This violates the global optimality of (\mathbf{a}_*, r_*, t_*) for (2.3). Hence $\mathbf{a}_* \in H_\ell$.

1. Note that the objective function of (2.6) is a scaled one of (P) with a factor $1/\ell$.

Conversely, if (β^*, s_*, t_*) is optimal for (\mathcal{K}) , then $(\mathbf{a}_*, r_*, t_*) = (\sum_{i=1}^{\ell} \beta_i^* \Phi(\mathbf{x}_i), \|\mathbf{a}_*\|^2 - s_*, t_*)$ must be optimal for (2.3). Otherwise, there exists $(\hat{\mathbf{a}}, \hat{r}, \hat{t}) \neq (\mathbf{a}_*, r_*, t_*)$ such that $g(\hat{\mathbf{a}}, \hat{r}, \hat{t}) < g(\mathbf{a}_*, r_*, t_*)$. However, we have shown there exists a decomposition $\hat{\mathbf{a}} = \sum_{i=1}^{\ell} \beta_i \Phi(\mathbf{x}_i)$ such that $(\hat{\beta}, \|\hat{\mathbf{a}}\|_2^2 - \hat{r}, \hat{t})$ yields a smaller objective (\mathcal{K}) —a contradiction. Therefore, (\mathbf{a}_*, r_*, t_*) is optimal for (2.3). \blacksquare

As we show in the next, for other choices of parameter values ν, μ and b , although the constraint $r \geq 0$ is needed, the global optimal solution of (2.3) can still be found by solving convex problems. This result is summarized in the following theorem. The proof is provided in Appendix B.

Theorem 4 *Suppose $\nu + \mu \geq m/\ell$ is satisfied.*

(i) *If $\mu = 0$, the infimum of (\mathcal{F}) is attained at*

$$\Omega := \left\{ (\hat{\mathbf{a}}, r, t) : 0 \leq r \leq \min_{y_i=1} \{\|\Phi(\mathbf{x}_i) - \hat{\mathbf{a}}\|^2\}, r+t \leq \min_{y_i=-1} \{\|\Phi(\mathbf{x}_i) - \hat{\mathbf{a}}\|^2\}, (\nu - m/\ell)r = 0, t \geq 0 \right\}.$$

In particular, $r_ = 0, t_* = \min_{y_i=-1} \|\Phi(\mathbf{x}_i) - \hat{\mathbf{a}}\|^2, \mathbf{a}_* = \hat{\mathbf{a}}$ is an optimal solution of (\mathcal{F}) .*

(ii) *If $0 < \mu < m/\ell$, (\mathbf{a}_*, z_*) is optimal for*

$$\begin{aligned} & \underset{\mathbf{a} \in H, z \in \mathbb{R}, \xi \in \mathbb{R}^n}{\text{minimize}} && \frac{\ell\lambda}{2} \|\mathbf{a}\|^2 + \frac{\ell\mu}{2} z - m \langle \hat{\mathbf{a}}, \mathbf{a} \rangle + b \sum_{i=m+1}^{\ell} \xi_i \\ & \text{subject to} && -\frac{k_{ii}}{2} + \langle \Phi(\mathbf{x}_i), \mathbf{a} \rangle - \frac{z}{2} - \xi_i \leq 0, \xi_i \geq 0, i \in [m+1, \ell] \end{aligned} \quad (2.7)$$

with $\lambda = m/\ell - \mu$, then the solution satisfies $\|\mathbf{a}_\|^2 \geq z_*$, and $(\mathbf{a}_*, 0, \|\mathbf{a}_*\|^2 - z_*)$ is optimal for (\mathcal{F}) . Conversely, if (\mathbf{a}_*, r_*, t_*) is optimal for (\mathcal{F}) , $(\mathbf{a}_*, \|\mathbf{a}_*\|^2 - (r_* + t_*))$ is optimal for (2.7).*

(iii) *If $\mu = m/\ell$, (\mathbf{a}_*, z_*) is optimal for*

$$\begin{aligned} & \underset{\mathbf{a} \in H, z \in \mathbb{R}, \xi \in \mathbb{R}^n}{\text{minimize}} && \frac{\ell\mu}{2} z - m \langle \hat{\mathbf{a}}, \mathbf{a} \rangle + b \sum_{i=m+1}^{\ell} \xi_i \\ & \text{subject to} && -\frac{k_{ii}}{2} + \langle \Phi(\mathbf{x}_i), \mathbf{a} \rangle - \frac{z}{2} - \xi_i \leq 0, \xi_i \geq 0, i \in [m+1, \ell], \end{aligned} \quad (2.8)$$

then the solution satisfies $\|\mathbf{a}_\|^2 \geq z_*$, and $(\mathbf{a}_*, 0, \|\mathbf{a}_*\|^2 - z_*)$ is optimal for (\mathcal{F}) . Conversely, if (\mathbf{a}_*, r_*, t_*) is optimal for (\mathcal{F}) , $(\mathbf{a}_*, \|\mathbf{a}_*\|^2 - (r_* + t_*))$ is optimal for (2.8).*

The margin parameter μ must be chosen carefully to balance the effects of the margin and hinge-loss terms. First, $\nu + \mu \geq m/\ell$ ensures the combined effect of volume and margin terms can be supported by the positive examples, preventing the objective from decreasing without bound. Second, $0 < \mu < m/\ell$ guarantees that the margin term contributes to discrimination while remaining small enough for the positive hinge losses to counterbalance it. Finally, $\mu = m/\ell$ represents a critical point where the margin term is exactly balanced by the positive examples' contribution, marking the boundary beyond which the objective would become unbounded.

The entire scheme of the proposed CSSLM is described in the flowchart in Figure 1. The procedure starts with choosing the appropriate parameter values of μ, ν and then decides which convex problems to solve, or the solution is trivially obtained. Figure 2 depicts the optimal r_* and t_* of (\mathcal{F}) dependent on different choices of μ and ν .

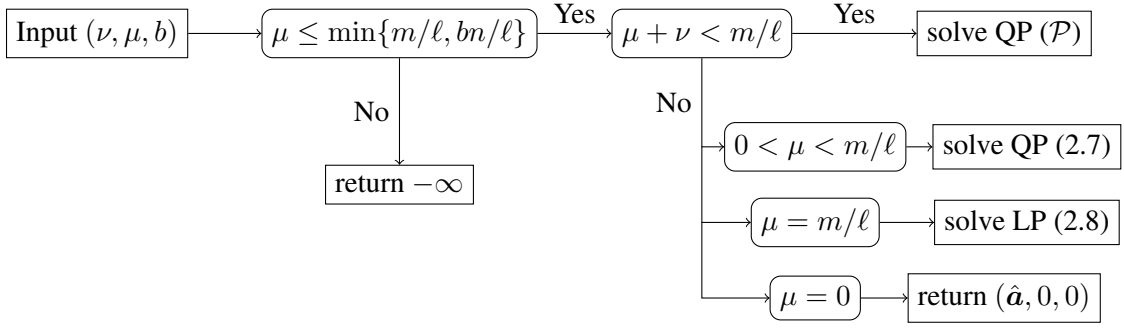
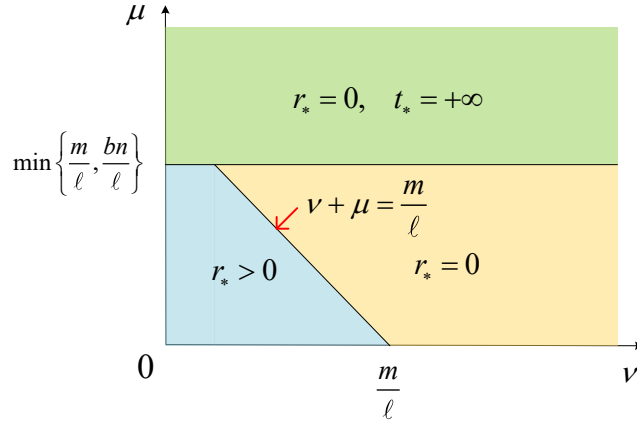


Figure 1: The flowchart of CSSLM.


 Figure 2: the optimal r_* and t_* of (\mathcal{F}) dependent on different choices of μ and ν .

3. Connection to SSLM and SVDD

We have proved that our proposed method can find the *global* minimizer (\mathbf{a}_*, r_*, t_*) of (\mathcal{F}) for different scenarios of parameters ν , μ and b . Now, we show our proposed model can easily render the *global* optimal solutions of SSLM (1.8) and SVDD (including the one-class soft-margin SVDD (1.1) and the two-class soft-margin SVDD (1.7)).

Letting $r = R^2$ (or, $R^2 = \|\mathbf{a}\|^2 - s$ in $f(\mathbf{a}, s, t)$) and $t = \rho^2$, (\mathcal{F}) reverts to

$$\min_{\mathbf{a} \in H, R, \rho \in \mathbb{R}} \frac{\nu}{2} \left(R^2 - \frac{\mu}{\nu} \rho^2 + \frac{1}{\nu \ell} \sum_{i=1}^m (d_i^2 - R^2)_+ + \frac{b}{\nu \ell} \sum_{i=m+1}^{\ell} (R^2 + \rho^2 - d_i^2)_+ \right) \quad (3.1)$$

with $d_i = \|\Phi(\mathbf{x}_i) - \mathbf{a}\|$. It is the unconstrained version of the SSLM problem (1.8) with parameters chosen as

$$(\bar{\nu}, \bar{\nu}_1, \bar{\nu}_2) = \left(\frac{\mu}{\nu}, \frac{\ell \nu}{m}, \frac{\ell \nu}{nb} \right) \quad (3.2)$$

where the objective of (3.1) is that of (1.8) scaled by $\nu/2$. Since (3.1) and (1.8) share the same global minimizers, then $(\mathbf{a}_*, R_*, \rho_*)$ is the *global* minimizer of (1.8) with

$$R_* = \sqrt{r_*} \text{ and } \rho_* = \sqrt{t_*}. \quad (3.3)$$

On the other hand, setting $\mu = 0, C = 1/(\nu\ell)$ and $D = b/(\nu\ell)$ in (3.1) yields

$$\min_{\mathbf{a} \in H, R, \rho \in \mathbb{R}} \hat{g}(\mathbf{a}, R, \rho) := \frac{\nu}{2} \left(R^2 + C \sum_{i=1}^m (d_i^2 - R^2)_+ + D \sum_{i=m+1}^{\ell} (R^2 + \rho^2 - d_i^2)_+ \right). \quad (3.4)$$

If $m = \ell$, then the last summation in the objective function of (3.4) disappears and (3.4) is obviously equivalent to the one-class soft-margin SVDD problem (1.1). Hence, $(\mathbf{a}_*, \sqrt{r_*})$ is the *global* minimizer of (1.1). If $m < \ell$, we show in the following theorem that (3.4) is equivalent to the unconstrained version of the two-class soft-margin SVDD (1.7), which can be rewritten as

$$\min_{\mathbf{a} \in H, R \in \mathbb{R}} \hat{g}(\mathbf{a}, R, 0). \quad (3.5)$$

Theorem 5 *Suppose $\mu = 0$ and $m < \ell$. For any $(\mathbf{a}_*, R_*, \rho_*)$ optimal for (3.4), (\mathbf{a}_*, R_*) is globally optimal for (3.5). Conversely, for any (\mathbf{a}_*, R_*) optimal for (3.5), $(\mathbf{a}_*, R_*, 0)$ is globally optimal for (3.4).*

Proof For any optimal solution $(\mathbf{a}_*, R_*, \rho_*)$ for (3.4), $\hat{g}(\mathbf{a}_*, R_*, \rho_*) \geq \hat{g}(\mathbf{a}_*, R_*, 0)$ since \hat{g} is increasing in $|\rho|$. Therefore, $(\mathbf{a}_*, R_*, 0)$ is also optimal for (3.4). It follows that $\hat{g}(\mathbf{a}, R, 0) \geq \hat{g}(\mathbf{a}_*, R_*, 0)$ for any (\mathbf{a}, R) , meaning (\mathbf{a}_*, R_*) is global optimal for (3.5).

