Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 19, No. 4, pp. 689–714, 2023

SPLITTING METHOD FOR SUPPORT VECTOR MACHINE WITH LOWER SEMI-CONTINUOUS LOSS[*]

Mingyu Mo and Qi Ye[†]

Abstract: In this paper, we study the splitting method for support vector machine in reproducing kernel Hilbert space with lower semi-continuous loss function. We equivalently transfer support vector machine in reproducing kernel Hilbert space with lower semi-continuous loss function to a finite-dimensional Optimization and propose the splitting method based on alternating direction method of multipliers. If the loss function is lower semi-continuous and subanalytic, we use the Kurdyka-Lojasiewicz property of the augmented Lagrangian function to show that the iterative sequence induced by this splitting method globally converges to a stationary point. The numerical experiments also demonstrate the effectiveness of the splitting method.

Key words: *support vector machine, lower semi-continuous loss function, reproducing kernel Hilbert space, splitting method, Kurdyka-Lojasiewicz property*

Mathematics Subject Classification: *68Q32, 49J52*

1 Introduction

Support vector machine (SVM) is a well-known model for binary classification in machine learning. In this paper, we discuss an Optimization induced from SVM (see [24, Chapter 5])

$$\inf_{f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{x}_i, y_i, f(\boldsymbol{x}_i)) + \lambda\|f\|_{\mathcal{H}}^2, \tag{1.1}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), $\{(\boldsymbol{x}_i, y_i) : i = 1, 2, ..., N\}$ is the training data, L is a loss function, and $\lambda > 0$ is the regularization paramater. The RKHS, the training data, the loss function and the regularization parameter in Optimization (1.1) are given before discussion. Optimization (1.1) is already achieved with convex loss function (see [24, Chapter 5]). After the success of convex loss function, people continue to investigate whether Optimization (1.1) can be feasible for nonconvex loss function. Recently, different nonconvex loss functions are proposed and used for traditional SVM (see [7, 9, 15, 18, 19, 22, 29]). But the algorithm of Optimization (1.1) with general nonconvex loss function is still lack of study. If the loss function is convex, then it is a lower semi-continuous loss function

[*]This program is supported by the National Natural Science Foundation of China under grants 12071157, 12026602, and the Natural Science Foundation of Guangdong 2020B1515310013.

[†]Corresponding author

while the lower semi-continuity can not imply the convexity. Currently, people are most interested in the infinite-dimensional spaces for application of machine learning such that the learning algorithm can be chosen from the enough large amounts of suitable solutions. In this paper, we consider Optimization (1.1) in an infinite-dimensional RKHS with lower semi-continuous loss function.

First we show that Optimization (1.1) has a minimizer in a finite-dimensional subspace spanned by the reproducing kernel basis related to the training data. Thus, Optimization (1.1) can be equivalently transferred to a finite-dimensional Optimization. From this equivalent Optimization, we discuss the splitting method based on alternating direction method of multipliers (ADMM) for Optimization (1.1). By this splitting method, we obtain two subproblems which are computable easily. Moreover, the convergence of ADMM is already guaranteed well for convex Optimizations (see [6]) and some special nonconvex Optimizations by the Kurdyka-Lojasiewicz (KL) property (see [11, 14]). To complete the proof, we reexchange the convergence theorems in [11, 14] and verify the convergence of this splitting method for Optimization (1.1) if loss function is lower semi-continuous and subanalytic for the global convergence to a stationary point and the error bound. Finally, we use the minimizer of Optimization (1.1) to build the SVM in RKHS to make prediction on some testing data.

This paper is organized as follows. We introduce the notations and preliminary materials of the SVM in RKHS in Section 2. Next, we discuss how to solve Optimization (1.1) by the splitting method based on ADMM in Section 3. Moreover, we discuss the global convergence and convergent rate of this splitting method for lower semi-continuous and subanalytic loss function in Section 4. Finally, we give some numerical examples of different loss functions and RKHSs for the synthetic data and the real data to show that the SVM in RKHS with lower semi-continuous and nonconvex loss function is better than the SVM in RKHS with convex loss function in some cases in Section 5.

2 Notations and Preliminaries

In this section, we review some notations and preliminaries of the SVM in RKHS. We denote the set of positive integers as $\mathbb{N}$ and the set of real numbers as $\mathbb{R}$, respectively. Also, we denote the d-dimensional Euclidean space as $\mathbb{R}^d$. For the sample space $X \subseteq \mathbb{R}^d$ and the label space $Y = \{+1, -1\}$, the training data

$$D := \{(\boldsymbol{x}_i, y_i) : i = 1, 2, ..., N\} \subseteq X \times Y$$

is composed of distinct input data $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \in X$ and output data $y_1, y_2, ..., y_N \in Y$. We will find a mapping $\mathcal{R} : X \to Y$ related to D such that $\mathcal{R}(\boldsymbol{x})$ is a good approximation of the response y to an arbitrary $\boldsymbol{x}$. For the rest of this paper, without specification, every vector is supposed to be a column vector.

SVM is an important class of mapping $\mathcal{R} : X \to Y$. The traditional SVM is to find a hyperplane in $\mathbb{R}^d$ that classifies all the training data in D correctly and creates the biggest margin. Then we construct $\mathcal{R} : X \to Y$ by this hyperplane and sign function. However, the hyperplane to separate D may not exist and we can only accept the hyperplane that misclassifies some training data. To help us define what we mean by "good", we introduce the loss function to find a hyperplane that achieves the smallest regularized empirical risk and its equivalent Optimization is considered in an RKHS consisting of all linear functions on X with an offset term (see [24, Section 1.3]). This RKHS is finite-dimensional which is isometrically isomorphic to $\mathbb{R}^d$.

Since the RKHS mentioned above is only a finite-dimensional space, in the rest of this paper, we shall discuss the SVM in an infinite-dimensional RKHS. Generally speaking, a Hilbert space $\mathcal{H}$ of functions $f : X \to \mathbb{R}$ equipped with the complete inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the reproducing kernel $K : X \times X \to \mathbb{R}$ is called an RKHS if it satisfies the following two conditions (see [26, Definition 10.1])

(i) $K(\boldsymbol{x}, \cdot) \in \mathcal{H}$ for all $\boldsymbol{x} \in X$,

(ii) $f(\boldsymbol{x}) = \langle f, K(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and all $\boldsymbol{x} \in X$.

In particular, for any $f, g \in \mathcal{H}$, the corresponding norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$ and the corresponding metric $\|f - g\|_{\mathcal{H}}$ are complete. Also, the reproducing kernel K of an RKHS $\mathcal{H}$ is uniquely determined. Conversely, for any kernel K on X, [24, Theorem 4.21] shows that there exists a unique RKHS $\mathcal{H}$ induced from K. Usually, we use some RKHSs induced from some common kernels, such as polynomial kernel, Gaussian kernel, Matérn kernel, and so on. For more flexible kernels such as those of Gaussian kernels, which belong to the most important kernels in practice, the offset term has neither a known theoretical nor an empirical advantage. In addition, the theoretical analysis is often substantially complicated by offset term. Thus, we decide to consider the SVM in RKHS without an offset term. Let us fix such an RKHS $\mathcal{H}$ and a real number $\lambda > 0$, we obtain Optimization (1.1)

$$\inf_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(\boldsymbol{x}_i, y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2,$$

where $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a loss function and $\lambda\|f\|_{\mathcal{H}}^2$ is the regularization term used to penalize f with the large RKHS norm. In the following, we will interpret $L(\boldsymbol{x}, y, f(\boldsymbol{x}))$ as the loss of predicting y by $f(\boldsymbol{x})$ if $\boldsymbol{x}$ is observed, that is, the smaller the value $L(\boldsymbol{x}, y, f(\boldsymbol{x}))$ is, the better $f(\boldsymbol{x})$ predicts y in the sense of L. It is easy to check that Optimization (1.1) is infinite-dimensional and nonnegative. We denote the minimizer of Optimization (1.1) as f_D. Next we use f_D to construct the SVM in RKHS as follows

$$\mathcal{R}(\boldsymbol{x}) = \begin{cases} +1, & f_D(\boldsymbol{x}) \geq 0, \\ -1, & f_D(\boldsymbol{x}) < 0. \end{cases}$$

Moreover, the classification rules constructed by different minimizers of Optimization (1.1) have no difference in performance. In conclusion, we just need to find a minimizer of Optimization (1.1). Next we discuss the splitting method for Optimization (1.1) in Section 3.

3 Splitting Method

In this section, we discuss the splitting method for Optimization (1.1). First we discuss the minimizer of Optimization (1.1). For binary classification, the most straightforward loss function is 0-1 loss function, which is an "ideal" loss function (see [8]). However, 0-1 loss function is bounded, nonconvex, and lower semi-continuous but discontinuous. Trying to optimize 0-1 loss function directly leads to a lower semi-continuous and nonconvex Optimization which is unable to be deal with by traditional optimization algorithms. Thus, some surrogate loss functions are proposed in the literature, such as convex loss function, that is, $t \mapsto L(\boldsymbol{x}, y, t)$ is a convex function on $\mathbb{R}$ for all $\boldsymbol{x} \in X$ and all $y \in Y$ (see [24, Definition 2.12]). Besides convexity, we can define other loss functions similarly, such as continuity, smoothness, lower semi-continuity, etc (see [24, Section 2.2]). Specially, if L is a convex loss function, then $t \mapsto L(\boldsymbol{x}, y, t)$ is continuous and thus lower semi-continuous on

$\mathbb{R}$ for all $\boldsymbol{x} \in X$ and all $y \in Y$. Hence, L is a lower semi-continuous loss function. But the lower semi-continuity can not imply the convexity.

Generally speaking, loss functions can be divided into two categories: convex loss functions and nonconvex loss functions. Convex loss functions including Hinge loss and square loss are the most commonly used. If L is a convex loss function, then the classical representer theorem [24, Theorem 5.5] assures that Optimization (1.1) has a unique minimizer f_D contained in a known finite-dimensional subspace spanned by the reproducing kernel basis and the training data, even if the space $\mathcal{H}$ itself is substantially larger. Thus, convex loss function is viewed as highly preferable in many publications because of its computational advantages (unique minimizer, ease-of-use, ability to be efficiently optimized by convex optimization tools, etc.).

However, the convexity also offers poor approximation to 0-1 loss function. Hence, different nonconvex loss functions, such as ramp loss, truncated logistic loss, truncated least square loss, truncated pinball loss, bi-truncated pinball loss, generalized exponential loss, generalized logistic loss and Sigmoid loss are proposed and used in SVM (see [7, 9, 15, 18, 19, 22, 29]). These loss functions mentioned above and 0-1 loss function are lower semi-continuous and nonconvex. Recently, [12, Proposition 3.3] generalizes the classical representer theorem [24, Theorem 5.5] to lower semi-continuous loss function. Moreover, by some preliminary numerical experiments, we find that the SVM in RKHS with lower semi-continuous and nonconvex loss function is better than convex loss function in some cases (see Section 5). Thus, we discuss the SVM in RKHS with lower semi-continuous loss function. Before we show our main result, we need some concepts of RKHS.

