Chin. Ann. Math. Ser. B
45(6), 2024, 823–854
DOI: 10.1007/s11401-024-0041-5

Chinese Annals of Mathematics, Series B

Splitting Method for Support Vector Machine in Reproducing Kernel Banach Space with a Lower Semi-continuous Loss Function*

Mingyu MO[1] Yimin WEI[2] Qi YE[3]

Abstract In this paper, the authors employ the splitting method to address support vector machine within a reproducing kernel Banach space framework, where a lower semi-continuous loss function is utilized. They translate support vector machine in reproducing kernel Banach space with such a loss function to a finite-dimensional tensor optimization problem and propose a splitting method based on the alternating direction method of multipliers. Leveraging Kurdyka-Lojasiewicz property of the augmented Lagrangian function, the authors demonstrate that the sequence derived from this splitting method is globally convergent to a stationary point if the loss function is lower semi-continuous and subanalytic. Through several numerical examples, they illustrate the effectiveness of the proposed splitting algorithm.

Keywords Support vector machine, Lower semi-continuous loss function, Reproducing kernel Banach space, Tensor optimization problem, Splitting method
2000 MR Subject Classification 68Q32, 68T05, 46E22, 68P01

1 Introduction

In this paper, we employ the splitting method to address support vector machine (SVM for short) in reproducing kernel Banach space (RKBS for short) with a lower semi-continuous loss function. SVM is a successful model in machine learning. First, we introduce the background of machine learning. We denote the sample space as X in the d-dimensional Euclidean space $\mathbb{R}^d$ and the label space as Y in the set of real numbers as $\mathbb{R}$, respectively. We have a truth $\mathcal{R} : X \to Y$ and the following training data

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

Manuscript received March 6, 2024. Revised August 27, 2024.
[1]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China; School of Mathematical Science, South China Normal University, Guangzhou 510631, China. E-mail: mmymaths@pku.edu.cn
[2]School of Mathematical Science, Key Laboratory of Mathematics for Nonlinear Sciences, Fudan University, Shanghai 200433, China. E-mail: ymwei@fudan.edu.cn
[3]Corresponding author. School of Mathematical Science, South China Normal University, Guangzhou 510631, China. E-mail: yeqi@m.scnu.edu.cn
*This work was supported by the National Natural Science Foundation of China (Nos. 12026602, 12071157, 12271108), the Natural Science Foundation of Guangdong Provience (No. 2024A1515012288), the Science and Technology Commission of Shanghai Municipality (No. 23JC1400501) and the Ministry of Science and Technology of China (No. G2023132005L).

For any $\boldsymbol{x} \in X\backslash\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$, $\mathcal{R}(\boldsymbol{x})$ is unknown. Based on the training data D, we learn a mapping $\mathcal{R}_D : X \to Y$ such that $\mathcal{R}_D(\boldsymbol{x})$ is a good approximation of $\mathcal{R}(\boldsymbol{x})$ to an arbitrary $\boldsymbol{x} \in X$.

One heuristic method for learning $\mathcal{R}_D$ is to choose the optimal one in the set of all mappings from X to Y according to a given criterion. Since X is uncountable, it is difficult to consider all mappings from X to Y. We can only consider a part of mappings from X to Y. Thus, different parts of mappings from X to Y and different criteria can obtain different kinds of $\mathcal{R}_D$, but we do not know which $\mathcal{R}_D$ is better without testing data. We hope that as many mappings from X to Y as possible can be considered and the corresponding $\mathcal{R}_D$ can be obtained easily.