Conversely, if (\mathbf{a}_*, R_*) is optimal for (3.5), then

$$\hat{g}(\mathbf{a}_*, R_*, 0) \leq \hat{g}(\mathbf{a}, R, 0) \leq \hat{g}(\mathbf{a}, R, \rho)$$

for any (\mathbf{a}, R, ρ) . Therefore, $(\mathbf{a}_*, R_*, 0)$ is optimal for (3.4). ■

Theorem 2 and Theorem 5 imply that when setting $\mu = 0, m < \ell$ and $\nu < m/\ell$, our proposed CSSLM renders the *global* minimizer $(\mathbf{a}_*, \sqrt{r_*})$ of (1.7) by obtaining an optimal (\mathbf{a}_*, r_*, t_*) to (2.3).

Remark 6 *Theorem 5 does not imply that any optimal ρ_* for (3.4) must be zero. In fact, if there exists $i_* \in [m+1, \ell]$ such that the optimal $(\mathbf{a}_*, R_*, \rho_*)$ for (3.4) satisfies*

$$\|\Phi(\mathbf{x}_{i_*}) - \mathbf{a}_*\| \leq R_*, \quad (3.6)$$

then $\hat{g}(\mathbf{a}_, R_*, 0) < \hat{g}(\mathbf{a}_*, R_*, \rho)$ for any $\rho > 0$; therefore, the optimal solution satisfies $\rho_* = 0$. On the other hand, if (3.6) is not satisfied, meaning*

$$\|\Phi(\mathbf{x}_i) - \mathbf{a}_*\| > R_*, \quad \text{for any } i \in [m+1, \ell],$$

then $(\mathbf{a}_, R_*, \rho_*)$ is globally optimal for (3.4) for any $\rho_* \in [0, \hat{\rho}]$ with $\hat{\rho} := \min_{i \in [m+1, \ell]} \{\|\Phi(\mathbf{x}_{i_*}) - \mathbf{a}_*\| - R_*\}$.*

We summarize the connections between our proposed CSSLM and the state-of-the-art SSLM (1.8), one-class SVDD (1.1) and two-class SVDD (1.7) in Figure 3. Therefore, the global minimizers of SSLM and SVDD can be found by applying polynomial algorithms to solve CSSLM.

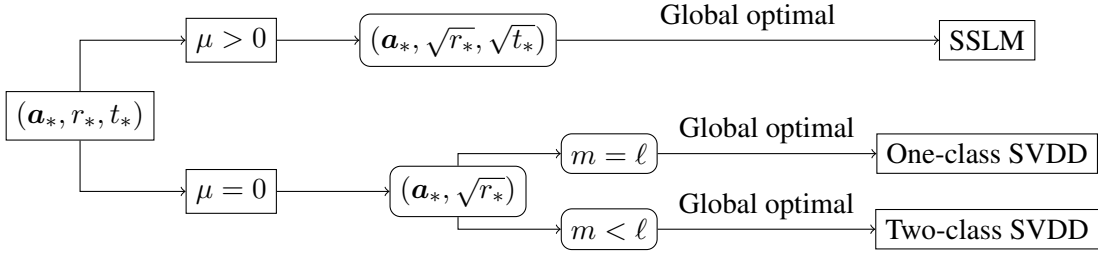


Figure 3: Connection to SSLM and SVDD

4. Existence and Uniqueness of Optimal Solutions

In this section, we state and prove our main results on the existence and uniqueness of optimal solutions of the problems (\mathcal{P}) , (2.7) and (2.8) involved in our proposed model.

4.1 Existence of Optimal Solutions

One possible and rather common approach for solving the convex optimization problem is to solve its dual problem. Elementary computation show that the Lagrangian dual of (\mathcal{P}) admits an explicit form, i.e.

$$\begin{aligned}
 & \underset{\alpha \in \mathbb{R}^\ell}{\text{maximize}} && -\frac{1}{2\ell\nu} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y_i y_j k_{ij} \alpha_i \alpha_j + \frac{1}{2} \sum_{i=1}^{\ell} y_i k_{ii} \alpha_i \\
 & \text{subject to} && \sum_{i=1}^{\ell} y_i \alpha_i = \ell\nu, \quad \sum_{i=m+1}^{\ell} \alpha_i \geq \ell\mu, \\
 & && 0 \leq \alpha_i \leq 1, i \in [1, m]; \quad 0 \leq \alpha_i \leq b, i \in [m+1, \ell].
 \end{aligned} \tag{D}$$

Theorem 7 *Suppose (2.2) holds. Problem (\mathcal{P}) and its dual problem (\mathcal{D}) are solvable and the strong duality holds at the primal-dual optimal solution. In particular, $(\mathbf{a}, s, t, \boldsymbol{\xi}, \boldsymbol{\alpha})$ is the primal-dual optimal solution if and only if it satisfies*

$$\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - r - 2\xi_i \leq 0, \xi_i \geq 0, i \in [1, m], \tag{4.1a}$$

$$q - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - 2\xi_i \leq 0, \xi_i \geq 0, i \in [m+1, \ell], \tag{4.1b}$$

$$\mathbf{a} = \frac{1}{\ell\nu} \sum_{i=1}^{\ell} y_i \alpha_i \Phi(\mathbf{x}_i), \tag{4.1c}$$

$$\sum_{i=1}^{\ell} y_i \alpha_i = \ell\nu, \quad \sum_{i=m+1}^{\ell} \alpha_i \geq \ell\mu, \tag{4.1d}$$

$$0 \leq \alpha_i \leq 1, i \in [1, m], \quad 0 \leq \alpha_i \leq b, i \in [m+1, \ell], \tag{4.1e}$$

$$\xi_i(1 - \alpha_i) = 0, i \in [1, m], \quad \xi_i(b - \alpha_i) = 0, i \in [m+1, \ell], \tag{4.1f}$$

$$(\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - r - 2\xi_i)\alpha_i = 0, i \in [1, m], \tag{4.1g}$$

$$(q - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - 2\xi_i)\alpha_i = 0, i \in [m+1, \ell], \tag{4.1h}$$

$$t \geq 0, t \left(\sum_{i=m+1}^{\ell} \alpha_i - \ell\mu \right) = 0 \tag{4.1i}$$

with $r = \|\mathbf{a}\|^2 - s$ and $q = r + t$.

Proof Problem (\mathcal{P}) is clearly feasible and convex. The dual problem (\mathcal{D}) is also a convex quadratic program. Moreover, define $\alpha_i = \ell(\nu + \mu)/m$ if $i \in [1, m]$; Otherwise $\alpha_i = \ell\mu/n$. Then this α is feasible to (\mathcal{D}) since (2.1) and (2.2). Owing to the well-known Frank-Wolfe theorem (Frank and Wolfe, 1956), both the quadratic programming problem (\mathcal{P}) as well as its dual problem (\mathcal{D}) are solvable. By Bertsekas (1999, Proposition 3.4.2(a)), strong duality holds at the optimal primal-dual solution. The necessary and sufficient optimality conditions can be derived by applying the Karush-Kuhn-Tucker optimality conditions to (\mathcal{P}) . \blacksquare

As for problem (2.7), its Lagrangian dual problem admits the following explicit form

$$\begin{aligned} & \underset{\alpha \in \mathbb{R}^n}{\text{maximize}} && -\frac{1}{2\ell\lambda} \sum_{i,j=m+1}^{\ell} k_{ij}\alpha_i\alpha_j - \frac{1}{2} \sum_{i=m+1}^{\ell} \left(k_{ii} - \frac{2}{\ell\lambda} \sum_{j=1}^m k_{ji} \right) \alpha_i - \frac{1}{2\ell\lambda} \sum_{i,j=1}^m k_{ij} \\ & \text{subject to} && \sum_{i=m+1}^{\ell} \alpha_i = \ell\mu, \\ & && 0 \leq \alpha_i \leq b, i \in [m+1, \ell]. \end{aligned} \quad (4.2)$$

We show in the next theorem that the choice of parameters guarantee that the optimal solutions of (2.7) always exist.

Theorem 8 *Suppose $\mu + \nu \geq m/\ell$ and $\mu < m/\ell$ holds. Problem (2.7) and its dual problem (4.2) are solvable and the strong duality holds at the primal-dual optimal solution. In particular, $(\mathbf{a}, z, \xi, \alpha)$ is the primal-dual optimal solution if and only if it satisfies*

$$\mathbf{a} = \frac{1}{\ell\lambda} \left[\sum_{i=1}^m \Phi(\mathbf{x}_i) - \sum_{i=m+1}^{\ell} \alpha_i \Phi(\mathbf{x}_i) \right], \quad (4.3a)$$

$$\sum_{i=m+1}^{\ell} \alpha_i = \ell\mu, \quad (4.3b)$$

$$t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - 2\xi_i \leq 0, \quad \xi_i \geq 0, \quad i \in [m+1, \ell], \quad (4.3c)$$

$$\xi_i(b - \alpha_i) = 0, \quad 0 \leq \alpha_i \leq b, \quad i \in [m+1, \ell], \quad (4.3d)$$

$$(t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - 2\xi_i)\alpha_i = 0, \quad i \in [m+1, \ell] \quad (4.3e)$$

with $t = \|\mathbf{a}\|^2 - z$.

Proof First of all, $\lambda = m/\ell - \mu > 0$ implies (2.7) is well defined and convex since $\mu < m/\ell$. Moreover, define $\alpha_i = \ell\mu/n, i \in [m+1, \ell]$. Then this α is feasible to (4.2) since (2.1). Owing to the well-known Frank-Wolfe theorem (Frank and Wolfe, 1956), both the quadratic programming problem (2.7) as well as its dual problem (4.2) are solvable. By Bertsekas (1999, Proposition 3.4.2(a)), strong duality holds at the optimal primal-dual solution. The necessary and sufficient optimality conditions can be derived by applying the Karush-Kuhn-Tucker optimality conditions to (2.7). \blacksquare

As for problem (2.8), its Lagrangian dual problem is given by

$$\begin{aligned}
 & \underset{\alpha \in \mathbb{R}^n}{\text{minimize}} && \sum_{i=m+1}^{\ell} k_{ii} \alpha_i \\
 & \text{subject to} && \sum_{i=m+1}^{\ell} \alpha_i = \ell \mu \\
 & && 0 \leq \alpha_i \leq b, \quad i \in [m+1, \ell] \\
 & && \sum_{i=m+1}^{\ell} \alpha_i \Phi(\mathbf{x}_i) = \sum_{i=1}^m \Phi(\mathbf{x}_i).
 \end{aligned} \tag{4.4}$$

It is trivial that (2.8) is feasible. If the dual problem (4.4) is infeasible, (2.8) is unbounded below, neither does problem (\mathcal{F}) . A necessary condition for (4.4) to be feasible is $\mu/b \leq n/\ell$. However, we show this condition is not sufficient for the existence of the optimal solution to (2.8) with a counterexample.

Example 1 Let $m = n = 1, \ell = 2, \mu = 1/2, b = 1, \mathbf{x}_1 = (1, 0)$ with label $y_1 = 1$ and $\mathbf{x}_2 = (0, 1)$ with label $y_2 = -1$. Considering the linear kernel $\Phi(\mathbf{x}) = \mathbf{x}$, the dual problem (4.4) is infeasible in this case since $(0, \alpha_2) \neq (1, 0)$ for any $\alpha_2 \in [0, 1]$. Therefore, the corresponding (2.8) is unbounded below.

4.2 Uniqueness of the Optimal Sphere Center

In this subsection, we characterize the optimal solution of the proposed (\mathcal{P}) . In particular, we show that the global optimal solution is unique. Such a result represents the stark contrast of (\mathcal{P}) , and is unrealistic to obtain for traditional methods such as (1.7) and (1.8) due to the lack of convexity.

First of all, the optimal center of the sphere is unique, since (\mathcal{P}) and (2.7) is partially strictly convex with respect to \mathbf{a} .