First an RKHS $\mathcal{H}$ can be seen as a Banach space. For any $\boldsymbol{x} \in X$, we denote $\delta_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$,

$$\delta_{\boldsymbol{x}}(f) := f(\boldsymbol{x})$$

as the point evaluation functional. [26, Theorem 10.2] shows that $\delta_{\boldsymbol{x}}$ is a linear continuous mapping from $\mathcal{H}$ to $\mathbb{R}$, that is, $\delta_{\boldsymbol{x}} \in \mathcal{H}^*$, where $\mathcal{H}^*$ denotes the dual space of $\mathcal{H}$. Moreover, the Riesz Representation Theorem shows that $\mathcal{H}^*$ is isometrically isomorphic to $\mathcal{H}$ which ensures that $\mathcal{H}$ is reflexive. For a function $\psi : \mathcal{H} \to \mathbb{R}$, ψ is Fréchet differentiable at $f \in \mathcal{H}$ if there exists a linear continuous functional $\nabla\psi(f) : \mathcal{H} \to \mathbb{R}$ such that

$$\lim_{h \to f} \frac{\psi(h) - \psi(f) - \nabla\psi(f)(h - f)}{\|h - f\|_{\mathcal{H}}} = 0.$$

We call $\nabla\psi(f)$ the Fréchet derivative of ψ at f (see [17, Page 19]). It is clear that Fréchet derivative is a generalization of gradient. Since $\mathcal{H}$ is also strictly convex and smooth, we obtain the following formula of the Fréchet derivative (see [27, Remark 2.24])

$$\nabla(\|\cdot\|_{\mathcal{H}})(f) = \left\langle \cdot, \frac{f}{\|f\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} \in \mathcal{H}^*, \ \forall f \neq 0. \tag{3.1}$$

Lemma 3.1. *If L is a lower semi-continuous loss function, then Optimization* (1.1) *has a minimizer f_D such that*

$$f_D \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\},$$

where $\operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$ *denotes the set of all finite linear combinations of* $\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$.

Proof. Let a generalized loss function $\bar{L} : \mathcal{H}^* \times Y \times \mathbb{R} \to [0, \infty)$,

$$\bar{L}(\delta_{\boldsymbol{x}}, y, t) := L(\boldsymbol{x}, y, t).$$

Therefore, [12, Proposition 3.3] ensures that Optimization (1.1) has a minimizer f_D. If $f_D \neq 0$, then [12, Proposition 3.3] and (3.1) show that

$$\nabla(\|\cdot\|_{\mathcal{H}})(f_D) = \left\langle \cdot, \frac{f_D}{\|f_D\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} \in \operatorname{span}\{\delta_{\boldsymbol{x}_1}, ..., \delta_{\boldsymbol{x}_N}\}. \tag{3.2}$$

From the reproducing property of the kernel K, for any $f \in \mathcal{H}$, we have that

$$\delta_{\boldsymbol{x}_i}(f) = f(\boldsymbol{x}_i) = \langle f, K(\boldsymbol{x}_i, \cdot)\rangle_{\mathcal{H}},\ i = 1, 2, ..., N.$$

Thus, $\delta_{\boldsymbol{x}_i} = \langle \cdot, K(\boldsymbol{x}_i, \cdot)\rangle_{\mathcal{H}},\ i = 1, 2, ..., N$. Moreover, from the linear property of $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, we see that

$$\operatorname{span}\{\delta_{\boldsymbol{x}_1}, ..., \delta_{\boldsymbol{x}_N}\} = \{\langle \cdot, f\rangle_{\mathcal{H}} : f \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}. \tag{3.3}$$

Since $\|f_D\|_{\mathcal{H}} > 0$, by (3.2) and (3.3), it follows that

$$f_D \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}.$$

If $f_D = 0$, then $0 \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$. In conclusion,

$$f_D \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}.$$

This proof is completed. □

Remark 3.2. If L is a lower semi-continuous and nonconvex loss function, then Optimization (1.1) may have more than one minimizer. Moreover, Lemma 3.1 guarantees that at least one of minimizers of Optimization (1.1) is contained in $\operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$. Thus, in this paper, we focus on finding the minimizer in $\operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$.

Next we show that Optimization (1.1) in $\operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$ also can be equivalently transferred to a finite-dimensional Optimization in $\mathbb{R}^N$. We denote the Gram matrix of kernel K for training data D as

$$A := \begin{pmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & K(\boldsymbol{x}_1, \boldsymbol{x}_2) & ... & K(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ K(\boldsymbol{x}_2, \boldsymbol{x}_1) & K(\boldsymbol{x}_2, \boldsymbol{x}_2) & ... & K(\boldsymbol{x}_2, \boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_N, \boldsymbol{x}_1) & K(\boldsymbol{x}_N, \boldsymbol{x}_2) & ... & K(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix}.$$

By [24, Theorem 4.16], A is a symmetric and positive definite matrix, that is,

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K(\boldsymbol{x}_j, \boldsymbol{x}_i),\ \ i, j = 1, 2, ..., N,$$

and for any $\boldsymbol{c} \in \mathbb{R}^N$, $\boldsymbol{c}^T A \boldsymbol{c} \geq 0$. For each $f \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$, there exists a vector $\boldsymbol{c} = (c_1, ..., c_N)^T \in \mathbb{R}^N$ such that f has the finite representation

$$f = \sum_{j=1}^{N} c_j K(\boldsymbol{x}_j, \cdot),$$

which ensures that

$$f(\boldsymbol{x}_i) = \sum_{j=1}^{N} c_j K(\boldsymbol{x}_j, \boldsymbol{x}_i) = \sum_{j=1}^{N} K(\boldsymbol{x}_i, \boldsymbol{x}_j) c_j = (A\boldsymbol{c})_i,\ i = 1, 2, ..., N.$$

On the other hand, combining the linear property of inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ with the reproducing property in the RKHS, it holds that

$$\|f\|_{\mathcal{H}}^2 = \langle f, \sum_{i=1}^{N} c_i K(\boldsymbol{x}_i, \cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^{N} c_i \langle f, K(\boldsymbol{x}_i, \cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^{N} c_i f(\boldsymbol{x}_i) = \boldsymbol{c}^T A \boldsymbol{c}. \tag{3.4}$$

So Optimization (1.1) can be equivalently transferred to the following Optimization in $\mathbb{R}^N$

$$\min_{\boldsymbol{c} \in \mathbb{R}^N} \ \frac{1}{N} \sum_{i=1}^{N} L(\boldsymbol{x}_i, y_i, (A\boldsymbol{c})_i) + \lambda \boldsymbol{c}^T A \boldsymbol{c}. \tag{3.5}$$

We denote the minimizer of Optimization (3.5) as $\boldsymbol{c}_D \in \mathbb{R}^N$, it follows that

$$f_D = \sum_{i=1}^{N} (\boldsymbol{c}_D)_i K(\boldsymbol{x}_i, \cdot).$$

This ensures that we can employ the finite suitable parameters to reconstruct the SVM in RKHS.

By this idea, we consider finding an algorithm based on Optimization (3.5) to solve Optimization (1.1) easily. At present, we mainly use subgradient method, Lagrangian multipliers method, and sequential minimal optimization (SMO) for SVM. These classical numerical algorithms are suitable for solving convex and smooth Optimizations. For notational convenience, let

$$F(\boldsymbol{\alpha}) := \frac{1}{N} \sum_{i=1}^{N} L(\boldsymbol{x}_i, y_i, \alpha_i), \ \ G(\boldsymbol{c}) := \lambda \boldsymbol{c}^T A \boldsymbol{c}.$$

Then Optimization (3.5) can be rewritten as

$$\min_{\boldsymbol{c} \in \mathbb{R}^N} \ F(A\boldsymbol{c}) + G(\boldsymbol{c}). \tag{3.5'}$$

Since L is a lower semi-continuous loss function, [2, Propositions 1.1.2 and 1.1.5] show that F is lower semi-continuous on $\mathbb{R}^N$. Moreover, [2, Propositions 1.1.2 and 1.1.4] assure that $\boldsymbol{c} \mapsto F(A\boldsymbol{c})$ is lower semi-continuous on $\mathbb{R}^N$. On the other hand, G is continuously differentiable on $\mathbb{R}^{\mathbb{N}}$ and for any $\boldsymbol{c} \in \mathbb{R}^N$,

$$\nabla G(\boldsymbol{c}) = 2\lambda A \boldsymbol{c}, \ \ \nabla^2 G(\boldsymbol{c}) = 2\lambda A, \tag{3.6}$$

where ∇ denotes the gradient and ∇^2 denotes the Hessian matrix, respectively. Since A is symmetric and positive definite, [2, Propositions 1.1.7 (a) and 1.1.10 (i)] assure that G is convex on $\mathbb{R}^N$ and

$$G(\boldsymbol{d}) \geq G(\boldsymbol{c}) + (\nabla G(\boldsymbol{c}))^T (\boldsymbol{d} - \boldsymbol{c}), \ \forall \boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^N. \tag{3.7}$$

In conclusion, Optimization (3.5) is lower semi-continuous which may be nonsmooth or nonconvex. Many classical algorithms are not suitable for Optimization (3.5) when L is a lower semi-continuous loss function.

The ADMM algorithm, as one of splitting techniques, has been successfully exploited in a wide range of structured regularization Optimizations in machine learning. ADMM can even be used to minimize nonsmooth or nonconvex function, which solves Optimization by breaking them into smaller pieces. Moreover, paper [25] discusses how to use ADMM for traditional SVM with 0-1 loss function. For the general SVM with lower semi-continuous loss function, we observe that the subproblems in ADMM for Optimization (3.5) can be transferred into some Optimizations in $\mathbb{R}$ and a well-posed linear system, each of them is easier to handle. Moreover, if the sample size N is small to moderate, then some preliminary numerical experiments show that the splitting method based on ADMM is a fast algorithm for Optimization (1.1) (see Section 5). Hence, we will study how to solve Optimization (1.1) by the splitting method based on ADMM.

To describe the algorithm, we first reformulate Optimization (3.5) as

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{c}} \quad & F(\boldsymbol{\alpha}) + G(\boldsymbol{c}), \\ \text{s.t.} \quad & \boldsymbol{\alpha} = A\boldsymbol{c}. \end{aligned} \tag{3.8}$$

Recall that the augmented Lagrangian function of Optimization (3.8) is defined as:

$$\mathcal{L}_\rho(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) := F(\boldsymbol{\alpha}) + G(\boldsymbol{c}) + \boldsymbol{\gamma}^T(\boldsymbol{\alpha} - A\boldsymbol{c}) + \frac{\rho}{2}\|\boldsymbol{\alpha} - A\boldsymbol{c}\|^2,$$

where the Lagrangian multiplier $\rho > 0$ and $\|\cdot\|$ denotes 2-norm in Euclidean space. The splitting method is presented as follows. Suppose that the algorithm is initialized at $(\boldsymbol{\alpha}^0, \boldsymbol{c}^0, \boldsymbol{\gamma}^0)$, its iterative scheme is

$$\boldsymbol{\alpha}^{k+1} \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\operatorname{argmin}}\ \mathcal{L}_\rho(\boldsymbol{\alpha}, \boldsymbol{c}^k, \boldsymbol{\gamma}^k), \tag{S-1}$$

$$\boldsymbol{c}^{k+1} \in \underset{\boldsymbol{c}\in\mathbb{R}^N}{\operatorname{argmin}}\ \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}, \boldsymbol{\gamma}^k), \tag{S-2}$$

$$\boldsymbol{\gamma}^{k+1} := \boldsymbol{\gamma}^k + \rho(\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c}^{k+1}), \tag{S-3}$$

$$s^{k+1} := \sum_{i=1}^N (\boldsymbol{c}^{k+1})_i K(\boldsymbol{x}_i, \cdot), \tag{S-4}$$

where k is an iteration counter. Since (S-1) only depends on $\boldsymbol{\alpha}$ and (S-2) only depends on $\boldsymbol{c}$, by combining the linear and quadratic terms in $\mathcal{L}_\rho$, we equivalently transfer (S-1) and (S-2) to

$$\boldsymbol{\alpha}^{k+1} \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\operatorname{argmin}}\ F(\boldsymbol{\alpha}) + \frac{\rho}{2}\|\boldsymbol{\alpha} - A\boldsymbol{c}^k + \frac{1}{\rho}\boldsymbol{\gamma}^k\|^2, \tag{S-1'}$$

$$\boldsymbol{c}^{k+1} \in \underset{\boldsymbol{c}\in\mathbb{R}^N}{\operatorname{argmin}}\ G(\boldsymbol{c}) + \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c} + \frac{1}{\rho}\boldsymbol{\gamma}^k\|^2. \tag{S-2'}$$

By definition, it is easy to check that (S-1') is lower semi-continuous on $\mathbb{R}^N$ and (S-2') is continuous on $\mathbb{R}^N$. Moreover, (S-1') and (S-2') are coercive (see [1, Definition 2.13]), that is,

$$\lim_{\|\boldsymbol{\alpha}\|\to\infty} F(\boldsymbol{\alpha}) + \frac{\rho}{2}\|\boldsymbol{\alpha} - A\boldsymbol{c}^k + \frac{1}{\rho}\boldsymbol{\gamma}^k\|^2 \to \infty, \quad \lim_{\|\boldsymbol{c}\|\to\infty} G(\boldsymbol{c}) + \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c} + \frac{1}{\rho}\boldsymbol{\gamma}^k\|^2 \to \infty.$$

Thus, Weierstrass Theorem [1, Theorem 2.14] assures that (S-1') and (S-2') both have a minimizer. As a consequence, this splitting method is well-defined and an infinite iterative sequence $\{(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k, s^k)\}$ is generated. Also, $\{s^k\}$ can be seen as a sequence to approximate the minimizer of Optimization (1.1).