In this paper, we consider some infinite-dimensional $\frac{m}{m-1}$-norm RKBSs $\mathcal{B}^{\frac{m}{m-1}}$ that consist of some mappings from X to $\mathbb{R}$, where m is an even integer. We will find a mapping in $\mathcal{B}^{\frac{m}{m-1}}$ that achieves the smallest regularized possible empirical risk (see [31, Section 5.5]), that is,

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

where $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a given lower semi-continuous loss function and $\lambda > 0$ is a given regularization paramater. 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. A loss function L is called lower semi-continuous if $t \mapsto L(\boldsymbol{x}, y, t)$ is lower semi-continuous on $\mathbb{R}$ for all $\boldsymbol{x} \in X$ and $y \in Y$. Similarly, we can define other properties of loss function, such as continuity, smoothness, convexity, etc. Finally, we use f_D to construct $\mathcal{R}_D$ according to the task requirement. For example, if $Y = \mathbb{R}$, then we construct $\mathcal{R}_D = f_D$ for regression. If $Y = \{+1, -1\}$, then we can construct

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

for binary classification (see [28, Sections 8–9]). In this paper, we call $\mathcal{R}_D$ constructed by optimization problem (1.1) the SVM in $\mathcal{B}^{\frac{m}{m-1}}$. Clearly, the key step in constructing the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ is to solve optimization problem (1.1).

When $m = 2$, $\frac{m}{m-1} = 2$ and $\mathcal{B}^2$ is a reproducing kernel Hilbert space (RKHS for short). The SVM in $\mathcal{B}^2$ is already achieved with convex loss function (see [28]). By definition, if L is a convex loss function, then L is a lower semi-continuous loss function. But the converse is not true. Recently, some nonconvex and lower semi-continuous loss functions have been used in SVM (see [7, 12, 17, 26, 30]). In [19], we show that when $m = 2$ and L is a lower semi-continuous loss function, optimization problem (1.1) can be equivalently transferred to a lower semi-continuous finite-dimensional optimization problem with linear constraint and a positive semi-definite matrix. Next, we discuss the splitting method based on the alternating direction method of multipliers (ADMM for short) for optimization problem (1.1). By this splitting method, we obtain two subproblems that are computable easily. The first subproblem can be equivalently transferred to some lower semi-continuous 1-dimensional optimization problems and the second one can be equivalently transferred to a well-posed linear equation. Also, we discuss the convergence of this splitting method. The convergence of ADMM is already guaranteed well for convex optimization problems (see [6]) and some special nonconvex optimization

problems by the Kurdyka-Lojasiewicz (KL for short) property of the augmented Lagrangian function (see [13, 15]). In [19], we discuss the global convergence of the splitting method based on ADMM to a stationary point by the KL property of the augmented Lagrangian function.

Since RKHS is a special case of RKBS, after the success of the SVM in RKHS, people naturally would like to know whether the SVM can be achieved in a general RKBS. When $m \geq 4$, $\mathcal{B}^{\frac{m}{m-1}}$ is just an RKBS, not an RKHS. If $m \geq 4$ and L is a convex loss function, [31, Section 5.5] shows that optimization problem (1.1) has a unique minimizer. Moreover, [31, Section 5.2] develops the fixed-point algorithm for optimization problem (1.1) with a convex and smooth loss function. Furthermore, a preliminary experiment in [31, Chapter 6] shows that if L is a convex and smooth loss function, the SVM in $\mathcal{B}^{\frac{4}{3}}$ can perform better than the SVM in $\mathcal{B}^2$. Recently, [14] shows that when $m \geq 4$ and L is a lower semi-continuous loss function, optimization problem (1.1) with another regularization term $\lambda\|f\|^2_{\mathcal{B}^{\frac{m}{m-1}}}$ has a minimizer. However, when $m \geq 4$ and L is a nonconvex and lower semi-continuous loss function, the algorithm for optimization problem (1.1) is still lack of study, no matter what kind of regularization term it contains.

Based on the situation above, when $m \geq 4$, there are still a couple of issues that need discussion:

- When L is a nonconvex and lower semi-continuous loss function, does optimization problem (1.1) have a minimizer? If the minimizer exists, how to solve it?
- If we find an algorithm for optimization problem (1.1) with a nonconvex and lower semi-continuous loss function, is this algorithm still effective when L is a convex loss function? Does the loss function need to be smooth?

In this paper, we answer these two questions. For simplicity, in the rest of this paper, without specification, we assume that m is an even integer, $m \geq 4$, and L is a lower semi-continuous loss function.