Theorem 9 *The optimal \mathbf{a}_* for (\mathcal{P}) and (2.7) is unique if $m/\ell > \mu$.*

Proof It suffices to show that for any optimal solutions (\mathbf{a}', s', t') and (\mathbf{a}'', s'', t'') for (2.6), it holds that $\mathbf{a}' = \mathbf{a}''$. Suppose by contradiction that $\mathbf{a}' \neq \mathbf{a}''$. By convexity,

$$\mathbf{u} := \frac{1}{2}(\mathbf{a}', s', t') + \frac{1}{2}(\mathbf{a}'', s'', t'')$$

is also optimal for (2.6). It follows from $\|\frac{1}{2}\mathbf{a}' + \frac{1}{2}\mathbf{a}''\|^2 < \frac{1}{2}\|\mathbf{a}'\|^2 + \frac{1}{2}\|\mathbf{a}''\|^2$ that

$$f(\mathbf{u}) < \frac{1}{2}f(\mathbf{a}', s', t') + \frac{1}{2}f(\mathbf{a}'', s'', t'') = f(\mathbf{a}', s', t'),$$

contradicting with the optimality of (\mathbf{a}', s', t') . Therefore, the optimal \mathbf{a}_* for (\mathcal{P}) is unique. It follows from the same argument that \mathbf{a}_* is also unique for (2.7) if $m/\ell > \mu$. \blacksquare

4.3 Uniqueness of the Optimal Radius and Margin

Since the sphere center is unique for (\mathcal{P}) , now we fix the sphere center and investigate the uniqueness of the radius and margin. For this purpose, notice that the strong duality holds for (\mathcal{P}) and (\mathcal{D}) .

A direct result from this property is that for any optimal primal variable $(\mathbf{a}, s, t, \boldsymbol{\xi})$ for (\mathcal{P}) and any optimal dual variable $\boldsymbol{\alpha}$ for (\mathcal{D}) , $(\mathbf{a}, s, t, \boldsymbol{\xi}, \boldsymbol{\alpha})$ together satisfies the KKT conditions (4.1). Therefore, our analysis is based on the optimality conditions (4.1) and the primal optimal solution is characterized via any given optimal dual variable.

Let $\boldsymbol{\alpha}^*$ be an optimal solution to the dual problem (\mathcal{D}) . Define the following index sets

$$\begin{aligned}\mathcal{T}_+ &:= \{i : y_i = 1, 0 \leq \alpha_i^* < 1\}, & \mathcal{T}_- &:= \{i : y_i = -1, 0 \leq \alpha_i^* < b\}, \\ \mathcal{SV}_+ &:= \{i : y_i = 1, 0 < \alpha_i^* \leq 1\}, & \mathcal{SV}_- &:= \{i : y_i = -1, 0 < \alpha_i^* \leq b\}.\end{aligned}$$

Proposition 10 *Suppose (2.2) holds. Let $\boldsymbol{\alpha}^*$ be an optimal solution to the dual problem (\mathcal{D}) . The unique optimal sphere center for (\mathcal{P}) is*

$$\mathbf{a}_* = \frac{1}{\ell\nu} \sum_{i=1}^{\ell} y_i \alpha_i^* \Phi(\mathbf{x}_i). \quad (4.5)$$

The optimal radius and margin for (\mathcal{P}) is

$$\Gamma_* = \{(r, t) : r_l \leq r \leq r_u, q_l \leq r + t \leq q_u, t \geq 0, t(\sum_{i=m+1}^{\ell} \alpha_i^* - \ell\mu) = 0\} \quad (4.6)$$

with $r_l := \max_{i \in \mathcal{T}_+} \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2$, $r_u := \min_{i \in \mathcal{SV}_+} \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2$, $q_l := \max_{i \in \mathcal{SV}_-} \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2$ and $q_u := \min_{i \in \mathcal{T}_-} \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2$, where by custom $\max \emptyset = -\infty$ and $\min \emptyset = \infty$.

Proof By duality, the primal-dual pair $(\mathbf{a}, s, t, \boldsymbol{\xi}, \boldsymbol{\alpha}^*)$ is optimal if and only if it satisfies the optimality conditions in (4.1) by Theorem 7. Hence, the vector \mathbf{a}_* in (4.5) is optimal and unique due to (4.1c) and Theorem 9.

Since \mathbf{a}_* is unique, we can treat it as fixed value and let $r = \|\mathbf{a}_*\|^2 - s$ and $q = r + t$. The complementary slackness condition (4.1f) implies that $\xi_i = 0$ for $i \in \mathcal{T}_+ \cup \mathcal{T}_-$, meaning $\mathbf{x}_i, i \in \mathcal{T}_+ \cup \mathcal{T}_-$ are separated correctly by the sphere. The complementary slackness conditions (4.1g) and (4.1h) imply $\|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 - r - 2\xi_i = 0, i \in \mathcal{SV}_+$ and $q - \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 - 2\xi_i = 0, i \in \mathcal{SV}_-$. Combining those conditions with the primal feasible conditions (4.1a) and (4.1b), (r, t) are optimal if and only if $t \geq 0$ and the following conditions are satisfied:

$$\begin{aligned}r &\geq \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \text{ with } \xi_i = 0, i \in \mathcal{T}_+, \\ r &\leq \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \text{ with } 2\xi_i = \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 - r, i \in \mathcal{SV}_+, \\ q &\leq \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \text{ with } \xi_i = 0, i \in \mathcal{T}_-, \\ q &\geq \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \text{ with } 2\xi_i = q - \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2, i \in \mathcal{SV}_-.\end{aligned}$$

Then (r, t) is optimal if and only if $(r, t) \in \Gamma_*$. ■

Remark 11 *Note that r always lies in a nonnegative finite interval. First, the dual feasibility $\sum_{i=1}^{\ell} y_i \alpha_i^* = \ell\nu$ and $\nu > 0$ imply $\mathcal{SV}_+ \neq \emptyset$. Hence $r_u \geq 0$. Secondly, if $\mathcal{T}_+ \neq \emptyset$ then $r_l \geq 0$. Otherwise, $\alpha_i^* = 1$ for all $y_i = 1$. By the equality in (4.1d) one has $\sum_{i=m+1}^{\ell} \alpha_i^* = m - \ell\nu$. Combining it with (2.2), $\sum_{i=m+1}^{\ell} \alpha_i^* > \ell\mu$, we have $\mathcal{SV}_- \neq \emptyset$ and $t_* = 0$ by the complementarity in (4.1i). Hence $q_l \geq 0$ and $r \geq \max\{r_l, q_l\}$.*

We can then use the above characterization of (r, t) to derive the following results on the uniqueness of (r, t) .

Theorem 12 *Suppose (2.2) holds. If there exist $i_+ \in [1, m]$ with $\alpha_{i_+}^* \in (0, 1)$ and $i_- \in [m+1, \ell]$ with $\alpha_{i_-}^* \in (0, b)$, then $r_* = \|\Phi(\mathbf{x}_{i_+}) - \mathbf{a}_*\|^2$ and $t_* = \|\Phi(\mathbf{x}_{i_-}) - \mathbf{a}_*\|^2 - \|\Phi(\mathbf{x}_{i_+}) - \mathbf{a}_*\|^2$ are the unique optimal radius and margin for (\mathcal{P}) , respectively. If $\sum_{i=m+1}^{\ell} \alpha_i^* > \ell\mu$ then the optimal margin is 0 and the optimal radii set is*

$$[\max\{r_l, q_l\}, \min\{r_u, q_u\}]. \quad (4.7)$$

In particular, if there exist $i_ \in [1, m]$ with $\alpha_{i_*}^* \in (0, 1)$ or $i_* \in [m+1, \ell]$ with $\alpha_{i_*}^* \in (0, b)$, then $r_* = \|\Phi(\mathbf{x}_{i_*}) - \mathbf{a}_*\|^2$ is the unique optimal radius for (\mathcal{P}) .*

Proof Under the assumption, one has $i_+ \in \mathcal{T}_+ \cap \mathcal{SV}_+$ and $i_- \in \mathcal{T}_- \cap \mathcal{SV}_-$. Then the optimal radii and margin for (\mathcal{P}) in (4.6) reduces to the single point

$$\Gamma_* = \{(\|\Phi(\mathbf{x}_{i_+}) - \mathbf{a}_*\|^2, \|\Phi(\mathbf{x}_{i_-}) - \mathbf{a}_*\|^2 - \|\Phi(\mathbf{x}_{i_+}) - \mathbf{a}_*\|^2)\}.$$

If $\sum_{i=m+1}^{\ell} \alpha_i^* > \ell\mu$ then the optimal margin is 0 by the complementary condition (4.1i). Bring $t = 0$ into (4.6), one has the desired optimal radii set (4.7).

If there exist $i_* \in [1, m]$ with $\alpha_{i_*}^* \in (0, 1)$ then $i_* \in \mathcal{T}_+ \cap \mathcal{SV}_+$. Hence the interval in (4.7) reduces to the point $\|\Phi(\mathbf{x}_{i_*}) - \mathbf{a}_*\|^2$. Similarly if there exists $i_* \in [m+1, \ell]$ with $\alpha_{i_*}^* \in (0, b)$, then $i_* \in \mathcal{T}_- \cap \mathcal{SV}_-$. Hence the interval in (4.7) also reduces to the point $\|\Phi(\mathbf{x}_{i_*}) - \mathbf{a}_*\|^2$. This completes the proof. \blacksquare

As for problem (2.7), we can also have the following similar result.

Proposition 13 *Suppose $0 < \mu < m/\ell$ and $\mu + \nu \geq m/\ell$ hold. Let α^* be an optimal solution to the dual problem (4.2). The unique optimal sphere center for (2.7) is*

$$\mathbf{a}_* = \frac{1}{\ell\lambda} \left(\sum_{i=1}^m \Phi(\mathbf{x}_i) - \sum_{i=m+1}^{\ell} \alpha_i^* \Phi(\mathbf{x}_i) \right), \quad (4.8)$$

and the optimal radius is $r_ = 0$. If $\sum_{i=1}^{\ell} \alpha_i^* > \ell\mu$ then the optimal margin is 0. Otherwise, the optimal margin set is*

$$T_* = \{t : \max_{i \in \mathcal{SV}_-} \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \leq t \leq \min_{i \in \mathcal{T}_-} \|\Phi(\mathbf{x}_i) - \mathbf{a}_*\|^2\}. \quad (4.9)$$

If there exist $i_- \in [m+1, \ell]$ with $\alpha_{i_-}^ \in (0, b)$, then $t_* = \|\Phi(\mathbf{x}_{i_-}) - \mathbf{a}_*\|^2$ is the unique optimal margin for (2.7).*

Proof First of all, Theorem 8 and Theorem 4 implies (2.7) is solvable and the optimal radius is $r_* = 0$. By duality, the primal-dual pair $(\mathbf{a}_*, t, \xi, \alpha^*)$ is optimal if and only if it satisfies the optimality conditions in (4.3). Hence, the vector \mathbf{a}_* in (4.8) is optimal and unique due to Theorem 9.