Next, we discuss how to solve subproblems (S-1') and (S-2'). As for (S-1'), by definition, F and $\|\cdot\|^2$ can be split of the variable into subvectors, that is,

$$\frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{x}_i, y_i, \alpha_i) + \frac{\rho}{2}\|\boldsymbol{\alpha} - A\boldsymbol{c}^k + \frac{1}{\rho}\boldsymbol{\gamma}^k\|^2 = \sum_{i=1}^{N} \frac{L(\boldsymbol{x}_i, y_i, \alpha_i)}{N} + \frac{\rho}{2}\left(\alpha_i - (A\boldsymbol{c}^k)_i + \frac{1}{\rho}(\boldsymbol{\gamma}^k)_i\right)^2,$$

and

$$\frac{L(\boldsymbol{x}_i, y_i, \alpha_i)}{N} + \frac{\rho}{2}\left(\alpha_i - (A\boldsymbol{c}^k)_i + \frac{1}{\rho}(\boldsymbol{\gamma}^k)_i\right)^2 \geq 0, \;\; i = 1, 2, ..., N.$$

Thus, we can equivalently transfer an Optimization in $\mathbb{R}^N$ to some Optimizations in $\mathbb{R}$, that is,

$$(\boldsymbol{\alpha}^{k+1})_i \in \operatorname*{argmin}_{\alpha_i \in \mathbb{R}} \; \frac{L(\boldsymbol{x}_i, y_i, \alpha_i)}{N} + \frac{\rho}{2}\left(\alpha_i - (A\boldsymbol{c}^k)_i + \frac{1}{\rho}(\boldsymbol{\gamma}^k)_i\right)^2, \; i = 1, 2, ..., N. \tag{S-1''}$$

In other words, we solve (S-1') in $\mathbb{R}^N$ by breaking it into N Optimizations (S-1") in $\mathbb{R}$ and each of them is easier to handle. For the general lower semi-continuous loss function L, the solution set of (S-1") may not be a singleton. In this case, we choose one of the elements in the solution set as $(\boldsymbol{\alpha}^{k+1})_i, \; i = 1, ..., N$.

As for (S-2'), since A is symmetric and positive definite and $\lambda, \rho > 0$, it is clear that (S-2') is nonnegative and continuously differentiable. By derivative rule and (3.6), we have that for any $\boldsymbol{c} \in \mathbb{R}^N$,

$$\begin{aligned}\nabla(G + \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A(\cdot) + \frac{1}{\rho}\boldsymbol{\gamma}^k\|^2)(\boldsymbol{c}) &= 2\lambda A\boldsymbol{c} - \rho A(\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c} + \frac{1}{\rho}\boldsymbol{\gamma}^k) \\ &= A\left((2\lambda I + \rho A)\boldsymbol{c} - \rho\boldsymbol{\alpha}^{k+1} - \boldsymbol{\gamma}^k\right),\end{aligned}$$

where I is the identity matrix with an order N and

$$\nabla^2(G + \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A(\cdot) + \frac{1}{\rho}\boldsymbol{\gamma}^k\|^2)(\boldsymbol{c}) = 2\lambda A + \rho A^2.$$

Since A is symmetric and positive definite, $2\lambda A + \rho A^2$ is also symmetric and positive definite. Hence, [2, Propositions 1.1.7 (a) and 1.1.10 (i)] assure that (S-2') is convex on $\mathbb{R}^N$. Moreover, since $2\lambda I + \rho A$ is symmetric and strictly positive definite and thus nonsingular, the following linear system

$$(2\lambda I + \rho A)\boldsymbol{c} = \rho\boldsymbol{\alpha}^{k+1} + \boldsymbol{\gamma}^k \tag{S-2''}$$

has a unique solution $\boldsymbol{c}^{k+1}$. Furthermore, since (S-2') is convex and differentiable on $\mathbb{R}^N$ and

$$A\left((2\lambda I + \rho A)\boldsymbol{c}^{k+1} - \rho\boldsymbol{\alpha}^{k+1} - \boldsymbol{\gamma}^k\right) = \mathbf{0},$$

it is clear that $\boldsymbol{c}^{k+1}$ is a stationary point of (S-2') and thus a minimizer of (S-2'). Next we consider using conjugate gradient method for the well-posed linear system (S-2") (see [13, Section 4.7.3]).

When $\boldsymbol{\alpha}^{k+1}$ and $\boldsymbol{c}^{k+1}$ are acquired, we can obtain $\boldsymbol{\gamma}^{k+1}$ by (S-3). However, we have a simpler one in mind that accomplishes the same goal. Substituting (S-3) into (S-2") and rearranging terms, we have that

$$\boldsymbol{\gamma}^{k+1} = 2\lambda\boldsymbol{c}^{k+1}. \tag{S-3'}$$

Moreover, combining (3.6) with (S-3'), it follows that

$$\nabla G(\boldsymbol{c}^{k+1}) = 2\lambda A\boldsymbol{c}^{k+1} = A\boldsymbol{\gamma}^{k+1}. \tag{3.9}$$

In conclusion, the splitting method for Optimization (1.1) can be represented as follows:

Algorithm 1 Splitting Method for the SVM in RKHS with Lower Semi-continuous Loss Function

```
input: initial value $(\boldsymbol{\alpha}^0, \boldsymbol{c}^0, \boldsymbol{\gamma}^0)$, the training data $D$, loss function $L$, the Gram matrix $A$, regularization parameter $\lambda > 0$, Lagrangian multiplier $\rho > 0$ and stopping threshold $\varepsilon_0 > 0$.
for $k = 0, 1, 2, ...$ do
    (1) Choose $\boldsymbol{\alpha}^{k+1}$ in $\operatorname*{argmin}_{\alpha_i} \dfrac{L(\boldsymbol{x}_i, y_i, \alpha_i)}{N} + \dfrac{\rho}{2}\|\alpha_i - (A\boldsymbol{c}^k)_i + \dfrac{1}{\rho}(\boldsymbol{\gamma}^k)_i\|^2,\ i = 1, 2, ..., N.$
    (2) Let $\boldsymbol{c}^{k(0)} = \boldsymbol{c}^k$, $\boldsymbol{u}^{k(0)} = (2\lambda I + \rho A)\boldsymbol{c}^{k(0)} - (\rho\alpha^{k+1} + \boldsymbol{\gamma}^k), \boldsymbol{d}^{k(0)} = -\boldsymbol{u}^{k(0)}$.
    for $j = 0, 1, 2, ...$ do
        (2-1) Set $\boldsymbol{c}^{k(j+1)} \leftarrow \boldsymbol{c}^{k(j)} + \dfrac{\|\boldsymbol{u}^{k(j)}\|^2}{(\boldsymbol{d}^{k(j)})^T(2\lambda I + \rho A)\boldsymbol{d}^{k(j)}}\boldsymbol{d}^{k(j)}$.
        (2-2) Set $\boldsymbol{u}^{k(j+1)} \leftarrow \boldsymbol{u}^{k(j)} - \dfrac{\|\boldsymbol{u}^{k(j)}\|^2}{(\boldsymbol{d}^{k(j)})^T(2\lambda I + \rho A)\boldsymbol{d}^{k(j)}}(2\lambda I + \rho A)\boldsymbol{d}^{k(j)}$.
        (2-3) Set $\boldsymbol{d}^{k(j+1)} \leftarrow -\boldsymbol{u}^{k(j+1)} + \dfrac{\|\boldsymbol{u}^{k(j+1)}\|^2}{\|\boldsymbol{u}^{k(j)}\|^2}\boldsymbol{d}^{k(j)}$.
        if $\boldsymbol{u}^{k(j+1)} = \boldsymbol{0}$ then stop.
    end for
    output: The approximate solution $\boldsymbol{c}^{k(j+1)}$ as $\boldsymbol{c}^{k+1}$.
    (3) Set $\boldsymbol{\gamma}^{k+1} \leftarrow 2\lambda\boldsymbol{c}^{k+1}$.
    (4) Set $s^{k+1} \leftarrow \sum_{i=1}^{N}(\boldsymbol{c}^{k+1})_i K(\boldsymbol{x}_i, \cdot)$.
    if $\|\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c}^{k+1}\| < \varepsilon_0$ then stop.
end for
output: The approximate solution $s^{k+1}$.
```

In Section 4, we verify that under some mild assumption, $\{s^k\}$ globally converges to a stationary point of Optimization (1.1). In particular, if L is a convex loss function, then $\{s^k\}$ is globally convergent to the minimizer f_D. If L is a lower semi-continuous and nonconvex loss function, then $\{s^k\}$ may converge to a stationary point rather than a minimizer. Hence, it is better to solve Optimization (1.1) repeatedly by selecting some initial values randomly and choosing the minimizer of these outputs as the approximate solution s_D of Optimization (1.1). Finally, we construct the classification rule by s_D, that is,

$$\mathcal{R}(\boldsymbol{x}) = \begin{cases} +1, & s_D(\boldsymbol{x}) \geq 0, \\ -1, & s_D(\boldsymbol{x}) < 0. \end{cases}$$

We will complete the convergence analysis of Algorithm 1 in Section 4.

4 Convergence Analysis

In this section, we discuss the convergence of $\{s^k\}$ inspired by the work [11, 14] and use similar line of arguments therein. To ensure the convergence, we need the following assumption of Optimization (1.1).

Assumption 4.1. For Optimization (1.1), the following conditions hold

(i) L is a lower semi-continuous and subanalytic loss function.
(ii) A is a symmetric and strictly positive definite matrix.