First, we show that optimization problem (1.1) has a minimizer (see Lemma 2.1) and it can be equivalently transferred to a lower semi-continuous finite-dimensional optimization problem with nonlinear constraint and a positive semi-definite tensor (see optimization problem (2.11)). Next, we discuss the splitting method based on ADMM for optimization problem (1.1) (see Algorithm 1). By Algorithm 1, we obtain two subproblems that are computable easily. The first subproblem can be equivalently transferred to some lower semi-continuous 1-dimensional optimization problems and the second one can be equivalently transferred to a well-posed tensor equation. Specially, two subproblems are nonconvex. To ensure the convergence of Algorithm 1, we need to add more conditions on loss functions, training data, and RKBSs (see Assumption 4.1) to establish Lemma 4.1 to Lemma 4.6. In Assumption 4.1, the loss function L can be nonconvex and nomsmooth. Thus, we reexchange the convergence theorems in [13, 15, 19] and verify the global convergence to a stationary point of Algorithm 1 for optimization problem (1.1) by KL property of the augmented Lagrangian function (see Theorem 4.1). Finally, we give some numerical examples for synthetic data and real data in Section 5 to show that Algorithm 1 for the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ with a lower semi-continuous loss function is feasible.

This paper is organized as follows. Section 2 introduces some preliminary materials of the SVM in $\mathcal{B}^{\frac{m}{m-1}}$. Next, we use Algorithm 1 for optimization problem (1.1) in Section 3. Moreover, we discuss the global convergence of Algorithm 1 in Section 4. Finally, we give some numerical examples for synthetic data and real data in Section 5.

2 Support Vector Machine in $\frac{m}{m-1}$-norm Reproducing Kernel Banach Space

In this section, we review some concepts of the SVM in $\mathcal{B}^{\frac{m}{m-1}}$, where m is an even integer and $m \geq 4$. We denote the set of positive integers as $\mathbb{N}$. In this paper, every vector is supposed to be a column vector.

Before introducing $\mathcal{B}^{\frac{m}{m-1}}$, we need some basic concepts of the spaces $\ell^{\frac{m}{m-1}}$ (see [18, Appendix C]) that consist of real sequences. We define a normed space

$$\ell^{\frac{m}{m-1}} := \Big\{\{a_n\} : \ a_n \in \mathbb{R}, \ \sum_{n\in\mathbb{N}} |a_n|^{\frac{m}{m-1}} < \infty\Big\},$$

equipped with the norm $\|\{a_n\}\|_{\frac{m}{m-1}} := \big(\sum_{n\in\mathbb{N}} |a_n|^{\frac{m}{m-1}}\big)^{\frac{m-1}{m}}$. Since $1 < \frac{m}{m-1} \leq \frac{4}{3}$, [18, Theorem C.10] shows that $\ell^{\frac{m}{m-1}}$ is a Banach space. Let $(\ell^{\frac{m}{m-1}})'$ be the dual space of $\ell^{\frac{m}{m-1}}$. Since $\frac{1}{\frac{m}{m-1}} + \frac{1}{m} = 1$, [18, Theorem C.12] ensures that $(\ell^{\frac{m}{m-1}})'$ is isometrically isomorphic to ℓ^m, where

$$\ell^m := \Big\{\{b_n\} : \ b_n \in \mathbb{R}, \ \sum_{n\in\mathbb{N}} |b_n|^m < \infty\Big\},$$

and ℓ^m is equipped with the norm $\|\{b_n\}\|_m := \big(\sum_{n\in\mathbb{N}} |b_n|^m\big)^{\frac{1}{m}}$. The isometric isomorphism from ℓ^m onto $(\ell^{\frac{m}{m-1}})'$ can be represented in the form $\{b_n\} \mapsto \langle\cdot, \{b_n\}\rangle_{\frac{m}{m-1}}$, where $\langle\cdot,\cdot\rangle_{\frac{m}{m-1}}$ is the dual bilinear product defined on $\ell^{\frac{m}{m-1}}$ and ℓ^m, that is,

$$\langle\{a_n\}, \{b_n\}\rangle_{\frac{m}{m-1}} = \sum_{n\in\mathbb{N}} a_n b_n.$$