Since \mathbf{a}_* is given and $r_* = 0$, it follows from the optimality condition in (4.3d) and (4.3e) that $\xi_i = 0$ and $\|\phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \leq t$ for any $i \in \mathcal{T}_-$ and $\|\phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 - t - 2\xi_i = 0$ for any $i \in \mathcal{SV}_-$. Therefore, t is optimal if and only if $t \geq 0$ and the following conditions are satisfied:

$$t \leq \|\phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \text{ with } \xi_i = 0, \quad i \in \mathcal{T}_-, \quad (4.10)$$

$$t \geq \|\phi(\mathbf{x}_i) - \mathbf{a}_*\|^2 \text{ with } 2\xi_i = t - \|\phi(\mathbf{x}_i) - \mathbf{a}_*\|^2, \quad i \in \mathcal{SV}_-. \quad (4.11)$$

Then t is optimal if and only if $t \in T_*$. Under the assumption, one has $i_- \in \mathcal{T}_- \cap \mathcal{SV}_-$. Then the optimal margin set T_* in (4.9) reduces to the single point $\{\|\phi(\mathbf{x}_{i_-}) - \mathbf{a}_*\|^2\}$. \blacksquare

4.4 ν -property

Following the terminology in Schölkopf et al. (2001) (see e.g., Scholkopf et al., 2000; Wu and Ye, 2009), A training example $\mathbf{x}_i, 1 \leq i \leq \ell$ is called a support vector (SV) if the associated multiplier $\alpha_i^* > 0$. Therefore, $\{\mathbf{x}_i : i \in \mathcal{SV}_- \cup \mathcal{SV}_+\}$ are the set of SVs. \mathbf{x}_i is a margin error (ME) if the corresponding $\xi_i^* > 0$. As for MEs, by complementary condition (4.1f),

$$\alpha_i^* = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ is a positive ME,} \\ b & \text{if } \mathbf{x}_i \text{ is a negative ME.} \end{cases} \quad (4.12)$$

Therefore, the MEs in the positive constitute a subset of \mathcal{SV}_+ and the MEs in the negative is a subset of \mathcal{SV}_- . Let m_+ and n_+ denote the number of MEs in the positive and negative classes, while s_+ and s_- denote the number of SVs in the positive and negative classes, respectively. This means

$$m_+ \leq |\mathcal{SV}_+|, \quad n_+ \leq |\mathcal{SV}_-|, \quad s_+ = |\mathcal{SV}_+|, \quad s_- = |\mathcal{SV}_-|.$$

The ν -property (Schölkopf et al., 1999, 2001; Scholkopf et al., 2000; Wu and Ye, 2009) established the connection between the model parameters and the fractions of SV and ME in the dataset. Our model (\mathcal{P}) involves a nonnegative parameter μ and two positive parameters ν and b . We can also derive a similar ν -property for these parameters in the following proposition.

Proposition 14 *Suppose (2.2) holds. If the optimal margin of (\mathcal{P}) is positive, then*

$$\frac{m_+}{\ell} \leq \nu + \mu \leq \frac{s_+}{\ell}, \quad \frac{n_+}{\ell} \leq \frac{\mu}{b} \leq \frac{s_-}{\ell}; \quad (4.13)$$

Otherwise,

$$\frac{m_+}{\ell} \leq \nu + b\frac{s_-}{\ell}, \quad \nu + b\frac{n_+}{\ell} \leq \frac{s_+}{\ell}, \quad \frac{\mu}{b} \leq \frac{s_-}{\ell}. \quad (4.14)$$

Proof From the optimality condition (4.1d), there exists an optimal solution α^* of (\mathcal{D}) satisfies

$$\sum_{i=1}^m \alpha_i^* - \sum_{i=m+1}^{\ell} \alpha_i^* = \ell\nu \quad (4.15)$$

and

$$\sum_{i=m+1}^{\ell} \alpha_i^* \geq \ell\mu. \quad (4.16)$$

If $t_* > 0$, by (4.1i), it holds that

$$\sum_{i=m+1}^{\ell} \alpha_i^* = \ell\mu \quad (4.17)$$

and

$$\sum_{i=1}^m \alpha_i^* = \ell(\nu + \mu). \quad (4.18)$$

Summing up α_i^* over the positive MEs leads to

$$m_+ = \sum_{\mathbf{x}_i \text{ is positive ME}} \alpha_i^* \leq \sum_{i \in \mathcal{SV}_+} \alpha_i^* = \sum_{i=1}^m \alpha_i^*, \quad (4.19)$$

where the last equality is by the fact that $\alpha_i^* = 0$ for $i \in [1, m] \setminus \mathcal{SV}_+$. On the other hand,

$$s_+ = |\mathcal{SV}_+| \geq \sum_{i \in \mathcal{SV}_+} \alpha_i^* = \sum_{i=1}^m \alpha_i^*. \quad (4.20)$$

Combining (4.18), (4.19) and (4.20) we can derive the first group inequalities in (4.13).

Summing up α_i^* over the negative MEs leads to

$$bn_+ = \sum_{\mathbf{x}_i \text{ is negative ME}} \alpha_i^* \leq \sum_{i \in \mathcal{SV}_-} \alpha_i^* = \sum_{i=m+1}^{\ell} \alpha_i^*, \quad (4.21)$$

where the last equality is by the fact that $\alpha_i^* = 0$ for $i \in [m+1, \ell] \setminus \mathcal{SV}_-$. On the other hand,

$$bs_- \geq \sum_{i \in \mathcal{SV}_-} \alpha_i^* = \sum_{i=m+1}^{\ell} \alpha_i^*. \quad (4.22)$$

The second group inequalities in (4.13) is then derived based on (4.17), (4.21) and (4.22).

If the margin $t_* = 0$, it follows from (4.19), (4.22) and (4.15) that

$$m_+ - bs_- \leq \sum_{i=1}^m \alpha_i^* - \sum_{i=m+1}^{\ell} \alpha_i^* = \ell\nu. \quad (4.23)$$

Similarly, it follows from (4.20), (4.21) and (4.15) that

$$s_+ - bn_+ \geq \sum_{i=1}^m \alpha_i^* - \sum_{i=m+1}^{\ell} \alpha_i^* = \ell\nu. \quad (4.24)$$

Combining (4.23) and (4.24) yields the first two inequalities in (4.14). The last inequality in (4.14) is true since by (4.16) one has

$$bs_- \geq \sum_{i \in \mathcal{SV}_-} \alpha_i^* = \sum_{i=m+1}^{\ell} \alpha_i^* \geq \ell\mu,$$

completing the proof. ■

In the absence of the negative data where $\mu = 0$ and all the terms regarding the negative data disappear in (\mathcal{P}) , then assumption (2.2) reduces to $\nu < 1$ and one can easily arrive at the following result for this particular case.

Corollary 15 *Assume $\nu \in (0, 1)$ and $m = \ell$. Then the solution of (\mathcal{P}) satisfies*

$$\frac{m_+}{m} \leq \nu \leq \frac{s_+}{m}, \quad (4.25)$$

that is ν is an upper bound on the fraction of errors, and is a lower bound on the fraction of SVs.

The following result for the degenerate case can also be easily derived by a similar argument based on (4.3b).

Proposition 16 *Suppose $\mu < m/\ell$ and $\mu + \nu \geq m/\ell$ holds. One has*

$$\frac{n_+}{\ell} \leq \frac{\mu}{b} \leq \frac{s_-}{\ell}. \quad (4.26)$$

Proof Summing up α_i^* over the negative MEs leads to

$$bn_+ = \sum_{\mathbf{x}_i \text{ is negative ME}} \alpha_i^* \leq \sum_{i \in \mathcal{SV}_-} \alpha_i^* \leq \sum_{i=m+1}^{\ell} \alpha_i^* = \ell\mu \quad (4.27)$$

by (4.3b) and (4.12). On the other hand,

$$bs_- \geq \sum_{i \in \mathcal{SV}_-} \alpha_i^* = \sum_{i=m+1}^{\ell} \alpha_i^* = \ell\mu. \quad (4.28)$$

Combining (4.27) and (4.28) we can derive (4.26). ■

5. Numerical Experiments

In this section, we report experimental results to demonstrate the effectiveness of the proposed method. We investigate the influence of nonconvexity in state-of-the-art models and compare the classification boundaries learned by our proposed primal CSSLM (\mathcal{P}) . We also compare the performance of the dual CSSLM (\mathcal{D}) —against several baselines, including primal SSLM (1.8), dual SSLM (1.9), SVDD, SVM (1.10), Decision Tree (DT), KNN, and DNN models (MLP and CNN).

Most experiments are executed on a machine with an Intel Core i7-13700H CPU and 16 GB of RAM, while the DNN models are trained on a separate server with an Intel Xeon Gold 6330 CPU, 100 GB of RAM, and an NVIDIA RTX 3090 GPU.

5.1 Nonconvexity Issues in Solving the Primal and Dual SSLM

We analyze the influence of nonconvexity in primal SSLM (1.8) and dual SSLM (1.9) respectively with the Banana Dataset², which is a synthetic 2D benchmark containing 5,300 points forming two banana-shaped clusters. Approximately 10% of points lie in an overlapping noisy region. To reduce noise, we relabel points satisfying $x_2 + \frac{3}{7}(x_1 - 3) > 0$ as -1 . The dataset is generated from Gaussian data nonlinearly transformed into separable banana classes, providing an intuitive yet challenging test for classifier performance. Four scenarios are randomly chosen from the training set, and Scenarios A-C address binary classification while Scenario D focuses on anomaly detection. Note that to ensure the reproducibility of the experimental results, once the data for each scenario are randomly selected from the training set, they become fixed. Table 1 is the summary of the banana data set.

	Training set	Scenario A	Scenario B	Scenario C	Scenario D	Test
(m, n)	(1630, 2670)	(3, 3)	(100, 100)	(768, 1232)	(1000, 100)	(381, 619)

Table 1: Summary of training sets across scenarios and test set of the Banana Dataset.

To illustrate the challenges of solving the primal and dual SSLM due to its nonconvexity, we applied two advanced nonlinear solvers available in `AMPL` to the 10-dimensional feature space in Scenarios A, B, and C using the cubic polynomial feature mapping:

$$\Phi(\mathbf{x}) = (1, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1x_2, x_1^3, x_2^3, \sqrt{3}x_1x_2^2, \sqrt{3}x_1^2x_2),$$

corresponding to the kernel function $k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T \mathbf{y})^3$.

For consistent comparison, the optimal values of SSLM and dual SSLM were scaled by a factor of $\ell\nu/2$ to match the optimal values of the primal CSSLM (P) and dual CSSLM (D). The parameters were set as $b = 1, \nu = 0.001$ and $\mu = 0.05$, satisfying conditions (2.1) and (2.2). SSLM-specific parameters $(\bar{\nu}, \bar{\nu}_1, \bar{\nu}_2)$ were configured according to (3.2).

The solvers employed include the global optimization solver `Gurobi` and the local optimization solver `Snopt`. Table 2 summarizes some initial points, optimal values, runtime, and optimality residuals obtained by each solver. Figure 4 illustrates the optimal values of the primal SSLM (1.8) obtained with random initializations under the conditions $R = 0, \rho = 0$ and $R = 0, \rho \neq 0$ in Scenario C on the Banana Dataset.

Our experiments yield several key observations. Across all three scenarios, `Gurobi` successfully identified the global optimum (highlighted in bold), though it required substantially more computational time compared to local solvers such as `Snopt`. Notably, `Snopt` consistently converged to a local optimum from different initial points, and the significant gap between local and global solutions underscores the risk of relying solely on local methods. Furthermore, under certain random initializations, the primal SSLM (1.8) also exhibited convergence to local optima, highlighting the challenge of nonconvex optimization even with advanced solvers.

As demonstrated, achieving the global optimum for the nonconvex primal SSLM (1.8) is challenging. A natural alternative is to solve the convex dual SSLM (1.9) instead of the primal problem. However, our experiment shows that the dual SSLM does not always yield the global optimal solution, reflecting the lack of strong duality. We used `quadprog` in Matlab to solve the dual SSLM

2. <https://leon.bottou.org/papers/bordes-ertekin-weston-bottou-2005>

Scenario	Initial point (\mathbf{a}, R, ρ)	Solver	Optimal value	Runtime	Optimality Error
A	$(\mathbf{0}, 0, 0)$	Gurobi	-225.8368	4.79	9.4735e-9
		Snopt	1.3246	0.00	7.0000e-7
	$(\mathbf{0}, 1, 0)$	Gurobi	-225.8368	3.81	9.4735e-9
		Snopt	0.0036	0.00	1.7000e-6
	$(\mathbf{1}, 0, 1)$	Gurobi	-225.8368	3.71	9.4735e-9
		Snopt	-9.9497	0.00	1.1000e-6
B	$(\mathbf{0}, 0, 0)$	Gurobi	121.9522	0.96	1.1057e-8
		Snopt	1586.8200	0.03	1.3000e-7
	$(\mathbf{0}, 1, 0)$	Gurobi	121.9522	0.83	1.1057e-8
		Snopt	121.9522	0.01	1.8000e-7
	$(\mathbf{1}, 0, 1)$	Gurobi	121.9522	0.80	1.1057e-8
		Snopt	1546.7500	0.07	1.7000e-6
C	$(\mathbf{0}, 0, 0)$	Gurobi	1747.4100	13.13	6.0332e-9
		Snopt	12024.7000	2.78	5.8000e-7
	$(\mathbf{0}, 1, 0)$	Gurobi	1747.4100	13.32	6.0332e-9
		Snopt	1747.4100	0.57	6.5000e-11
	$(\mathbf{1}, 0, 1)$	Gurobi	1747.4100	13.66	6.0332e-9
		Snopt	11679.5000	3.68	3.3000e-7

Table 2: Initial points and results of Gurobi and Snopt for the primal SSLM (1.8) in the 10-dimensional feature space of Scenarios A, B, and C on the Banana Dataset.