Before we show our main result in this section, we discuss what conclusions can be drawn under Assumption 4.1. Subanalytic functions are quite wide, including semi-algebraic, analytic, and semi-analytic functions (see [10, 6.6 Analytic Problems]). More precisely, polynomial functions and piecewise polynomial functions are subanalytic functions. However, subanalyticity does not even imply continuity. Moreover, some margin-based loss functions satisfy Assumption 4.1 (i), such as the least square loss, the Hinge loss, the truncated least square loss, logistic loss, and so on (see [24, Section 2.3]). Since F, G and $(\boldsymbol{\alpha}, \boldsymbol{c}) \mapsto \frac{\rho}{2}\|\boldsymbol{\alpha} - A\boldsymbol{c}\|^2$ are nonnegative, lower semi-continuous and subanalytic and $(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) \mapsto \boldsymbol{\gamma}^T(\boldsymbol{\alpha} - A\boldsymbol{c})$ is continuous, subanalytic and bounded for any bounded set in $\mathbb{R}^{3N}$, [23, (I.2.1.9)] shows that $\mathcal{L}_\rho$ is lower semi-continuous and subanalytic. Moreover, [3, 4, 28] assure that $\mathcal{L}_\rho$ is a KL function on $\mathbb{R}^{3N}$, that is, $\mathcal{L}_\rho$ has KL property at each point in $\mathbb{R}^{3N}$ (see [5, Section 2.4]). The KL property of $\mathcal{L}_\rho$ plays a crucial role in estimating the error bound of the iterative sequence.

In some cases, A can be symmetric and strictly positive definite. For instance, the Gram matrix A of strictly positive definite kernel for any training data D is always symmetric and strictly positive definite. Gaussian kernel and Matérn kernel are the most common strictly positive definite kernels. Suppose that A is symmetric and strictly positive definite and for any $f \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$, there exists $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^N$ such that

$$f = \sum_{i=1}^{N} c_i K(\boldsymbol{x}_i, \cdot) = \sum_{i=1}^{N} d_i K(\boldsymbol{x}_i, \cdot),$$

which ensures that

$$(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), ..., f(\boldsymbol{x}_N))^T = A\boldsymbol{c} = A\boldsymbol{d}.$$

Since A is symmetric and strictly positive definite and thus nonsingular, it follows that $\boldsymbol{c} = \boldsymbol{d}$. Thus, f has the unique finite representation

$$f = \sum_{i=1}^{N} c_i K(\boldsymbol{x}_i, \cdot).$$

Moreover, we denote $\boldsymbol{\delta} : \mathcal{H} \to \mathbb{R}^N$,

$$\boldsymbol{\delta}(f) := (\delta_{\boldsymbol{x}_1}(f), \delta_{\boldsymbol{x}_2}(f), ..., \delta_{\boldsymbol{x}_N}(f))^T.$$

We see that for any $f \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$,

$$\boldsymbol{\delta}(f) = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), ..., f(\boldsymbol{x}_N))^T = A\boldsymbol{c}.$$

Hence, it is easy to check that $\boldsymbol{\delta}$ is a linear mapping from $\operatorname{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$ onto $\mathbb{R}^N$ and thus an isomorphism by [16, 1.4.15 Theorem]. Since $s^k = \sum_{i=1}^{N}(\boldsymbol{c}^k)_i K(\boldsymbol{x}_i, \cdot)$ for any $k \in \mathbb{N}$, it is clear that

$$\boldsymbol{\delta}(s^{k+1} - s^k) = \boldsymbol{\delta}\left(\sum_{i=1}^{N}(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)_i K(\boldsymbol{x}_i, \cdot)\right) = A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k).$$

Therefore, [16, 1.4.14 Proposition (i)] shows that there exists $0 < w_1 \leq w_2$ such that

$$w_1\|s^{k+1} - s^k\|_{\mathcal{H}} \leq \|\boldsymbol{\delta}(s^{k+1} - s^k)\| = \|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\| \leq w_2\|s^{k+1} - s^k\|_{\mathcal{H}}. \tag{4.1}$$

Now we begin to show our main result in this section.

Theorem 4.1. *Suppose that Assumption* 4.1 *holds and Algorithm* 1 *is initialized at* $(\boldsymbol{\alpha}^0, \boldsymbol{c}^0, \boldsymbol{\gamma}^0)$. *If* $\rho > 4\lambda\|A^{-1}\|$, *then* $\{s^k\}$ *converges to a stationary point* s^* *of Optimization* (1.1).

Before we prove our main result, we need two lemmas about $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ and $\{s^k\}$.

Lemma 4.2. *If the conditions in Theorem* 4.1 *hold, then there exists* $\zeta_1 > 0$ *such that*

$$\zeta_1\|s^{k+1} - s^k\|_{\mathcal{H}}^2 \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^{k+1}).$$

Proof. From (S-1), we know that $\boldsymbol{\alpha}^{k+1}$ is the minimizer of $\boldsymbol{\alpha} \mapsto \mathcal{L}_\rho(\boldsymbol{\alpha}, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)$, that is,

$$0 \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^k, \boldsymbol{\gamma}^k). \tag{4.2}$$

Similarily from the definition of $\mathcal{L}_\rho$ and using (S-3) and (3.9), we see that

$$\begin{aligned}
&\mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^k)\\
=&G(\boldsymbol{c}^k) - G(\boldsymbol{c}^{k+1}) + (\boldsymbol{\gamma}^k)^T A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k) + \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c}^k\|^2 - \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c}^{k+1}\|^2\\
=&G(\boldsymbol{c}^k) - G(\boldsymbol{c}^{k+1}) + (\boldsymbol{\gamma}^k)^T A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k) + \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c}^{k+1}\|^2 + \frac{\rho}{2}\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2\\
&+ \rho(\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c}^{k+1})^T A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k) - \frac{\rho}{2}\|\boldsymbol{\alpha}^{k+1} - A\boldsymbol{c}^{k+1}\|^2\\
=&G(\boldsymbol{c}^k) - G(\boldsymbol{c}^{k+1}) + (\boldsymbol{\gamma}^{k+1})^T A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k) + \frac{\rho}{2}\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2\\
=&G(\boldsymbol{c}^k) - G(\boldsymbol{c}^{k+1}) - (2\lambda A\boldsymbol{c}^{k+1})^T(\boldsymbol{c}^k - \boldsymbol{c}^{k+1}) + \frac{\rho}{2}\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2\\
=&G(\boldsymbol{c}^k) - G(\boldsymbol{c}^{k+1}) - (\nabla G(\boldsymbol{c}^{k+1}))^T(\boldsymbol{c}^k - \boldsymbol{c}^{k+1}) + \frac{\rho}{2}\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2.
\end{aligned}$$

From the equation above and (3.7), we have that

$$\frac{\rho}{2}\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2 \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^k). \tag{4.3}$$

Furthermore, from the definition of $\mathcal{L}_\rho$ and using (S-3), it follows that

$$-\frac{1}{\rho}\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k\|^2 = \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^{k+1}).$$

By (S-3'), it is easy to check that

$$\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k\| = 2\lambda\|\boldsymbol{c}^{k+1} - \boldsymbol{c}^k\| = 2\lambda\|A^{-1}A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\| \leq 2\lambda\|A^{-1}\|\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|.$$

Combining with two relations above, we see that

$$-\frac{4\lambda^2\|A^{-1}\|^2}{\rho}\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2 \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^{k+1}). \tag{4.4}$$

Hence, (4.2), (4.3) and (4.4) show that

$$\left(\frac{\rho}{2} - \frac{4\lambda^2\|A^{-1}\|^2}{\rho}\right)\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2 \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^{k+1}).$$

Since $\rho > 4\lambda\|A^{-1}\|$, we have that

$$\frac{\rho}{2} - \frac{4\lambda^2\|A^{-1}\|^2}{\rho} = \frac{\rho^2 - 8\lambda^2\|A^{-1}\|^2}{2\rho} > \frac{16\lambda^2\|A^{-1}\|^2 - 8\lambda^2\|A^{-1}\|^2}{2\rho} = \frac{4\lambda^2\|A^{-1}\|^2}{\rho} > 0.$$

Thus,

$$\frac{4\lambda^2\|A^{-1}\|^2}{\rho}\|A(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)\|^2 \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^{k+1}).$$

Let $\zeta_1 := \dfrac{4\lambda^2\|A^{-1}\|^2(w_1)^2}{\rho}$. Then $\zeta_1 > 0$. Therefore, the inequality above and (4.1) show that

$$\zeta_1\|s^{k+1} - s^k\|_{\mathcal{H}}^2 \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \mathcal{L}_\rho(\boldsymbol{\alpha}^{k+1}, \boldsymbol{c}^{k+1}, \boldsymbol{\gamma}^{k+1}).$$

This proof is completed. □

Lemma 4.3. *If the conditions in Theorem* 4.1 *hold, then* $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ *is monotonically decreasing, bounded and convergent.*

Proof. First Lemma 4.2 shows that $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ is monotonically decreasing and for any $k \in \mathbb{N}$,

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^0, \boldsymbol{c}^0, \boldsymbol{\gamma}^0) \geq \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) = F(\boldsymbol{\alpha}^k) + G(\boldsymbol{c}^k) - \frac{1}{2\rho}\|\boldsymbol{\gamma}^k\|^2 + \frac{\rho}{2}\|\boldsymbol{\alpha}^k - A\boldsymbol{c}^k + \frac{1}{\rho}(\boldsymbol{\gamma}^k)\|^2. \quad (4.5)$$

Moreover, since A is symmetric and strictly positive definite, the minimum eigenvalue of A is $\dfrac{1}{\|A^{-1}\|}$ which is its largest possible strong convexity parameter. Thus, [1, Example 5.19 and Theorem 5.24 (iii)] show that

$$G(\boldsymbol{c}^k) = \lambda(\boldsymbol{c}^k - \mathbf{0})^T(A\boldsymbol{c}^k - \mathbf{0}) \geq \frac{\lambda}{\|A^{-1}\|}\|\boldsymbol{c}^k\|^2.$$

Since $F(\boldsymbol{\alpha}^k) \geq 0$ and $\rho > 4\lambda\|A^{-1}\|$, (S-3') provides that

$$G(\boldsymbol{c}^k) - \frac{1}{2\rho}\|\boldsymbol{\gamma}^k\|^2 \geq \frac{\lambda}{\|A^{-1}\|}\|\boldsymbol{c}^k\|^2 - \frac{2\lambda^2}{\rho}\|\boldsymbol{c}^k\|^2 \geq \frac{\lambda}{2\|A^{-1}\|}\|\boldsymbol{c}^k\|^2 \geq 0. \quad (4.6)$$

Hence, (4.5) and (4.6) show that

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^0, \boldsymbol{c}^0, \boldsymbol{\gamma}^0) \geq \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) \geq 0,$$

which ensures that $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ is bounded. Thus, [21, Theorem 3.24] shows that $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ is convergent. This proof is completed. □

We denote the residual of $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ as

$$r^k := \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - \lim_{k\to\infty}\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k).$$

Then Lemma 4.2 shows that $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ is monotonically decreasing and the descent inequality can be rewritten as

$$\zeta_1\|s^{k+1} - s^k\|_{\mathcal{H}}^2 \leq r^k - r^{k+1}. \quad (4.7)$$

Also, for any $k \in \mathbb{N}$, $r^k \geq 0$ and $\{r^k\}$ is monotonically decreasing and convergent to 0.

We are now ready for proving the main result of this section.

Proof of Theorem 4.1. We consider the following two cases:

(I) If there exists $k_0 \in \mathbb{N}$ for which $r^{k_0} = 0$, then for any $k > k_0 + 1$, $r^k = r^{k+1} = 0$. Therefore, (4.7) shows that

$$\zeta_1 \|s^{k+1} - s^k\|_{\mathcal{H}}^2 \leq r^k - r^{k+1} = 0.$$

Since $\zeta_1 > 0$, it means that $s^{k+1} = s^k$. Hence, $\{s^k\}$ is convergent.