Also, the Hölder's inequality shows that

$$\langle\{a_n\}, \{b_n\}\rangle_{\frac{m}{m-1}} \leq \sum_{n\in\mathbb{N}} |a_n b_n| \leq \|\{a_n\}\|_{\frac{m}{m-1}} \|\{b_n\}\|_m.$$

Moreover, we need some basic properties of $\ell^{\frac{m}{m-1}}$. For example, [18, Theorem C.14] shows that $\ell^{\frac{m}{m-1}}$ is reflexive, [18, Proposition 5.2.6, Corollary 5.2.12] show that $\ell^{\frac{m}{m-1}}$ is strictly convex and [18, Proposition 5.5.7, Corollary 5.5.17] show that $\ell^{\frac{m}{m-1}}$ is smooth.

2.1 $\frac{m}{m-1}$-norm reproducing kernel Banach space

We now introduce some basic concepts of $\mathcal{B}^{\frac{m}{m-1}}$ (see [31, Section 3.2]), where m is an even integer. Let $\{\phi_n\}$ be a series of continuous functions from X to $\mathbb{R}$ such that $\sum_{n\in\mathbb{N}} |\phi_n(\boldsymbol{x})| < \infty$ for all $\boldsymbol{x} \in X$. We define a normed space

$$\mathcal{B}^{\frac{m}{m-1}} := \Big\{f := \sum_{n\in\mathbb{N}} a_n\phi_n : \ a_n \in \mathbb{R}, \ \sum_{n\in\mathbb{N}} |a_n|^{\frac{m}{m-1}} < \infty\Big\}$$

equipped with the norm $\|f\|_{\mathcal{B}^{\frac{m}{m-1}}} := \|\{a_n\}\|_{\frac{m}{m-1}}$. This construction of $\mathcal{B}^{\frac{m}{m-1}}$ ensures that $\mathcal{B}^{\frac{m}{m-1}}$ is isometrically isomorphic to $\ell^{\frac{m}{m-1}}$. Since $\ell^{\frac{m}{m-1}}$ is a Banach space, $\mathcal{B}^{\frac{m}{m-1}}$ is also a Banach space.

Let $(\mathcal{B}^{\frac{m}{m-1}})'$ be the dual space of $\mathcal{B}^{\frac{m}{m-1}}$. Since $(\ell^{\frac{m}{m-1}})'$ is isometrically isomorphic to ℓ^m, it is clear that $(\mathcal{B}^{\frac{m}{m-1}})'$ is isometrically isomorphic to $\mathcal{B}^m$, where

$$\mathcal{B}^m := \Big\{ g := \sum_{n\in\mathbb{N}} b_n\phi_n : \ b_n \in \mathbb{R}, \ \sum_{n\in\mathbb{N}} |b_n|^m < \infty \Big\}$$

and $\mathcal{B}^m$ is equipped with the norm $\|g\|_{\mathcal{B}^m} := \|\{b_n\}\|_m$. The isometric isomorphism from $\mathcal{B}^m$ onto $(\mathcal{B}^{\frac{m}{m-1}})'$ can be represented in the form $g \mapsto \langle \cdot, g\rangle_{\mathcal{B}^{\frac{m}{m-1}}}$, where $\langle\cdot,\cdot\rangle_{\mathcal{B}^{\frac{m}{m-1}}}$ is the dual bilinear product defined on $\mathcal{B}^{\frac{m}{m-1}}$ and $\mathcal{B}^m$, that is,

$$\langle f, g\rangle_{\mathcal{B}^{\frac{m}{m-1}}} := \sum_{n\in\mathbb{N}} a_n b_n.$$

Also, for any $f \in \mathcal{B}^{\frac{m}{m-1}}$ and $g \in \mathcal{B}^m$, we have that

$$\langle f, g\rangle_{\mathcal{B}^{\frac{m}{m-1}}} \le \|f\|_{\mathcal{B}^{\frac{m}{m-1}}} \|g\|_{\mathcal{B}^m}.$$

Remark 2.1 Following the isometric isomorphism from $\mathcal{B}^m$ onto $(\mathcal{B}^{\frac{m}{m-1}})'$, in this paper, we use the elements in $\mathcal{B}^m$ to represent the corresponding elements in $(\mathcal{B}^{\frac{m}{m-1}})'$.