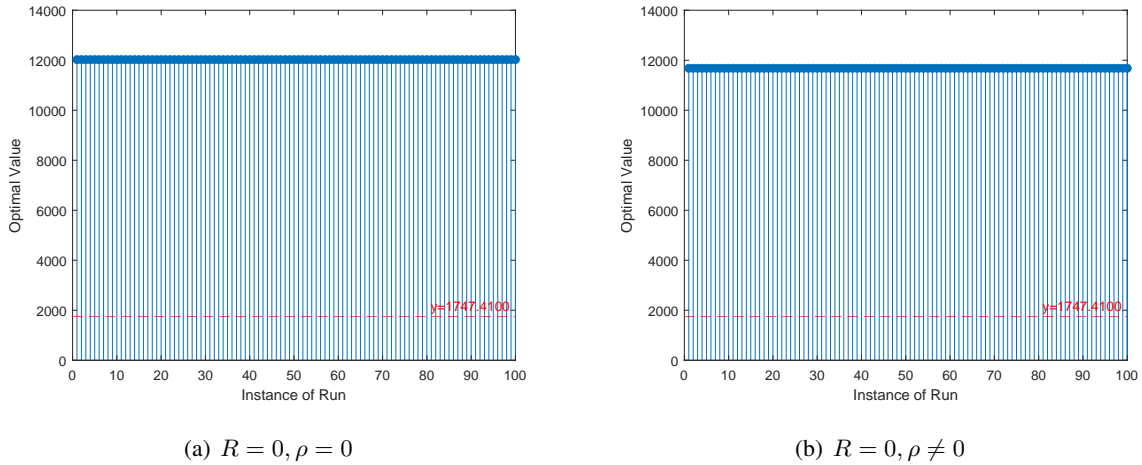


Figure 4: Optimal values of the primal SSLM (1.8) under 100 random initializations on the Banana Dataset (Scenario C) with $R = 0, \rho = 0$ and $R = 0, \rho \neq 0$. The dashed horizontal line marks the global optimal value ($y = 1747.41$), which was NOT achieved in the vast majority of runs, demonstrating the sensitivity of the algorithm to initial conditions.

(1.9), primal CSSLM (\mathcal{P}), and its dual problem (\mathcal{D}) in Scenarios A, B, and C on Banana Dataset, all of which are convex QPs. Table 3 lists the optimal objectives, exitflags³, runtime, and optimality errors.

Scenario	Model	Optimal value	Exitflag	Runtime	Optimality Error
A	Dual SSLM (1.9)	-225.8368	1	0.0161	2.7149e-11
	CSSLM (\mathcal{P})	-225.8368	1	0.0072	4.1457e-9
	CSSLM (\mathcal{D})	-225.8368	1	0.0048	1.3184e-10
B	Dual SSLM (1.9)	121.7294	1	0.0067	1.7144e-11
	CSSLM (\mathcal{P})	121.9522	1	0.0425	6.5391e-11
	CSSLM (\mathcal{D})	121.9522	1	0.0066	1.9341e-9
C	Dual SSLM (1.9)	1733.1000	1	2.0308	4.1751e-9
	CSSLM (\mathcal{P})	1747.4000	1	16.0189	1.6268e-10
	CSSLM (\mathcal{D})	1747.4000	1	1.4373	4.0643e-12

Table 3: The results of solvers for the dual SSLM (1.9), proposed problem (\mathcal{P}) and its dual problem (\mathcal{D}) in 10-dimensional feature space of Scenarios A, B, C on the Banana Dataset.

We observe that the dual optimal solution for SSLM does not always correspond to the global optimal solution of the primal SSLM (in bold) in Scenarios B and C (in bold). This highlights the absence of strong duality in nonconvex SSLM, indicating that solving the dual SSLM does not resolve nonconvexity issues present in the primal problem. Our proposed primal and dual CSSLM consistently achieved the global optimal solution (as verified by the global value determined by Gurobi in Table 2), which confirms strong duality in our model, a clear advantage over nonconvex alternatives.

5.2 Comprehensive Performance Evaluation Across Datasets

We then compare the CSSLM with other existing methods, which include DT, KNN, SVM, SVDD, on the Multivariate Cauchy (MC) Dataset, Blood Transfusion Service Center (BTSC), Breast Cancer Wisconsin Dataset (BCW) and KDD-Cup’99 Dataset. For each dataset, we consider two scenarios, i.e. the binary classification (BC) and anomaly detection (AD). Table 4 gives the summary of each data set and each scenario.

(m, n)	MC	BTSC	BCW	KDD-Cup’99
Training set	(250, 750)	(125, 399)	(149, 250)	(97278, 396743)
Test set	(250, 750)	(53, 171)	(63, 107)	(60593, 250436)
BC	(200, 200)	(100, 100)	(100, 100)	(1000, 1000)
AD	(200, 20)	(100, 10)	(100, 10)	(1000, 100)

Table 4: Summary of each data set and each scenario in performance evaluation.

As for the evaluation metric, we adopt the Area Under the Precision–Recall Curve (AUPRC) (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015), which is particularly suitable for im-

3. `exitflag = 1` in `quadprog` means the optimal solution is found

balanced datasets and emphasizes the performance of the positive class. Following this setting, we compute AUPRC by treating negative samples as the positive class.

For parameter selection, we employ five-fold stratified cross-validation following Wu and Ye (2009), ensuring that each fold preserves the positive-to-negative ratio of the original training set. Specifically we use grid search to select the hyperparameters that yield the highest average AUPRC across the folds. Table 5 gives the grids of parameters in each model, where the Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2}$ is used in SVM, SVDD and CSSLM, MinSamplesLeaf and MaxNumSplits are the minimum number of observations per leaf and the maximum number of splits in DT, respectively. For CSSLM, the parameter μ is selected from $\{0, 0.01m/\ell, 0.05m/\ell, 0.1m/\ell, m/\ell\}$ and delete the value according to (2.1). Then the parameter ν is selected from $\{0.01m/\ell, 0.1m/\ell\}$ according to (2.2). These constraints significantly reduce the computational overhead typically associated with cross-validation.

Parameter	Grid	Model
γ	$\{\frac{2^k}{\sigma^2} : k \in [-4, 4]\}^\dagger$	SVM, SVDD, CSSLM
b	$\{\frac{m}{4n}, \frac{m}{2n}, \frac{m}{n}, \frac{2m}{n}, \frac{4m}{n}\}$	SVM, SVDD, CSSLM
$\frac{1}{\nu\ell}$	$\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500\}$	SVM(1.10)
μ	$\{0, \frac{0.01m}{\ell}, \frac{0.05m}{\ell}, \frac{0.1m}{\ell}, \frac{m}{\ell}\}$	CSSLM
ν	$\{\frac{0.01km}{\ell} : k \in [1, 9]\} \cup \{\frac{0.1km}{\ell} : k \in [1, 9]\}$	SVDD (3.4)
ν	$\{\frac{0.01m}{\ell}, \frac{0.1m}{\ell}\}$	CSSLM
MinSamplesLeaf	$\{1, 2, 5, 10, 20, 50\}$	DT
MaxNumSplits	$\{5, 10, 20, 50, 100, 200, 500\}$	DT
k	$\{1, 2, 5, 10, 20, 25, 50, 100, 500\}$	KNN

Table 5: The grid for each parameters.

\dagger Here σ^2 is the mean squared norm of the training data.

Test errors on the positive and negative classes are denoted by e_+ and e_- , respectively. We also employ the geometric mean of sensitivity and specificity (G-means) metric, errors on the positive and negative classes as performance measures to show the overall accuracy and the trade-off for false-positive and false-negative rates. G-means is defined as $g = \sqrt{(1 - e_+)(1 - e_-)}$, where $1 - e_+$ and $1 - e_-$ are the so-called accuracy on positive and negative class.

Note that to quantify the central tendency and variability of algorithm performance, the data for each scenario are randomly selected from the training set, and calculate the AUPRC, G-means, and classification errors (e_+ , e_-) on the test set. Tables 6 report the average AUPRC, G-means, and classification errors (e_+ , e_-) with standard deviations across 10 runs. The best results in each scenario are highlighted in bold. A method is considered to have broken down if its average G-means falls below 50%. Below we describe each dataset and summarize dataset-specific findings.

The MC dataset is a 2D synthetic dataset where positive points are sampled from a multivariate Gaussian distribution and negative points from a multivariate Cauchy distribution (see Figure 5). The dataset’s heavy-tailed, non-Gaussian characteristics provide a challenging test case for evaluating robustness against outliers and heavy-tailed noise in statistical learning. For MC dataset the results in Table 6 show that in BC, CSSLM achieves the highest AUPRC (0.9944 ± 0.0002), slightly higher than SVM (0.9942 ± 0.0006), while attaining the lowest negative-class error ($e_- = 3.60\% \pm 0.49$). In AD, CSSLM secures the best G-means ($87.61\% \pm 8.60$) together with the lowest