(II) If $r^k > 0$ for any $k \in \mathbb{N}$, then we verify the convergence of $\{s^k\}$ by the KL property of $\mathcal{L}_\rho$. First we show that $\{(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ is a bounded sequence. From (4.5) and (4.6), we have that

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^0, \boldsymbol{c}^0, \boldsymbol{\gamma}^0) \geq \frac{\lambda}{2\|A^{-1}\|} \|\boldsymbol{c}^k\|^2 \geq 0.$$

It means that $\{\boldsymbol{c}^k\}$ is bounded which ensures $\{\boldsymbol{\gamma}^k\}$ is also bounded by (S-3'). Furthermore, (S-3) shows that

$$\|\boldsymbol{\alpha}^k\| \leq \|A\boldsymbol{c}^k\| + \frac{1}{\rho}\|\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1}\| \leq \|A\|\|\boldsymbol{c}^k\| + \frac{1}{\rho}(\|\boldsymbol{\gamma}^k\| + \|\boldsymbol{\gamma}^{k-1}\|).$$

Hence, $\{\boldsymbol{\alpha}^k\}$ is bounded. In conclusion, $\{(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ is bounded. Let S be the set of subsequential limits of $\{(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$. [21, Theorems 3.6 and 3.7] show that S is nonempty compact, and

$$\lim_{k\to\infty} \operatorname{dist}\big((\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k), S\big) = 0, \tag{4.8}$$

where $\operatorname{dist}(\cdot,\cdot)$ denotes Euclidean distance. Since $\mathcal{L}_\rho$ is a KL function on $\mathbb{R}^{3N}$, by the definition of KL function, $\mathcal{L}_\rho$ is a KL function on S. Moreover, we show that $\mathcal{L}_\rho$ is constant on S. For any $(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*) \in S$, there exists a subsequence $\{(\boldsymbol{\alpha}^{k_j}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j})\}$ that converges to $(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*)$. Since $\mathcal{L}_\rho$ is lower semi-continuous on $\mathbb{R}^{3N}$, the lower semi-continuity of $\mathcal{L}_\rho$ at $(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*)$ and Lemma 4.3 show that

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*) \leq \liminf_{j\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}) = \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k). \tag{4.9}$$

Conversely, since $\boldsymbol{\alpha}^{k_j+1}$ is a minimizer of $\boldsymbol{\alpha} \mapsto \mathcal{L}_\rho(\boldsymbol{\alpha}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j})$, it shows that

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^*, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}) \geq \mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j+1}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}). \tag{4.10}$$

From the continuity of $\mathcal{L}_\rho$ with respect to $\boldsymbol{c}$ and $\boldsymbol{\gamma}$, it holds that

$$\lim_{j\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^*, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}) = \mathcal{L}_\rho(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*). \tag{4.11}$$

By (4.2), (4.3) and (4.4), we have that

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j+1}, \boldsymbol{c}^{k_j+1}, \boldsymbol{\gamma}^{k_j+1}) \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j+1}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}) \leq \mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}).$$

Since $\lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}) = \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j+1}, \boldsymbol{c}^{k_j+1}, \boldsymbol{\gamma}^{k_j+1}) = \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)$, we verify that

$$\lim_{j\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^{k_j+1}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j}) = \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k). \tag{4.12}$$

By [21, Theorem 3.19], (4.10), (4.11) and (4.12) show that

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*) \geq \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k). \tag{4.13}$$

Finally, (4.9) and (4.13) assure that

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*) = \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k).$$

Hence, $\mathcal{L}_\rho$ is constant on S.

Next, we use the uniformized KL property of $\mathcal{L}_\rho$ on S to derive an important inequality of the sequence $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$. [5, Lemma 3.6] assures that there exist $\varepsilon > 0$, $\eta > 0$ and a nonnegative continuous concave function $\varphi : [0, \eta) \to (0, +\infty)$ related to KL property such that

(i) $\varphi(0) = 0$ and φ is continuously differentiable on $(0, \eta)$ with positive derivatives φ'.

(ii) if $\operatorname{dist}((\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}), S) < \varepsilon$ and $\lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) < \mathcal{L}_\rho(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) < \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) +$ η, then

$$\varphi^{'}(\mathcal{L}_\rho(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) - \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)) \operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma})) \geq 1.$$

where ∂ denotes the limiting subdifferential (see [20, Definition 8.3]). Since $r^k > 0$, from Lemma 4.3 and (4.8), it suffices to show that for $\varepsilon > 0$ and $\eta > 0$ above, there exists $k_1 \in \mathbb{N}$ such that for any $k > k_1$,

$$\varphi^{'}(r^k) \operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)) \geq 1. \tag{4.14}$$

From the concavity of φ, we get that

$$\varphi^{'}(r^k)(r^k - r^{k+1}) \leq \varphi(r^k) - \varphi(r^{k+1}).$$

Multiplying $\operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k))$ on both side of the above inequality and using (4.14), we obtain

$$r^k - r^{k+1} \leq \operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k))(\varphi(r^k) - \varphi(r^{k+1})). \tag{4.15}$$

Next we make an estimation on the upper bound of $\operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k))$. By [20, 8.8 Exercise (c) and 10.5 Proposition], it follows that

$$\partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) = \{\partial F(\boldsymbol{\alpha}^k) + \boldsymbol{\gamma}^k + \rho(\boldsymbol{\alpha}^k - A\boldsymbol{c}^k),\ 2\lambda A\boldsymbol{c}^k - A\boldsymbol{\gamma}^k - \rho A(\boldsymbol{\alpha}^k - A\boldsymbol{c}^k),\ \boldsymbol{\alpha}^k - A\boldsymbol{c}^k\}.$$

Invoking the optimality condition for (S-1'), we have that

$$-\rho\left(\boldsymbol{\alpha}^k - A\boldsymbol{c}^{k-1} + \frac{1}{\rho}\boldsymbol{\gamma}^{k-1}\right) \in \partial F(\boldsymbol{\alpha}^k). \tag{4.16}$$

We denote

$$\begin{aligned}
\boldsymbol{\alpha}_k^\# &:= -\rho\left(\boldsymbol{\alpha}^k - A\boldsymbol{c}^{k-1} + \frac{1}{\rho}\boldsymbol{\gamma}^{k-1}\right) + \boldsymbol{\gamma}^k + \rho(\boldsymbol{\alpha}^k - A\boldsymbol{c}^k),\\
\boldsymbol{c}_k^\# &:= 2\lambda A\boldsymbol{c}^k - A\boldsymbol{\gamma}^k - \rho A(\boldsymbol{\alpha}^k - A\boldsymbol{c}^k),\\
\boldsymbol{\gamma}_k^\# &:= \boldsymbol{\alpha}^k - A\boldsymbol{c}^k.
\end{aligned}$$

Hence, $(\boldsymbol{\alpha}_k^\#, \boldsymbol{c}_k^\#, \boldsymbol{\gamma}_k^\#) \in \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)$, which ensures that

$$\operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)) = \inf_{\substack{(\boldsymbol{\alpha},\boldsymbol{c},\boldsymbol{\gamma})\in\\ \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k,\boldsymbol{c}^k,\boldsymbol{\gamma}^k)}} \|(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma})\| \leq \|(\boldsymbol{\alpha}_k^\#, \boldsymbol{c}_k^\#, \boldsymbol{\gamma}_k^\#)\| \leq \|\boldsymbol{\alpha}_k^\#\| + \|\boldsymbol{c}_k^\#\| + \|\boldsymbol{\gamma}_k^\#\|. \tag{4.17}$$

Inserting (S-3) and (S-3') into $\boldsymbol{\alpha}_k^{\#}$, $\boldsymbol{c}_k^{\#}$ and $\boldsymbol{\gamma}_k^{\#}$, we verify that

$$\begin{aligned}\boldsymbol{\alpha}_k^{\#} &= -\rho A(\boldsymbol{c}^k - \boldsymbol{c}^{k-1}) + \boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1},\\ \boldsymbol{c}_k^{\#} &= -A(\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1}),\\ \boldsymbol{\gamma}_k^{\#} &= \frac{1}{\rho}(\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1}).\end{aligned}$$

From (S-3') and (4.1), we see that

$$\|\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1}\| = 2\lambda\|\boldsymbol{c}^k - \boldsymbol{c}^{k-1}\| \le 2\lambda\|A^{-1}\|\|A(\boldsymbol{c}^k - \boldsymbol{c}^{k-1})\| \le 2\lambda w_2\|A^{-1}\|\|s^k - s^{k-1}\|_{\mathcal{H}},$$
$$\|A(\boldsymbol{c}^k - \boldsymbol{c}^{k-1})\| \le w_2\|s^k - s^{k-1}\|_{\mathcal{H}}.$$

Hence,

$$\begin{aligned}\|\boldsymbol{\alpha}_k^{\#}\| &\le \left(\rho w_2 + 2\lambda w_2\|A^{-1}\|\right)\|s^k - s^{k-1}\|_{\mathcal{H}},\\ \|\boldsymbol{c}_k^{\#}\| &\le 2\lambda w_2\|A\|\|A^{-1}\|\|s^k - s^{k-1}\|_{\mathcal{H}},\\ \|\boldsymbol{\gamma}_k^{\#}\| &\le \frac{2\lambda w_2\|A^{-1}\|}{\rho}\|s^k - s^{k-1}\|_{\mathcal{H}}.\end{aligned} \tag{4.18}$$

Combining (4.17) with (4.18), we see that there exists $\zeta_2 > 0$ such that

$$\operatorname{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)) \le \zeta_2\|s^k - s^{k-1}\|_{\mathcal{H}}. \tag{4.19}$$

Finally, from (4.7), (4.15) and (4.19), we obtain an inportant inequality of $\{s^k\}$ which can be used to verify the convergence of $\{s^k\}$. Since $\zeta_1, \zeta_2 > 0$, by rearranging terms, whenever $k > k_1 + 1$, (4.7), (4.15) and (4.19) assure that

$$\begin{aligned}\|s^{k+1} - s^k\|_{\mathcal{H}} &\le \sqrt{\frac{\zeta_2}{\zeta_1}\|s^k - s^{k-1}\|_{\mathcal{H}}\left(\varphi(r^k) - \varphi(r^{k+1})\right)}\\ &\le \frac{\|s^k - s^{k-1}\|_{\mathcal{H}} + \frac{\zeta_2}{\zeta_1}\left(\varphi(r^k) - \varphi(r^{k+1})\right)}{2}.\end{aligned}$$

By rearranging terms, we obtain further that

$$\|s^{k+1} - s^k\|_{\mathcal{H}} \le \|s^k - s^{k-1}\|_{\mathcal{H}} - \|s^{k+1} - s^k\|_{\mathcal{H}} + \frac{\zeta_2}{\zeta_1}(\varphi(r^k) - \varphi(r^{k+1})). \tag{4.20}$$

For any $l > k_1$, summing up the above relation from $k = k_1 + 1, ..., l$, since $\|s^{l+1} - s^l\|_{\mathcal{H}} \ge 0$ and $\varphi(r^{l+1}) > 0$, we see that

$$\begin{aligned}\sum_{k=k_1+1}^{l}\|s^{k+1} - s^k\|_{\mathcal{H}} &\le \|s^{k_1+1} - s^{k_1}\|_{\mathcal{H}} - \|s^{l+1} - s^l\|_{\mathcal{H}} + \frac{\zeta_2}{\zeta_1}(\varphi(r^{k_1+1}) - \varphi(r^{l+1}))\\ &\le \|s^{k_1+1} - s^{k_1}\|_{\mathcal{H}} + \frac{\zeta_2}{\zeta_1}\varphi(r^{k_1+1}) < \infty.\end{aligned}$$

Therefore,

$$\sum_{k=k_1+1}^{\infty}\|s^{k+1} - s^k\|_{\mathcal{H}} \le \|s^{k_1+1} - s^{k_1}\|_{\mathcal{H}} + \frac{\zeta_2}{\zeta_1}\varphi(r^{k_1+1}) < \infty,$$

which ensures that

$$\sum_{k=1}^{\infty}\|s^{k+1}-s^k\|_{\mathcal{H}}=\sum_{k=1}^{k_1}\|s^{k+1}-s^k\|_{\mathcal{H}}+\sum_{k=k_1+1}^{\infty}\|s^{k+1}-s^k\|_{\mathcal{H}}<\infty. \tag{4.21}$$

Moreover, we see that

$$\lim_{j\to\infty}\sum_{k=j}^{\infty}\|s^{k+1}-s^k\|_{\mathcal{H}}=0.$$

For any $\varepsilon_1>0$, there exists $k_2\in\mathbb{N}$ such that for any $l>j>k_2$,

$$\|s^l-s^j\|_{\mathcal{H}}\le\sum_{k=j}^{l-1}\|s^{k+1}-s^k\|_{\mathcal{H}}\le\sum_{k=j}^{\infty}\|s^{k+1}-s^k\|_{\mathcal{H}}<\varepsilon_1.$$

Thus, $\{s^k\}$ is a Cauchy sequence. Since $\mathcal{H}$ is a Hilbert space which is a complete metric space, it means that $\{s^k\}$ is convergent.