Dataset	Scenario	Model	AUPRC	G-means (%)	e_+ (%)	e_- (%)
MC	BC	DT	0.9821 ± 0.0106	92.60 ± 0.92	4.88 ± 2.77	9.79 ± 2.56
		KNN	0.9918 ± 0.0016	91.15 ± 0.92	1.12 ± 0.88	15.95 ± 2.35
		SVM	0.9942 ± 0.0006	93.39 ± 0.81	2.92 ± 2.75	10.07 ± 3.85
		SVDD	0.9930 ± 0.0010	92.92 ± 1.30	8.08 ± 4.05	5.97 ± 2.20
		CSSLM	0.9944 ± 0.0002	92.09 ± 0.78	11.92 ± 3.00	3.60 ± 0.49
	AD	DT	0.9697 ± 0.0058	87.14 ± 2.75	0.24 ± 0.34	23.81 ± 4.96
		KNN	0.9807 ± 0.0063	65.63 ± 34.66	0.00 ± 0.00	46.12 ± 28.65
		SVM	0.9910 ± 0.0083	71.84 ± 38.19	0.48 ± 0.88	34.85 ± 35.43
		SVDD	0.9892 ± 0.0036	85.93 ± 12.83	16.20 ± 23.62	9.16 ± 4.89
		CSSLM	0.9881 ± 0.0082	87.61 ± 8.60	17.96 ± 14.95	5.59 ± 2.97
BTSC	BC	DT	0.8943 ± 0.0207	64.98 ± 2.89	41.87 ± 10.33	26.02 ± 8.11
		KNN	0.8816 ± 0.0002	6.61 ± 20.90	3.02 ± 9.55	93.74 ± 19.79
		SVM	0.8954 ± 0.0096	57.63 ± 20.56	33.77 ± 21.06	37.13 ± 26.62
		SVDD	0.8698 ± 0.0213	58.14 ± 10.62	39.62 ± 25.45	34.50 ± 21.65
		CSSLM	0.8681 ± 0.0273	60.02 ± 7.95	36.23 ± 24.27	36.20 ± 19.35
	AD	DT	0.8822 ± 0.0016	2.73 ± 8.64	0.19 ± 0.60	99.24 ± 2.40
		KNN	0.8419 ± 0.0146	0.00 ± 0.00	0.00 ± 0.00	100.00 ± 0.00
		SVM	0.8802 ± 0.0199	48.78 ± 26.46	47.55 ± 32.97	30.41 ± 22.60
		SVDD	0.8646 ± 0.0251	49.18 ± 21.02	46.04 ± 35.42	32.98 ± 26.27
		CSSLM	0.8665 ± 0.0267	62.04 ± 5.72	28.87 ± 16.50	42.87 ± 15.30
BCW	BC	DT	0.9215 ± 0.0452	87.27 ± 1.58	11.43 ± 4.22	13.83 ± 4.64
		KNN	0.9755 ± 0.0043	88.75 ± 1.52	13.33 ± 3.53	9.07 ± 1.40
		SVM	0.9882 ± 0.0025	91.62 ± 1.59	10.64 ± 5.19	5.89 ± 3.18
		SVDD	0.9607 ± 0.0091	85.69 ± 1.45	18.73 ± 2.22	9.63 ± 1.40
		CSSLM	0.9850 ± 0.0022	83.90 ± 3.14	29.05 ± 5.35	0.65 ± 1.08
	AD	DT	0.8819 ± 0.0749	78.81 ± 7.01	8.10 ± 2.64	31.68 ± 12.86
		KNN	0.9544 ± 0.0196	44.89 ± 38.89	1.91 ± 2.57	65.05 ± 31.04
		SVM	0.9658 ± 0.0102	41.90 ± 44.24	2.54 ± 3.01	62.90 ± 39.45
		SVDD	0.9466 ± 0.0222	76.07 ± 26.86	13.49 ± 5.95	24.21 ± 27.18
		CSSLM	0.9678 ± 0.0186	81.91 ± 3.25	30.79 ± 4.62	2.99 ± 1.96
KDD-Cup'99	BC	DT	0.9893 ± 0.0013	94.50 ± 0.45	0.92 ± 0.98	9.86 ± 0.25
		KNN	0.9909 ± 0.0002	94.81 ± 0.11	0.33 ± 0.13	9.81 ± 0.14
		SVM	0.9911 ± 0.0060	90.64 ± 3.63	0.80 ± 0.65	17.06 ± 6.67
		SVDD	0.9952 ± 0.0013	86.22 ± 6.20	21.98 ± 14.87	3.57 ± 4.14
		CSSLM	0.9954 ± 0.0014	87.40 ± 3.10	17.22 ± 8.70	3.78 ± 1.78
	AD	DT	0.9883 ± 0.0025	93.80 ± 1.43	0.18 ± 0.43	11.84 ± 2.64
		KNN	0.9767 ± 0.0008	87.07 ± 0.32	0.83 ± 0.48	23.55 ± 0.47
		SVM	0.9971 ± 0.0006	81.19 ± 8.98	0.01 ± 0.03	33.34 ± 14.80
		SVDD	0.9948 ± 0.0015	85.67 ± 6.40	23.04 ± 15.08	3.45 ± 4.08
		CSSLM	0.9967 ± 0.0019	87.81 ± 5.48	20.32 ± 11.75	2.56 ± 2.99

Table 6: Average AUPRC, G-means, test errors (e_+ , e_-) and standard deviations for models on each scenario on each Dataset.

negative-class error ($e_- = 5.59\% \pm 2.97$). These results highlight the advantage of CSSLM in controlling negative-class errors while remaining competitive in AUPRC and G-means. Key observations: (i) CSSLM dominates in e_- across both scenarios; (ii) improvements are consistent across runs with relatively low variance.

The BTSC dataset⁴ contains 4-dimensional donor data from Hsin-Chu City, Taiwan. This real-world data set captures the demographic and behavioral patterns of blood donors, making it particularly valuable for predicting donor return behavior and evaluating classification methods. Before modeling, we apply logarithmic transformation to the features. On this more challenging dataset, many methods degrade significantly. In BC, CSSLM maintains stable performance (AUPRC = 0.8681 ± 0.0273 , G-means = $60.02\% \pm 7.95$), whereas KNN collapses with G-means

4. <https://archive.ics.uci.edu/dataset/176/blood+transfusion+service+center>

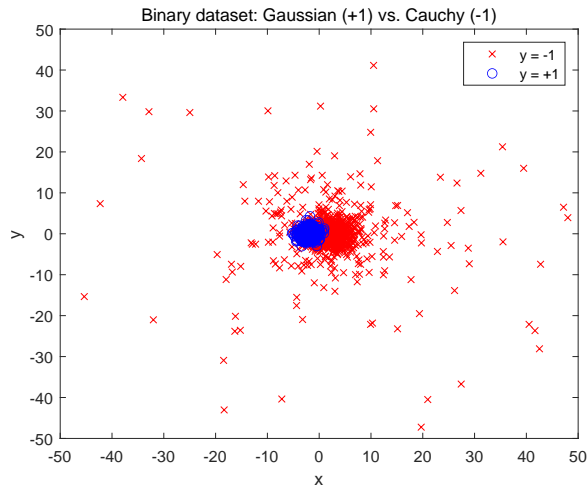


Figure 5: The data distribution of the Multivariate Cauchy Dataset.

$= 6.61\% \pm 20.90$. In AD, CSSLM again sustains performance (AUPRC = 0.8665 ± 0.0267 , G-means = $62.04\% \pm 5.72$), while other methods all break down. Key observations: (i) CSSLM is the only method that avoids collapse in both scenarios; (ii) it consistently yields relatively lower e_- .

The BCW dataset⁵ contains 30 features derived from digitized images of fine needle aspirates of breast masses, describing characteristics of cell nuclei. This medical dataset is widely used to benchmark classification methods in cancer diagnosis. Prior to modeling, we apply log transformation to the data. In BC, CSSLM performs competitively (AUPRC = 0.9850 ± 0.0022), close to SVM (0.9882 ± 0.0025), while achieving the lowest negative-class error ($0.65\% \pm 1.08$). In AD, CSSLM achieves the best AUPRC (0.9678 ± 0.0186), best G-means (81.91 ± 3.25) and again the lowest negative-class error ($2.99\% \pm 1.96$). Meanwhile, both KNN and SVM suffer breakdown under AD. Key observations: (i) CSSLM balances e_+ and e_- more reliably than baselines; (ii) KNN minimizes e_+ but incurs substantially higher e_- .

The KDD-Cup’99 dataset⁶ is a 41-dimensional benchmark for network intrusion detection, containing simulated traffic data with normal connections and various attacks. We use the 10% training subset “kddcup.data_10_percent” and the full test set “kddcup.data.corrected”. Categorical attributes are converted via one-hot encoding, followed by min-max normalization. In BC, CSSLM achieves the best AUPRC (0.9954 ± 0.0014) and a slightly higher negative-class error ($3.78\% \pm 1.78$) compared to SVDD ($3.57\% \pm 4.14$). In AD, CSSLM again dominates with the highest AUPRC (0.9967 ± 0.0019) and lowest e_- ($2.56\% \pm 2.99$), whereas SVM attains a low e_+ but at the cost of very high e_- . The key observation is CSSLM consistently secures both the best AUPRC and relatively low e_- .

For comparison with deep models we additionally train MLP and CNN on KDD-Cup’99 with *all* the training set ($(m, n) = (97278, 396743)$) of the KDD-Cup’99 Dataset. Here we adopt a 6-layer MLP and a 1-layer CNN following Vinayakumar et al. (2017), trained with learning rate 0.1, batch size 128, and 50 epochs. Table 7 shows the test results. Even when MLP (AUPRC = 0.9973) and CNN (0.9940) are trained on the *full* training set, their performance remains comparable to CSSLM

5. <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

6. <https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

trained on much smaller samples. This underscores the efficiency of support vector-based methods, whose performance depends primarily on the number of support vectors rather than the training set size. Hence the support vector-based approaches remain highly sample-efficient compared to deep neural models.

Model	AUPRC	G-means	e_+ (%)	e_- (%)
MLP	0.9973	96.58	1.33	5.47
CNN	0.9940	96.48	1.01	5.96

Table 7: The AUPRC, G-means, test errors (e_+ , e_-) for MLP and CNN on the KDD-Cup’99 Dataset.

5.3 Decision Boundary Visualization on the Banana Dataset

We examine the differences in the decision boundaries learned by the SVM-based models. To this end, we apply the Gaussian-kernel models (SVM, SVDD, and CSSLM) to Scenarios C and D of the banana dataset in Table 1. The SVDD and CSSLM are solved via their dual formulations using `quadprog` in Matlab, while SVM is trained with `libsvm`. For comparison, the performance of DT and KNN are included and we adopted the built-in `fitctree` and `fitcknn` functions in Matlab solving DT and KNN, respectively. Table 8 reports the parameters chosen together with the test performance. Figure 6 presents the AUPRC curves of all the methods in the test set. Figure 7 illustrates the decision boundaries learned by SVM, SVDD, and CSSLM on the training data. Figure 8 further overlays these boundaries on the test data for direct comparison.

Scenario	Model	Hyperparameter				Evaluation metric			
		μ	ν	b	γ	AUPRC	G-means (%)	e_+ (%)	e_- (%)
C	DT	-	-	-	-	0.9642	90.94	12.34	5.65
	KNN	-	-	-	-	0.9908	92.52	10.50	4.36
	SVM	-	5.0000	0.6234	0.3762	0.9920	93.08	6.56	7.27
	SVDD	0	0.0038	0.3117	0.1881	0.9871	92.63	4.97	9.69
	CSSLM	0.0384	0.0038	0.1558	0.3762	0.9925	92.91	8.66	5.49
D	DT	-	-	-	-	0.9535	83.62	0.26	29.87
	KNN	-	-	-	-	0.9889	78.25	0.00	38.77
	SVM	-	0.5000	5.0000	0.4385	0.9920	91.28	1.58	15.35
	SVDD	0	0.0091	2.5000	0.2193	0.9887	92.30	2.89	12.28
	CSSLM	0.0909	0.0091	2.5000	0.4385	0.9921	92.42	10.24	4.85

Table 8: Hyperparameters and evaluation metrics of each models on Scenarios C and D of the Banana Dataset.

In Scenario C, CSSLM achieves a false-positive rate 5.49%, which is significantly lower than SVM (7.27%) and SVDD (9.69%), at the cost of a slightly higher false-negative rate 8.66% compared to SVM (6.56%) and SVDD (4.97%). This indicates that CSSLM adopts a conservative strategy to reduce false positives, which explains its slightly higher overall test error. In Scenario D, where negative samples are scarce, CSSLM attains a low false-negative rate 10.24% along with a very competitive false-positive rate 4.85%, demonstrating a more balanced performance compared to SVM ($e_+ = 1.58\%$, $e_- = 15.35\%$).

Under these metrics, CSSLM consistently achieves the best or near-best performance, with the highest AUPRC in Scenario C (0.9925) and Scenario D (0.9921). Across both scenarios, CSSLM consistently achieves lower negative-class errors. Figure 8 illustrates SVDD and CSSLM consis-

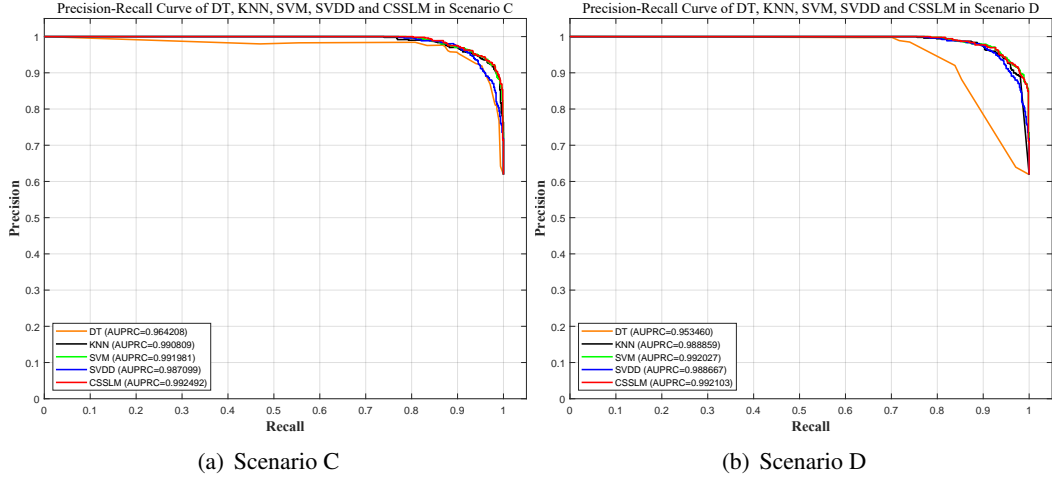


Figure 6: Comparison of precision–recall curves for DT, KNN, SVM, SVDD, and CSSLM in Scenarios C and D on the Banana Dataset.