Combining (I) with (II), we conclude that $\{s^k\}$ is convergent. We denote

$$s^*:=\lim_{k\to\infty}s^k.$$

Next we show that s^* is a stationary point of Optimization (1.1). We denote the objective function of Optimization (1.1) as $\mathcal{T}:\mathcal{H}\to\mathbb{R}$. Hence, the objective function of Optimization (1.1) can be rewritten as

$$\mathcal{T}(f)=\frac{1}{N}\sum_{i=1}^{N}L(\boldsymbol{x}_i,y_i,\delta_{\boldsymbol{x}_i}(f))+\lambda\|f\|_{\mathcal{H}}^2=(F\circ\boldsymbol{\delta})(f)+\lambda\|f\|_{\mathcal{H}}^2,$$

where $\circ$ denotes the composition. Moreover, [17, Definition 1.77 and Proposition 1.107] assure that

$$\partial\mathcal{T}(s^*)=\partial(F\circ\boldsymbol{\delta})(s^*)+\nabla(\lambda\|\cdot\|_{\mathcal{H}}^2)(s^*). \tag{4.22}$$

For any $h\in\mathcal{H}$, since

$$\boldsymbol{\delta}(h)-\boldsymbol{\delta}(s^*)=((h-s^*)(\boldsymbol{x}_1),...,(h-s^*)(\boldsymbol{x}_N))^T=(\langle h-s^*,K(\boldsymbol{x}_1,\cdot)\rangle_{\mathcal{H}},...,\langle h-s^*,K(\boldsymbol{x}_N,\cdot)\rangle_{\mathcal{H}})^T,$$

from the definition of Fréchet derivative, we have that

$$\lim_{h\to s^*}\frac{\boldsymbol{\delta}(h)-\boldsymbol{\delta}(s^*)-(\langle h-s^*,K(\boldsymbol{x}_1,\cdot)\rangle_{\mathcal{H}},...,\langle h-s^*,K(\boldsymbol{x}_N,\cdot)\rangle_{\mathcal{H}})^T}{\|h-s^*\|_{\mathcal{H}}}=\lim_{h\to s^*}\frac{0}{\|h-s^*\|_{\mathcal{H}}}=0,$$

which ensures that

$$\nabla\boldsymbol{\delta}(s^*)=\left(\langle\cdot,K(\boldsymbol{x}_1,\cdot)\rangle_{\mathcal{H}},...,\langle\cdot,K(\boldsymbol{x}_N,\cdot)\rangle_{\mathcal{H}}\right)^T.$$

Since $\{s^k\}\subseteq\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),...,K(\boldsymbol{x}_N,\cdot)\}$ and $\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),...,K(\boldsymbol{x}_N,\cdot)\}$ is closed by [16, 1.4.20 Corollary], we show that $s^*\in\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),...,K(\boldsymbol{x}_N,\cdot)\}$. Hence, there exists a unique $\boldsymbol{c}^*\in\mathbb{R}^N$ such that s^* has the finite representation

$$s^*=\sum_{i=1}^{N}(\boldsymbol{c}^*)_iK(\boldsymbol{x}_i,\cdot),$$

which ensures that

$$\boldsymbol{\delta}(s^*) = (\delta_{\boldsymbol{x}_1}(s^*), ..., \delta_{\boldsymbol{x}_N}(s^*))^T = (s^*(\boldsymbol{x}_1), ..., s^*(\boldsymbol{x}_N))^T = A\boldsymbol{c}^*.$$

From the chain rule, we see that

$$\partial(F \circ \boldsymbol{\delta})(s^*) = (\nabla\boldsymbol{\delta}(s^*))^T \partial F(\boldsymbol{\delta}(s^*)) = (\langle \cdot, K(\boldsymbol{x}_1, \cdot)\rangle_{\mathcal{H}}, ..., \langle \cdot, K(\boldsymbol{x}_N, \cdot)\rangle_{\mathcal{H}})^T \, \partial F(A\boldsymbol{c}^*). \quad (4.23)$$

Since $\boldsymbol{\delta}$ is an isomorphism from $\mathrm{span}\{K(\boldsymbol{x}_1, \cdot), ..., K(\boldsymbol{x}_N, \cdot)\}$ onto $\mathbb{R}^N$, $\lim\limits_{k\to\infty} s^k = s^*$ and A is nonsingular, we observe that $\lim\limits_{k\to\infty} \boldsymbol{c}^k = \boldsymbol{c}^*$. We denote

$$\boldsymbol{\alpha}^* := A\boldsymbol{c}^*, \;\; \boldsymbol{\gamma}^* := 2\lambda\boldsymbol{c}^*.$$

From (S-3) and (S-3'), it follows that

$$\lim_{k\to\infty} (\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) = (\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*),$$

that is, $(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*) \in S$. Therefore,

$$\mathcal{L}_\rho(\boldsymbol{\alpha}^*, \boldsymbol{c}^*, \boldsymbol{\gamma}^*) = F(\boldsymbol{\alpha}^*) + G(\boldsymbol{c}^*) = \lim_{k\to\infty} \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k).$$

From the continuity of G, we observe further that

$$\begin{aligned}\lim_{k\to\infty} F(\boldsymbol{\alpha}^k) &= \lim_{k\to\infty} \Big(\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k) - G(\boldsymbol{c}^k) - (\boldsymbol{\gamma}^k)^T(\boldsymbol{\alpha}^k - A\boldsymbol{c}^k) - \frac{\rho}{2}\|\boldsymbol{\alpha}^k - A\boldsymbol{c}^k\|^2\Big) \\ &= F(\boldsymbol{\alpha}^*).\end{aligned}$$

In the view of (4.16), by [20, Proposition 8.7] and passing to the limit along $\{(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$, it follows that

$$-2\lambda\boldsymbol{c}^* = -\boldsymbol{\gamma}^* \in \partial F(\boldsymbol{\alpha}^*) = \partial F(A\boldsymbol{c}^*). \quad (4.24)$$

Hence, (4.23) and (4.24) show that

$$(\langle \cdot, K(\boldsymbol{x}_1, \cdot)\rangle_{\mathcal{H}}, ..., \langle \cdot, K(\boldsymbol{x}_N, \cdot)\rangle_{\mathcal{H}})^T \,(-2\lambda\boldsymbol{c}^*) = -\langle \cdot, 2\lambda s^*\rangle_{\mathcal{H}} \in \partial(F \circ \boldsymbol{\delta})(s^*). \quad (4.25)$$

On the other hand, from the definition of Fréchet derivative, we have that

$$\lim_{h\to s^*} \frac{\lambda\|h\|^2_{\mathcal{H}} - \lambda\|s^*\|^2_{\mathcal{H}} - \langle h - s^*, 2\lambda s^*\rangle_{\mathcal{H}}}{\|h - s^*\|_{\mathcal{H}}} = \lim_{h\to s^*} \frac{\lambda\|h - s^*\|^2_{\mathcal{H}}}{\|h - s^*\|_{\mathcal{H}}} = \lim_{h\to s^*} \lambda\|h - s^*\|_{\mathcal{H}} = 0,$$

which ensures that

$$\nabla(\lambda\|\cdot\|^2_{\mathcal{H}})(s^*) = \langle \cdot, 2\lambda s^*\rangle_{\mathcal{H}}. \quad (4.26)$$

From (4.22), (4.25) and (4.26), we conclude that

$$0 \in \partial\mathcal{T}(s^*) = \partial(F \circ \boldsymbol{\delta})(s^*) + \nabla(\lambda\|\cdot\|^2_{\mathcal{H}})(s^*),$$

that is, s^* is a stationary point of Optimization (1.1). This proof is completed. □

Next we analyze the convergent rate of $\{s^k\}$. By [5, Example 5.3] and [3, Theorem 3.1], we know that φ has the following form

$$\varphi(z) = ez^{1-\theta}, \quad \text{for } e > 0, \; \theta \in [0, 1).$$

Moreover, we have the following proposition about convergent rate.

Proposition 4.4. *If the conditions in Theorem* 4.1 *hold, then we have the following estimations:*

(i) *If* $\theta = 0$*, then there exists* $n_1 \in \mathbb{N}$ *such that for* $k > n_1$*,*

$$s^k = s^*.$$

(ii) *If* $\theta \in (0, \frac{1}{2}]$*, then there exists* $n_2 \in \mathbb{N}$*,* $C_1 > 0$ *and* $\xi \in [0, 1)$ *such that for* $k > n_2$*,*

$$\|s^k - s^*\|_{\mathcal{H}} \le C_1 \xi^k.$$

(iii) *If* $\theta \in (\frac{1}{2}, 1)$*, then there exists* $n_3 \in \mathbb{N}$ *and* $C_2 > 0$ *such that for* $k > n_3$*,*

$$\|s^k - s^*\|_{\mathcal{H}} \le C_2 k^{\frac{1-\theta}{1-2\theta}}.$$

Proof. First we consider the case that $\theta = 0$. Suppose that $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ satisfies the case (II) in Theorem 4.1. From (4.21), we have that

$$\lim_{k \to \infty} \|s^k - s^{k-1}\|_{\mathcal{H}} = 0. \tag{4.27}$$

Moreover, (4.19) and (4.27) assure that when k is sufficient largely,

$$\mathrm{dist}(\mathbf{0}, \partial \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)) \le \zeta_2 \|s^k - s^{k-1}\|_{\mathcal{H}} < \frac{1}{e}. \tag{4.28}$$

On the other hand, by the definition of φ, we have that $\varphi'(r^k) = e$. Hence, (4.14) shows that

$$e \cdot \mathrm{dist}(\mathbf{0}, \partial \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)) \ge 1. \tag{4.29}$$

Clearly, (4.28) and (4.29) are contradiction. Therefore, $\{\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)\}$ satisfies the case (I) in Theorem 4.1, that is, there exists $n_1 \in \mathbb{N}$ such that whenever $k > n_1$, $s^k = s^*$. Item (i) holds.