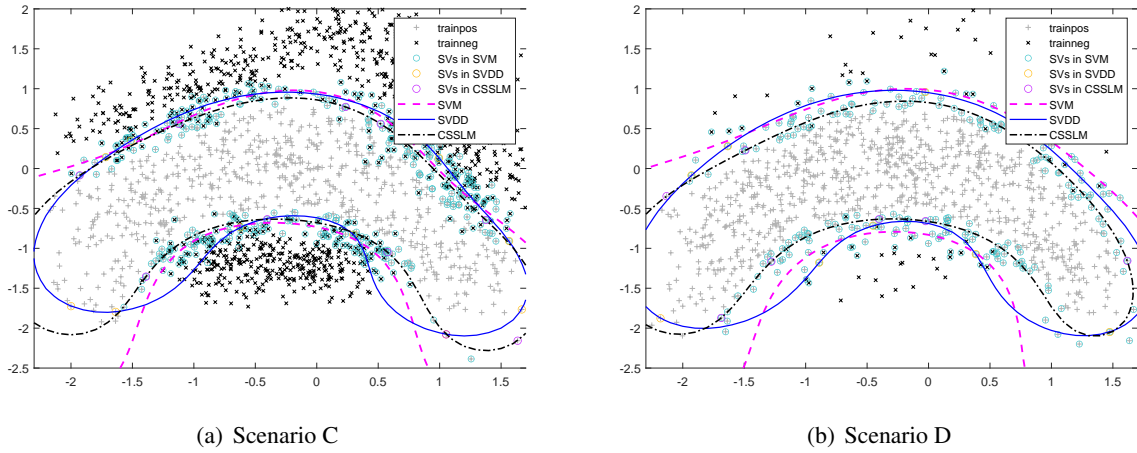


Figure 7: Decision boundaries learned by SVM, SVDD, and CSSLM on the training data in Scenarios C and D on the Banana Dataset. Data points inside the circles indicate the supporting vectors.

tently identify similar banana-shaped boundaries across datasets, whereas SVM with a Gaussian kernel tends to overfit positive samples, explaining its lower positive-class error but higher negative-class error. Overall, these results suggest that although CSSLM may not always minimize raw test error in heavily imbalanced settings, it provides more balanced error trade-offs and superior robustness across scenarios.

5.4 Solving Primal CSSLM via Stochastic Subgradient Method

We address the computational challenge of solving the proposed CSSLM in this subsection. The dual CSSLM (\mathcal{D}) involves one linear equality and inequality constraints with box constraints. It is

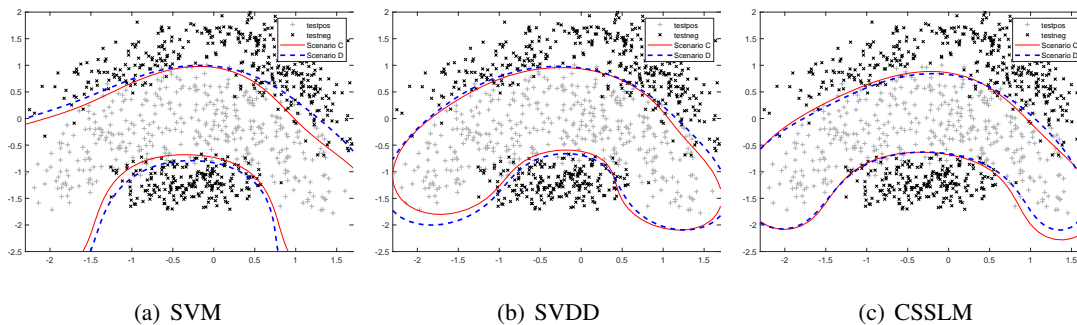


Figure 8: Decision boundaries learned by SVM, SVDD, and CSSLM on the test data in Scenarios C and D on the Banana Dataset.

more simple than the primal CSSLM (\mathcal{K}). This is a common phenomenon in SVM-like models. Similar to SVM (Chang and Lin, 2011), various coordinate descent methods and stochastic coordinate descent methods can be designed to solve the dual CSSLM (\mathcal{D}). However it's not easy to describe them clearly in a limited space.

Now we turn to solve the primal CSSLM (\mathcal{K}) which is a composition optimization, the sum of the smooth regularization term and the nonsmooth empirical loss function. Motivated by the well-known Pegasos (Shalev-Shwartz et al., 2011), a stochastic sub-gradient method for SVM, we propose a stochastic subgradient method (SSGM) to solve (\mathcal{K}). This approach is especially advantageous in large-scale settings. The SSGM operates with a batch size of 1 and adopts a decreasing learning rate $\eta_k = 1/(\nu \cdot k)$ at iteration k . A data point with index $i_k \in [1, \ell]$ is selected. The subgradient of the algorithm at $\mathbf{w}_k = (\mathbf{a}_k, s_k, t_k)$ is expressed as

$$\nabla_{\mathbf{w}_k} := \left(\nu \mathbf{a}_k, -\frac{\nu}{2}, -\frac{\mu}{2} \right) + \nabla \phi_{i_k}(\mathbf{w}_k),$$

where the example-dependent component $\nabla \phi_{i_k}(\mathbf{w}_k)$ is defined as

$$\nabla \phi_{i_k}(\mathbf{w}_k) = \begin{cases} (-\Phi(\mathbf{x}_{i_k}), \frac{1}{2}, 0) \cdot \mathbb{1} \left[\langle \mathbf{a}_k, \Phi(\mathbf{x}_{i_k}) \rangle < \frac{\|\Phi(\mathbf{x}_{i_k})\|^2 + s_k}{2} \right], & y_{i_k} = 1, \\ b (\Phi(\mathbf{x}_{i_k}), -\frac{1}{2}, \frac{1}{2}) \cdot \mathbb{1} \left[\langle \mathbf{a}_k, \Phi(\mathbf{x}_{i_k}) \rangle > \frac{\|\Phi(\mathbf{x}_{i_k})\|^2 + s_k - t_k}{2} \right], & y_{i_k} = -1, \end{cases}$$

with $\mathbf{a}_k = \sum_{j=1}^{\ell} \beta_j^k \Phi(\mathbf{x}_j)$ and $\mathbb{1}$ representing the indicator function.

We compare SSGM against the `quadprog` solver in MATLAB on the KDD-Cup'99 Dataset with 20k samples ($m = 10000, n = 10000$). The parameters in (\mathcal{K}) were set as $b = 1, \nu = 0.001$ and $\mu = 0.05$. Again the gaussian kernel is used with $\gamma = 1$. Figure 9 illustrates how the relative error changes with the number of data points accessed where the optimal value means the result obtained by solving (\mathcal{D}) using `quadprog`. As shown in Figure 9, SSGM significantly reduces computational cost compared to `quadprog`, especially as the dataset size increases. While `quadprog` computes an exact solution at high computational expense, SSGM achieves a rapidly decaying optimization error with far fewer data accesses. This demonstrates that SSGM is a scalable and practical alternative for large-scale applications of CSSLM.

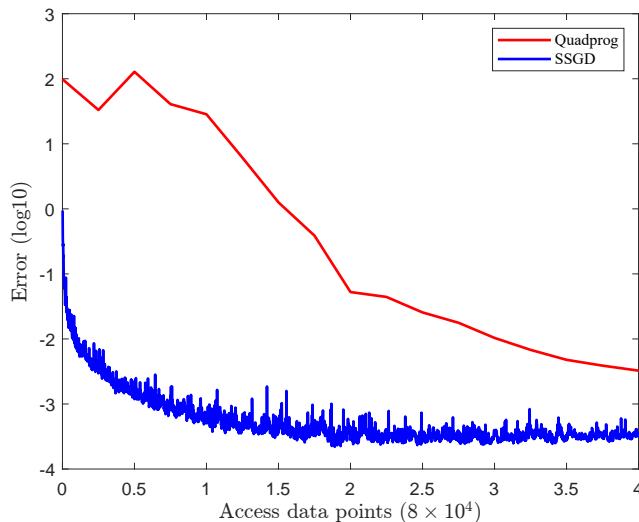


Figure 9: Comparison of optimization error between SSGM for solving CSSLM (\mathcal{K}) and quadprog for solving the dual CSSLM (\mathcal{D}).

6. Conclusion

We proposed a novel convex SSLM formulation that, for specific hyperparameter choices, reduces to a convex quadratic programming problem. The convexity of our model enables rigorous analysis and derivation of results that are challenging or impossible for traditional nonconvex SSLM models, notably simplifying the study of hypersphere-based SVM solutions. We systematically investigated the influence of hyperparameters on the optimal solution and established explicit connections with traditional SSLM, identifying conditions under which the solution is unique. Moreover, we derived the ν -property, providing insight into how hyperparameters govern the distribution of support vectors and margin errors across positive and negative classes.

Extensive numerical experiments demonstrate that our method consistently outperforms both the primal SSLM and its dual in terms of efficiency and accuracy. By addressing the challenges of nonconvexity, our work paves the way for developing more robust algorithms for hypersphere-based SVMs. Future work will explore additional theoretical properties and algorithmic strategies that leverage the convexity of our formulation.

Acknowledgments

This work was initiated through discussions between the first author and Prof. Shuisheng Zhou at Xidian University on the role of nonconvexity in anomaly detection. The authors gratefully acknowledge Prof. Zhou for his many insightful suggestions and constructive comments during the preparation of this paper. The importance of convex modeling in the context of bilevel optimization was inspired by discussions with Dr. Samuel Ward from the University of Southampton. The authors sincerely thank Dr. Ward for highlighting this point and for his valuable feedback.

Hongying Liu is supported by the National Natural Science Foundation of China under Grant 12171021 and 12131004.

Hao Wang is supported by the National Natural Science Foundation of China under Grant 12571326.

A. Proof of Lemma 1 in the Ill-Posed Cases

Proof If $\mu > m/\ell$, setting $(\mathbf{a}, r, t) = (\alpha\tilde{\mathbf{a}}, 0, \alpha^2\|\tilde{\mathbf{a}}\|^2)$ with any $\alpha > 0$ and $\tilde{\mathbf{a}} \neq \mathbf{0}$ yields

$$\begin{aligned}
 & g(\alpha\tilde{\mathbf{a}}, 0, \alpha^2\|\tilde{\mathbf{a}}\|^2) \\
 = & -\frac{1}{2}\mu\alpha^2\|\tilde{\mathbf{a}}\|^2 + \frac{1}{2\ell}\sum_{i=1}^m\|\Phi(\mathbf{x}_i) - \alpha\tilde{\mathbf{a}}\|^2 + \frac{b}{2\ell}\sum_{i=m+1}^{\ell}(\alpha^2\|\tilde{\mathbf{a}}\|^2 - \|\Phi(\mathbf{x}_i) - \alpha\tilde{\mathbf{a}}\|^2)_+ \\
 = & \frac{1}{2}\left(\frac{m}{\ell} - \mu\right)\alpha^2\|\tilde{\mathbf{a}}\|^2 - \frac{1}{\ell}\sum_{i=1}^m\langle\Phi(\mathbf{x}_i), \tilde{\mathbf{a}}\rangle\alpha + \frac{b}{\ell}\sum_{i=m+1}^{\ell}\left(\langle\Phi(\mathbf{x}_i), \tilde{\mathbf{a}}\rangle\alpha - \frac{k_{ii}}{2}\right)_+ + \frac{1}{2\ell}\sum_{i=1}^m k_{ii} \\
 \leq & \frac{1}{2}\left(\frac{m}{\ell} - \mu\right)\alpha^2\|\tilde{\mathbf{a}}\|^2 - \frac{1}{\ell}\left[\sum_{i=1}^m\langle\Phi(\mathbf{x}_i), \tilde{\mathbf{a}}\rangle - b\sum_{i=m+1}^{\ell}(\langle\Phi(\mathbf{x}_i), \tilde{\mathbf{a}}\rangle)_+\right]\alpha + \frac{1}{2\ell}\sum_{i=1}^m k_{ii},
 \end{aligned}$$

where the inequality is by $k_{ii} \geq 0$. It follows that $g(\alpha\tilde{\mathbf{a}}, 0, \alpha^2\|\tilde{\mathbf{a}}\|^2) \rightarrow -\infty$ as $\alpha \rightarrow \infty$, since $\mu > m/\ell$.