Next we consider the case that $\theta \in (0, 1)$. We denote

$$\Delta_k := \sum_{j=k}^{\infty} \|s^{j+1} - s^j\|_{\mathcal{H}}.$$

Since $\Delta_{k-1} - \Delta_k = \|s^k - s^{k-1}\|_{\mathcal{H}}$, for any $k > k_1$, summing up (4.20) from $j = k, k+1, ...$, the triangle inequality ensures that

$$\begin{aligned} \|s^k - s^*\|_{\mathcal{H}} &\le \sum_{j=k}^{\infty} \|s^{j+1} - s^j\|_{\mathcal{H}} \\ &= \Delta_k \le \Delta_{k-1} - \Delta_k + \frac{\zeta_2}{\zeta_1} \varphi(r^k) \\ &= \Delta_{k-1} - \Delta_k + \frac{\zeta_2}{\zeta_1} e (r^k)^{1-\theta}. \end{aligned} \tag{4.30}$$

Since $\varphi'(r^k) = e(1-\theta)(r^k)^{-\theta}$, (4.14) shows that

$$(r^k)^\theta \le e(1-\theta)\mathrm{dist}(\mathbf{0}, \partial \mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k)).$$

Hence, the inequality above and (4.19) assure that

$$\begin{aligned}(r^k)^{1-\theta} = ((r^k)^\theta)^{\frac{1-\theta}{\theta}} &\leq \left(e(1-\theta)\mathrm{dist}(\mathbf{0}, \partial\mathcal{L}_\rho(\boldsymbol{\alpha}^k, \boldsymbol{c}^k, \boldsymbol{\gamma}^k))\right)^{\frac{1-\theta}{\theta}} \\ &\leq \left(e(1-\theta)\zeta_2(\Delta_{k-1} - \Delta_k)\right)^{\frac{1-\theta}{\theta}}.\end{aligned} \tag{4.31}$$

Let $q := \dfrac{\zeta_2}{\zeta_1} e[e(1-\theta)\zeta_2]^{\frac{1-\theta}{\theta}}$. Thus $q > 0$. Combining (4.30) with (4.31), we have that

$$\begin{aligned}\Delta_k &\leq \Delta_{k-1} - \Delta_k + \frac{\zeta_2}{\zeta_1} e[e(1-\theta)\zeta_2]^{\frac{1-\theta}{\theta}} (\Delta_{k-1} - \Delta_k)^{\frac{1-\theta}{\theta}} \\ &= \Delta_{k-1} - \Delta_k + q(\Delta_{k-1} - \Delta_k)^{\frac{1-\theta}{\theta}}.\end{aligned} \tag{4.32}$$

Moreover, (4.27) shows that there exists $k_3 \in \mathbb{N}$ such that for any $k > k_3$,

$$\|s^k - s^{k-1}\|_{\mathcal{H}} = \Delta_{k-1} - \Delta_k < 1.$$

If $\theta \in (0, \frac{1}{2}]$, then $\frac{1-\theta}{\theta} \geq 1$. We denote $n_2 := \max\{k_1, k_3\}$. If $k > n_2$, then (4.32) shows that

$$\Delta_k \leq (1+q)(\Delta_{k-1} - \Delta_k). \tag{4.33}$$

This implies that $\Delta_k \leq \frac{q}{1+q}\Delta_{k-1}$. Let $C_1 := (\frac{q}{1+q})^{-1-n_2}\Delta_{n_2}$ and $\xi := \frac{q}{1+q}$. Thus $C_1 > 0$, and $\xi \in [0, 1)$. Combining (4.30) with (4.33), we show that

$$\|s^k - s^*\|_{\mathcal{H}} \leq \Delta_k \leq \left(\frac{q}{1+q}\right)^{k-1-n_2} \Delta_{n_2} = C_1\xi^k.$$

Item (ii) holds.

If $\theta \in (\frac{1}{2}, 1)$, then $\frac{1-\theta}{\theta} < 1$. Whenever $k > n_2$, it follows that

$$1 \leq (1+q)^{\frac{\theta}{1-\theta}}(\Delta_{k-1} - \Delta_k)\Delta_k^{-\frac{\theta}{1-\theta}}.$$

Let $v > 1$. First we assume that $\Delta_k^{-\frac{\theta}{1-\theta}} \leq v\Delta_{k-1}^{-\frac{\theta}{1-\theta}}$, it holds that

$$\begin{aligned}(\Delta_{k-1} - \Delta_k)\Delta_k^{-\frac{\theta}{1-\theta}} &\leq v(\Delta_{k-1} - \Delta_k)\Delta_{k-1}^{-\frac{\theta}{1-\theta}} \\ &\leq v\int_{\Delta_k}^{\Delta_{k-1}} z^{-\frac{\theta}{1-\theta}}\mathrm{d}z \\ &= \frac{1-\theta}{2\theta-1}v[\Delta_k^{\frac{1-2\theta}{1-\theta}} - \Delta_{k-1}^{\frac{1-2\theta}{1-\theta}}].\end{aligned}$$

Combining with two inequalities above, we find that

$$\frac{2\theta-1}{(1-\theta)v}(1+q)^{-\frac{\theta}{1-\theta}} \leq \Delta_k^{\frac{1-2\theta}{1-\theta}} - \Delta_{k-1}^{\frac{1-2\theta}{1-\theta}}.$$

Next we assume that $\Delta_k^{-\frac{\theta}{1-\theta}} > v\Delta_{k-1}^{-\frac{\theta}{1-\theta}}$. Since $-\frac{1-2\theta}{\theta} > 0$, we have $\Delta_k^{\frac{1-2\theta}{1-\theta}} > v^{\frac{2\theta-1}{\theta}}\Delta_{k-1}^{\frac{1-2\theta}{1-\theta}}$. This ensures that

$$(v^{\frac{2\theta-1}{\theta}} - 1)\Delta_{n_2}^{\frac{1-2\theta}{1-\theta}} \leq (v^{\frac{2\theta-1}{\theta}} - 1)\Delta_{k-1}^{\frac{1-2\theta}{1-\theta}} \leq \Delta_k^{\frac{1-2\theta}{1-\theta}} - \Delta_{k-1}^{\frac{1-2\theta}{1-\theta}}.$$

Let $\mu := \min\left\{\frac{2\theta-1}{(1-\theta)v}(1+q)^{-\frac{\theta}{1-\theta}}, (v^{\frac{2\theta-1}{\theta}}-1)\Delta_{n_2}^{\frac{1-2\theta}{1-\theta}}\right\}$. Thus $\mu > 0$, and

$$\mu \leq \Delta_k^{\frac{1-2\theta}{1-\theta}} - \Delta_{k-1}^{\frac{1-2\theta}{1-\theta}}. \tag{4.34}$$

Since $\frac{1-\theta}{1-2\theta} < 0$, summing up the above relation from $j = n_2, ..., k-1$ and rearranging terms, (4.30) and (4.34) show that there exists $C_2 > 0$ such that if $k > n_3 = n_2$,

$$\|s^k - s^*\|_{\mathcal{H}} \leq \Delta_k \leq [\Delta_{n_2}^{\frac{1-2\theta}{1-\theta}} + (k - n_2 + 1)\mu]^{\frac{1-\theta}{1-2\theta}} \leq C_2 k^{\frac{1-\theta}{1-2\theta}}.$$

Item (iii) follows immediately. This proof is completed. □

5 Numerical Examples

In this section, we test Algorithm 1 by the synthetic data and the real data. We choose some training data and testing data, RKHSs, and loss functions to test Algorithm 1. Let K_1 be Gaussian kernel, that is,

$$K_1(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma_1 \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2), \text{ for } \sigma_1 > 0$$

and K_2 be Matérn 1-norm kernel, that is,

$$K_2(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma_2 \|\boldsymbol{x} - \boldsymbol{x}'\|_1), \text{ for } \sigma_2 > 0,$$

where $\|\cdot\|_1$ denotes 1-norm in Euclidean space. Moreover, these two kernels are symmetric and strictly positive definite. In this section, we use the RKHSs induced from the kernels K_1 and K_2.

On the other hand, let L_1, L_2, L_3 and L_4 be four loss functions used in our experiments, that is,

$$L_1(\boldsymbol{x}, y, t) = \begin{cases} 1 - yt, & yt - 1 < 0 \\ 0, & yt - 1 \geq 0 \end{cases}, \quad L_2(\boldsymbol{x}, y, t) = \begin{cases} -yt + 2, & yt - 1 < -1 \\ -2yt + 2, & -1 \leq yt - 1 < 0 \\ 0, & yt - 1 \geq 0 \end{cases},$$

and

$$L_3(\boldsymbol{x}, y, t) = \begin{cases} log(2 - yt), & yt - 1 < 0 \\ 0, & yt - 1 \geq 0 \end{cases}, \quad L_4(\boldsymbol{x}, y, t) = \begin{cases} 1, & yt - 1 < -1 \\ 1 - yt, & -1 \leq yt - 1 < 0 \\ 0, & yt - 1 \geq 0 \end{cases}.$$

We see that L_1 is convex Hinge loss, L_2 is a nonconvex linear piecewise loss function, L_3 is a nonconvex piecewise logarithmic loss function and L_4 is a nonconvex ramp loss function. These four loss functions satisfy Assumption 4.1 (i). Here are the graphs of these loss functions above.

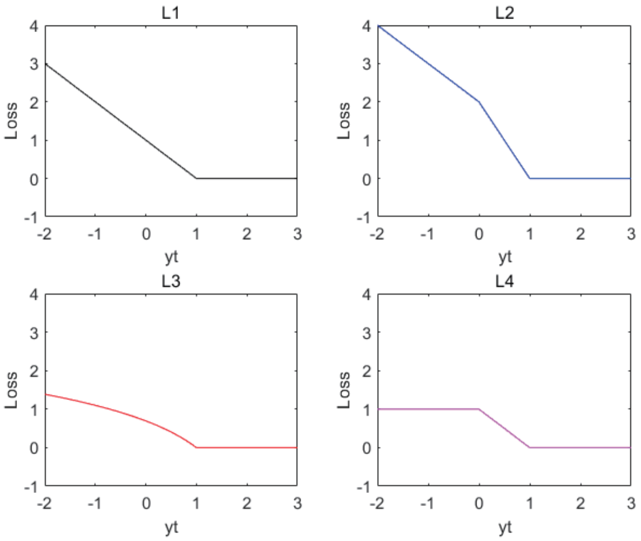


Figure 1: For the graphs of loss functions L_1, L_2, L_3 and L_4, we replace yt to t because of the symmetry of $y = +1$ and $y = -1$.

To illustrate how to solve Optimization (S-1”), we give a simple example. As for L_1, by simple algebra, the corresponding Optimization (S-1”) on $\mathbb{R}$ has the following minimizer. If $y_i = +1$, then

$$(\boldsymbol{\alpha}^{k+1})_i = \begin{cases} \nu_i + \frac{1}{\rho N}, & \nu_i < 1 - \frac{1}{\rho N}, \\ 1, & 1 - \frac{1}{\rho N} \leq \nu_i < 1, \\ \nu_i, & \nu_i \geq 1, \end{cases}$$

where $\nu_i = (A\boldsymbol{c}^k)_i - \dfrac{1}{\rho}(\boldsymbol{\gamma}^k)_i$. If $y_i = -1$, then

$$(\boldsymbol{\alpha}^{k+1})_i = \begin{cases} \nu_i, & \nu_i < -1, \\ -1, & -1 \leq \nu_i < -1 + \frac{1}{\rho N}, \\ \nu_i - \frac{1}{\rho N}, & \nu_i \geq -1 + \frac{1}{\rho N}. \end{cases}$$

As for L_2, L_3 and L_4, similarily, we can use some simple algebra in $\mathbb{R}$ to solve the corresponding Optimization (S-1”). Next we introduce our test results on synthetic data and real data.

$\boxed{5.1}$ Examples on Synthetic Data

We sample from $\Omega_1 = [-3, 10] \times [-3, 10]$ labeled by $+1$ and $\Omega_2 = [-10, 3] \times [-10, 3]$ labeled by -1 randomly to obtain different training sets and testing sets. The data labeled by $+1$ are equal to the data labeled by -1 in each training set or testing set. Here is an example of sampling. In the following figures, two subdatasets are colored in blue and red.

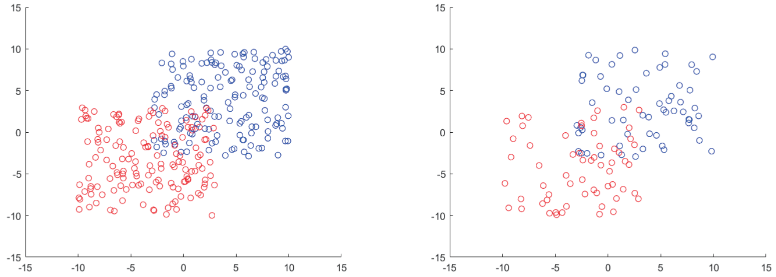


Figure 2: Example of Sampling ($N = 300$).

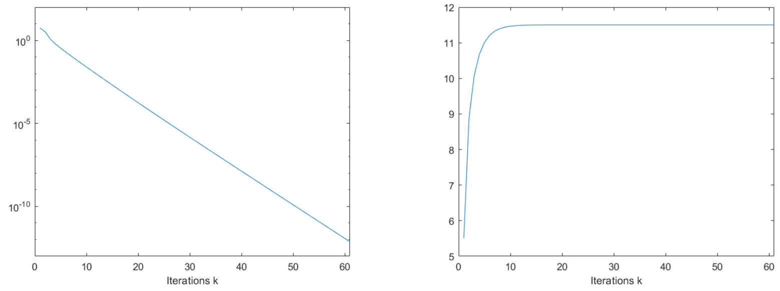
First we show the convergence of Algorithm 1 for nonconvex loss function L_3. Some parameters and results of the numerical experiment be represented as follows.

- Gaussian kernel K_1, where $\sigma_1 = 1$.
- The nonconvex loss function L_3.
- $N = 300$, $\lambda = 0.1$, $\rho = 0.05$ and $\varepsilon_0 = 10^{-12}$.
- Choose 20 initial values randomly in $[-10, 10]^N$.

From (4.21), we can show the convergence of $\{s^k\}$ by $\sum\limits_{k=1}^{\infty} \|s^{k+1} - s^k\|_{\mathcal{H}}$. By definition of $\{s^k\}$ and (3.4), we have that

$$\|s^{k+1} - s^k\|_{\mathcal{H}} = \sqrt{\sum_{i=1}^{N} (\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)_i (s^{k+1} - s^k)(\boldsymbol{x}_i)} = \sqrt{(\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)^T A (\boldsymbol{c}^{k+1} - \boldsymbol{c}^k)}.$$

In the following two pictures, we show the convergence of Algorithm 1 by $\sum\limits_{k=1}^{\infty} \|s^{k+1} - s^k\|_{\mathcal{H}}$.



(a) $\|s^{k+1} - s^k\|_{\mathcal{H}}$ (b) $\sum_{k=1}^{\infty} \|s^{k+1} - s^k\|_{\mathcal{H}}$

Figure 3: Convergence of Algorithm 1 with Nonconvex Loss Function L_3.

Figure 3 shows that for the training data and parameters, Algorithm 1 converges in 61 iterations. This shows the effectiveness of Algorithm 1. Next, we use different sizes of training sets and testing sets to test Algorithm 1. Some parameters and results of the numerical experiment are represented as follows.

- Gaussian kernel K_1, where $\sigma_1 = 1$.
- The nonconvex loss function L_2.
- $N = 300$, $\lambda = 0.1$, $\rho = 1$ and $\varepsilon_0 = 10^{-12}$.
- Choose 20 initial values randomly in $[-10, 10]^N$.

Table 1: Comparison of Different Sizes of Data.

Training Data	Testing Data	Time(s)	Training Accuracy	Testing Accuracy
100	40	3.546	98%	85%
200	80	7.302	94%	87.5%
300	120	10.524	93.7%	90%
400	160	14.833	93%	87.5%
500	200	18.343	93.2%	90%
600	240	21.813	92.7%	87.1%
700	280	27.469	92.4%	89.3%
800	320	33.776	92.1%	90.9%
900	360	44.653	90.3%	90%
1000	400	56.510	92.3%	89%

Table 1 shows that solving Optimization (1.1) by Algorithm 1 is feasible in terms of running time and accuracy. Next we will show that choosing different kinds of loss functions and kernels have different accuracy.

Now we sample from the area Ω_1 and Ω_2 randomly to obtain a training set with 300 points and a testing set with 120 points. Some parameters of these experiments are represented as follows.

- The kernels K_1 and K_2, where $\sigma_1 = 2$ and $\sigma_2 = 1$.
- The loss functions L_1, L_2, L_3 and L_4.
- $\lambda = 0.5$, $\rho = 5$ and $\varepsilon_0 = 10^{-12}$.
- Choose 20 initial values randomly in $[-10, 10]^N$.

In each experiment, we will choose a loss function and a kernel. The results of these experiments represent as follows.

Table 2: Numerical Results of Using Different Loss Functions and Kernels.

Loss Function	Kernel	Training Accuracy	Testing Accuracy
L_1	K_1	95.7%	87.5%
L_2	K_1	95%	88.3%
L_3	K_1	95.7%	88.3%
L_4	K_1	95%	87.5%
L_1	K_2	92%	90.8%
L_2	K_2	93.7%	90%
L_3	K_2	94.7%	90.8%
L_4	K_2	93%	90%

From Table 2, it is easy to see that nonconvex loss function L_3 performs better than L_1, L_2 and L_4 in these experiments. It shows that the SVM in RKHS with nonconvex loss function is better than the SVM in RKHS with convex loss function in some cases. Next, we introduce the numerical experiment result on a real-world benchmark dataset.

5.2 Examples on UCI Machine Learning Repository

The vinho verde data in UCI machine learning repository has two kinds of wine samples. We will identify them based on physicochemical tests. There are 11 input variables about them, which are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. We have 2000 wine samples in training set, and a half of them are labeled by +1 and the others are labeled by -1. Moreover, we have 718 wine samples in testing set, and a half of them are labeled by +1 and the others are labeled by -1. Next, we introduce some parameters of these experiments as follows.

- The kernels K_1 and K_2, where $\sigma_1 = 5$ and $\sigma_2 = 5$.
- The loss functions L_1, L_2, L_3 and L_4.
- $N = 2000$, $\lambda = 0.5$, $\rho = 1$ and $\varepsilon_0 = 10^{-12}$.
- Choose 20 initial values randomly in $[-10, 10]^N$.

In each experiment, we will choose a loss function and a kernel and we have the following results.

Table 3: Numerical Results on Vinho Verde Data.

Loss Function	Kernel	Training Accuracy	Testing Accuracy
L_1	K_1	99,9%	90.0%
L_2	K_1	100%	90.8%
L_3	K_1	99.9%	90.3%
L_4	K_1	99.9%	90.3%
L_1	K_2	100%	88.3%
L_2	K_2	100%	91.8%
L_3	K_2	100%	87.6%
L_4	K_2	100%	91.2%

From Table 3, we check that L_2 performs better than L_1, L_3 and L_4. It shows that in some cases nonconvex loss function is more suitable than convex loss function, which is our motivation for this paper.

In Section 5, we demonstrate the effectiveness of solving Optimization (1.1) by Algorithm 1. In addition, we give some examples to show that in some cases, the SVM in RKHS with nonconvex loss function is better than the SVM in RKHS with convex loss function. Therefore, we should reconsider not only convex loss function but also nonconvex loss function.

References

[1] A. Beck, *First-Order Methods in Optimization*, SIAM, Philadelphia, 2017.

[2] D. Bertsekas, *Convex Optimization Theory*, Athena Scientific, Nashua, 2009.

[3] J. Bolte, A. Daniilidis and A. Lewis, The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM J. Optimiz.* 17 (2007) 1205–1223.

[4] J. Bolte, A. Daniilidis, A. Lewis and M. Shiota, Clarke subgradients of stratifiable functions, *SIAM J. Optimiz.* 18 (2007) 556–572.

[5] J. Bolte, S. Sabach and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Math. Program.* 146 (2014) 459–494.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends. Mach. Le.* 3 (2011) 1–122.

[7] J. Brooks, Support vector machines with ramp loss and the hard margin loss, *Oper. Res.* 59 (2011) 467–479.

[8] C. Cortes and V. Vapnik, Support vector network, *Mach. Learn.* 20 (1995) 273–297.

[9] Y. Feng, Y. Yang, S. Huang, S. Mehrkanoon and J. Suykens, Robust support vector machines for classification with nonconvex and smooth losses, *Neural Comput.* 28 (2016) 1217–1247.

[10] F. Facchinei and J. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems, vol.I*, Springer, Berlin, 2003.

[11] K. Guo, D. Han and T. Wu, Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints, *Int. J. Comput. Math.* 94 (2016) 1–18.

[12] L. Huang, C. Liu, L. Tan and Q. Ye, Generalized representer theorems in Banach spaces, *Anal. Appl.* 19 (2021) 125–146.

[13] D. Kincaid and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing, 3rd Edition*, American Mathematical Society, Providence, 2002.

[14] G. Li and T. Pong, Global convergence of splitting methods for nonconvex composite optimization, *SIAM J. Optimiz.* 25 (2015) 2434–2460.

[15] D. Liu, Y. Shi, Y. Tian and X. Huang, Ramp loss least squares support vector machine, *J. Comput. Sci-Neth.* 14 (2016) 61–68.

[16] R. Megginson, *An Introduction to Banach Space Theory*, Springer-Verlag, New York, 1998.

[17] B. Mordukhovich, *Variational Analysis and Generalized Differentiation. I: Basic Theory*, Grundlehren Series (Fundamental Principles of Mathematical Sciences) Springer, Berlin, 2006.

[18] S. Park and Y. Liu, Robust penalized logistic regression with truncated loss functions, *Can. J. Stat.* 39 (2011) 300–323.

[19] F. Pérez-Cruz, A. Navia-Vázquez, A. Figueiras-Vidal and A. Artés-Rodríguez, Empirical risk minimization for support vector classifiers, *IEEE Trans. Neural Netw.* 14 (2003) 296–303.

[20] R. Rockafellar and R. Wets, *Variational Analysis*, Springer Berlin Heidelberg, Berlin, 1998.

[21] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, Inc, New York, 1976.

[22] X. Shen, L. Niu, Z. Qi. and Y. Tian, Support vector machine classifier with truncated pinball loss, *Pattern Recogn.* 68 (2017) 199–210.

[23] M. Shiota, *Geometry of Subanalytic and Semialgebraic Sets*, Birkhäuser, Boston, 1998.

[24] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer-Verlag, New York, 2008.

[25] H. Wang, Y. Shao, S. Zhou, C. Zhang and N. Xiu, Support vector machine classifier via L0/1 soft-margin loss, *IEEE T. Pattern Anal.* 44 (2022) 7253–7265.

[26] H. Wendland, *Scattered Data Approximation*, Cambridge University Press, Cambridge, 2005.

[27] Y. Xu and Q. Ye, Generalized Mercer kernels and reproducing kernel Banach spaces, *Mem. Am. Math. Soc.* 258 (2019) 1–122.

[28] Y. Xu and W. Yin, A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion, *SIAM J. Imaging Sci.* 6 (2013) 1758–1789.

[29] L. Yang and H. Dong, Support vector machine with truncated pinball loss and its application in pattern recognition, *Chemometr. Intell. Lab.* 177 (2018) 89–99.

Manuscript received 12 August 2022
revised 3 January 2023
accepted for publication 1 March 2023

Mingyu Mo
School of Mathematical Sciences
South China Normal University
Guangzhou, 510631, Guangdong, P.R. China
E-mail address: mmymaths@qq.com

Qi Ye
School of Mathematical Sciences
South China Normal University
Guangzhou, 510631, Guangdong, P.R. China
E-mail address: yeqi@m.scnu.edu.cn