If $\mu > bn/\ell$, setting $(\mathbf{a}, r, t) = (0, 0, t)$ with sufficiently large $t > 0$ yields

$$g(0, 0, t) = \frac{1}{2}\left(\frac{bn}{\ell} - \mu\right)t + \frac{1}{2\ell}\sum_{i=1}^m\|\Phi(\mathbf{x}_i)\|^2 - \frac{b}{2\ell}\sum_{i=m+1}^{\ell}\|\Phi(\mathbf{x}_i)\|^2,$$

which implies $g(0, 0, t) \rightarrow -\infty$ as $t \rightarrow \infty$. ■

B. Proof of Theorem 4

Proof (i) Substituting $\mu = 0$ into $g(\mathbf{a}, r, t)$, one has

$$\begin{aligned}
 g(\mathbf{a}, r, t) &= \frac{\nu}{2}r + \frac{1}{2\ell}\sum_{i=1}^m(\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - r)_+ + \frac{b}{2\ell}\sum_{i=m+1}^{\ell}(r + t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2)_+ \\
 &\geq \frac{1}{2}\left(\nu - \frac{m}{\ell}\right)r + \frac{1}{2\ell}\sum_{i=1}^m\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \\
 &\geq \frac{1}{2\ell}\sum_{i=1}^m\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \\
 &\geq \frac{1}{2\ell}\sum_{i=1}^m\|\Phi(\mathbf{x}_i) - \hat{\mathbf{a}}\|^2.
 \end{aligned}$$

Here the first inequality holds as equality if and only if $r \leq \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2, i \in [1, m]$ and $r + t \leq \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2, i \in [m + 1, \ell]$; the second inequality holds because of $\mu = 0, \nu \geq m/\ell$ and the inequality holds as equality if and only if $\mathbf{a} = \hat{\mathbf{a}}$ and $(\nu - m/\ell)r/2 = 0$. The last inequality is by the definition (2.4) of $\hat{\mathbf{a}}$. Therefore, $g(\mathbf{a}, r, t)$ reaches its lower bound if and only if $(\mathbf{a}, r, t) \in \Omega$.

(ii) Suppose $\nu + \mu \geq m/\ell$ is satisfied. We first prove (\mathcal{F}) is equivalent to

$$\inf_{\mathbf{a} \in H, t \in \mathbb{R}} g(\mathbf{a}, 0, t). \tag{B.1}$$

Moreover, if $\inf_{\mathbf{a} \in H, t \in \mathbb{R}} g(\mathbf{a}, 0, t) > -\infty$, then the global minimizer satisfies $r = 0$ and $t \geq 0$. By $\nu + \mu \geq m/\ell$, one has $0 < \mu \leq m/\ell$. Hence for any $\mathbf{a} \in F, r \geq 0$,

$$\begin{aligned} g(\mathbf{a}, r, t) &\geq \frac{1}{2}(\nu r - \mu t) + \frac{1}{2\ell} \sum_{i=1}^m (\|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 - r) + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (r + t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2)_+ \\ &= \frac{1}{2}(\nu - \frac{m}{\ell})r - \frac{\mu}{2}t + \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (r + t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2)_+ \\ &\geq -\frac{\mu}{2}(r + t) + \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (r + t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2)_+ \\ &= g(\mathbf{a}, 0, r + t) \\ &\geq \inf_{\mathbf{a} \in H, t} g(\mathbf{a}, 0, t) \end{aligned}$$

where the first inequality comes from $(z)_+ \geq z$ and the second inequality is due to condition $\nu + \mu \geq m/\ell$. On the other hand, for any $t < 0$,

$$g(\mathbf{a}, 0, t) = -\frac{\mu}{2}t + \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \geq \frac{1}{2\ell} \sum_{i=1}^m \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 = g(\mathbf{a}, 0, 0) \quad (\text{B.2})$$

by $\mu \geq 0$. It follows that $g(\mathbf{a}, r, t) \geq \inf_{\mathbf{a} \in H, t \in \mathbb{R}} g(\mathbf{a}, 0, t) \geq \inf_{\mathbf{a} \in H, t \geq 0} g(\mathbf{a}, 0, t)$. It implies

$$\inf_{\mathbf{a} \in H, t \in \mathbb{R}} g(\mathbf{a}, 0, t) = \inf_{\mathbf{a} \in H, t \geq 0} g(\mathbf{a}, 0, t).$$

Moreover, (B.2) indicates that if $\inf_{\mathbf{a} \in H, t} g(\mathbf{a}, 0, t) > -\infty$, $(\mathbf{a}_*, t_*) \in \arg \min_{\mathbf{a} \in H, t \in \mathbb{R}} g(\mathbf{a}, 0, t)$ also satisfies $t^* \geq 0$. The former result implies that problem (\mathcal{F}) in this case is equivalent to the unconstrained problem (B.1), which is equivalent to a *convex* problem by setting $z = \|\mathbf{a}\|^2 - t$ in g :

$$\min_{\mathbf{a} \in H, z \in \mathbb{R}} \tilde{f}(\mathbf{a}, z), \quad (\text{B.3})$$

with (see Appendix C for deriving $\tilde{f}(\mathbf{a}, z)$)

$$\tilde{f}(\mathbf{a}, z) = \frac{\ell\lambda}{2}\|\mathbf{a}\|^2 + \frac{\ell\mu}{2}z - m\langle \hat{\mathbf{a}}, \mathbf{a} \rangle + b \sum_{i=m+1}^{\ell} \left(-\frac{z+k_{ii}}{2} + \langle \mathbf{a}, \Phi(\mathbf{x}_i) \rangle \right)_+.$$

If $\mu < m/\ell \leq \nu + \mu$, then (B.3) reverts to the convex quadratic program problem (2.7).

(iii) If $\mu = m/\ell$, then (B.3) reverts to the linear programming problem (2.8). \blacksquare

C. Derivation of the Objective Function in the Degenerate Case

It holds that

$$\begin{aligned} g(\mathbf{a}, 0, t) &= -\frac{\mu}{2}t + \frac{1}{2\ell} \sum_{i=1}^m \|\mathbf{a} - \Phi(\mathbf{x}_i)\|^2 + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (t - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|_2^2)_+ \\ &= -\frac{\mu}{2}t + \frac{1}{2\ell} \sum_{i=1}^m (\|\mathbf{a}\|^2 - 2\langle \mathbf{a}, \Phi(\mathbf{x}_i) \rangle + k_{ii}) + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (t - \|\mathbf{a}\|^2 + 2\langle \mathbf{a}, \Phi(\mathbf{x}_i) \rangle - k_{ii})_+ \\ &= -\frac{\mu}{2}t + \frac{m}{2\ell} \|\mathbf{a}\|^2 - \frac{m}{\ell} \langle \hat{\mathbf{a}}, \mathbf{a} \rangle + \frac{b}{2\ell} \sum_{i=m+1}^{\ell} (t - \|\mathbf{a}\|^2 + 2\langle \mathbf{a}, \Phi(\mathbf{x}_i) \rangle - k_{ii})_+ + \frac{1}{2\ell} \sum_{i=1}^m k_{ii} \end{aligned}$$

where $\hat{\mathbf{a}} = \frac{1}{m} \sum_{i=1}^m \Phi(\mathbf{x}_i)$. Letting $z = \|\mathbf{a}\|^2 - t$ in f , it follows that

$$g(\mathbf{a}, 0, t) = \frac{1}{\ell} \tilde{f}(\mathbf{a}, z) + \frac{1}{2\ell} \sum_{i=1}^m k_{ii}$$

where

$$\tilde{f}(\mathbf{a}, z) = \frac{\ell}{2} \left(\frac{m}{\ell} - \mu \right) \|\mathbf{a}\|^2 + \frac{\ell\mu}{2} z - m \langle \hat{\mathbf{a}}, \mathbf{a} \rangle + b \sum_{i=m+1}^{\ell} \left(-\frac{z+k_{ii}}{2} + \langle \mathbf{a}, \Phi(\mathbf{x}_i) \rangle \right)_+.$$

References

- Dimitri P. Bertsekas. *Nonlinear Programming, 2nd Edition*. Athena Scientific, 1999.
- Yuzhou Cao, Yitian Xu, and Junling Du. Multi-variable estimation-based safe screening rule for small sphere and large margin support vector machine. *Knowledge-Based Systems*, 191:105223, 2020.
- Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:1–27, 2011.
- Wei-Cheng Chang, Ching-Pei Lee, and Chih-Jen Lin. A revisit to support vector data description. 2014. URL <https://www.csie.ntu.edu.tw/~cjlin/papers/svdd.pdf>.
- Olivier Chapelle. Training a support vector machine in the primal. *Neural Comput.*, 19(5):1155–1178, 2007.
- Guijun Chen, Xueying Zhang, Zizhong John Wang, and Fenglian Li. Robust support vector data description for outlier detection with noise or uncertain data. *Knowledge-Based Systems*, 90:129–137, 2015.
- Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.
- Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.
- Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Res Log Quart.*, 3:95–110, 1956.
- Yu Guo, Jin Meng, Yaxing Li, Songhu Ge, Jinling Xing, and Hao Wu. Incremental small sphere and large margin for online recognition of communication jamming. *Applied Intelligence*, 50:3429–3440, 2020.
- Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35:17248–17262, 2022.

- Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.
- Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik. *Extracting Support Data for a Given Task*. In U.M. Fayyad & R. Uthurusamy (Eds.) Proceedings, First International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA:AAAI Press, 1995.
- Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS'99*, pages 582–588, Cambridge, MA, USA, 1999. MIT Press.
- Bernhard Schölkopf, Alex Smola, Robert Williamson, and Peter Bartlett. New support vector algorithms. *Neural computation*, 12:1207–45, 06 2000.
- Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443 – 1471, 2001.
- Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):p.3–30, 2011.
- Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer New York, 2008.
- Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(8):211–232, 2005.
- David M. J. Tax and Robert P. W. Duin. Data domain description using support vectors. In *Verleysen, M. (Ed.), Proc. European Symposium Artificial Neural Networks 1999. D. Facto, Brussel*, pages 251–256, 1999a.
- David M. J. Tax and Robert P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11):1191–1199, 1999b.
- 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, 12 2001.
- David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54: 45–66, 2004.
- Vladimir N. Vapnik. The nature of statistical learning theory. *Springer*, 1995.
- R Vinayakumar, K Poornachandran Soman, and Prabaharan Poornachandran. Applying convolutional neural network for network intrusion detection. In *2017 International conference on advances in computing, communications and informatics (ICACCI)*, pages 1222–1228. IEEE, 2017.
- Hongmei Wang and Yitian Xu. A safe double screening strategy for elastic net support vector machine. *Information Sciences*, 582:382–397, 2022.

- Xiaoming Wang, Fulai Chung, and Shitong Wang. Theoretical analysis for solution of support vector data description. *Neural Networks*, 24(4):360–369, 2011.
- Yahui Wang and Yitian Xu. A non-convex robust small sphere and large margin support vector machine for imbalanced data classification. *Neural Computing and Applications*, 35(4):3245–3261, 2023.
- Mingrui Wu and Jieping Ye. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2088–2092, 2009.
- Min Yuan and Yitian Xu. Bound estimation-based safe acceleration for maximum margin of twin spheres machine with pinball loss. *Pattern Recognition*, 114:107860, 2021.
- Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024.
- Tianyi Zhou, Matthew Lau, Jizhou Chen, Wenke Lee, and Xiaoming Huo. Optimal classification-based anomaly detection with neural networks: Theory and practice, 2024. URL <https://arxiv.org/abs/2409.08521>.