Since $\ell^{\frac{m}{m-1}}$ is reflexive, strictly convex and smooth, $\mathcal{B}^{\frac{m}{m-1}}$ is also reflexive, strictly convex and smooth. For any $f \neq 0$, [31, Theorem 5.2] shows that

$$d_G(\|\cdot\|_{\mathcal{B}^{\frac{m}{m-1}}})(f) = \frac{(\mathcal{J}_m)^{-1}(f)}{\|f\|^{\frac{1}{m-1}}_{\mathcal{B}^{\frac{m}{m-1}}}}, \tag{2.1}$$

where d_G denotes Gâteaux derivative and $\mathcal{J}_m : \mathcal{B}^m \to \mathcal{B}^{\frac{m}{m-1}}$ denotes duality mapping from $\mathcal{B}^m$ to $\mathcal{B}^{\frac{m}{m-1}}$ (see [29, Definition 2]), respectively. The duality mapping $\mathcal{J}_m$ can be represented in the form

$$\mathcal{J}_m(g) := \sum_{n\in\mathbb{N}} (b_n)^{m-1}\phi_n.$$

Since $m \ge 4$, $\mathcal{J}_m$ is a nonlinear but continuous mapping. Also, [8, Chapter II, Proposition 4.8] shows that $\mathcal{J}_m$ is a homeomorphism. Hence, $\mathcal{J}_m$ has the inverse mapping

$$(\mathcal{J}_m)^{-1}(f) = \sum_{n\in\mathbb{N}} (a_n)^{\frac{1}{m-1}}\phi_n.$$

Next, we check the reproducing property of $\mathcal{B}^{\frac{m}{m-1}}$ by the well-defined kernel $K : X\times X \to \mathbb{R}$,

$$K(\boldsymbol{x}, \boldsymbol{x}') := \sum_{n\in\mathbb{N}} \phi_n(\boldsymbol{x})\phi_n(\boldsymbol{x}'), \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in X.$$

Since $\sum\limits_{n\in\mathbb{N}} |\phi_n(\boldsymbol{x})| < \infty$ for all $\boldsymbol{x} \in X$, it follows that $\sum\limits_{n\in\mathbb{N}} |\phi_n(\boldsymbol{x})|^m < \infty$. Hence,

$$K(\boldsymbol{x}, \cdot) = \sum_{n\in\mathbb{N}} \phi_n(\boldsymbol{x})\phi_n \in \mathcal{B}^m, \quad \langle f, K(\boldsymbol{x}, \cdot)\rangle_{\mathcal{B}^{\frac{m}{m-1}}} = \sum_{n\in\mathbb{N}} a_n\phi_n(\boldsymbol{x}) = f(\boldsymbol{x})$$

for all $\boldsymbol{x} \in X$ and $f \in \mathcal{B}^{\frac{m}{m-1}}$. Thus, $\mathcal{B}^{\frac{m}{m-1}}$ is a right-sided RKBS (see [31, Definition 2.1]). For any even integer $m \ge 4$, $\mathcal{B}^{\frac{m}{m-1}}$ has the same reproducing kernel K.

2.2 Support vector machine in $\frac{m}{m-1}$-norm reproducing kernel Banach space

In this subsection, we introduce some basic properties of optimization problem (1.1). For simplicity, we denote

$$\boldsymbol{\delta}(f) := (f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_N))^{\mathrm{T}}, \quad F(\boldsymbol{\alpha}) := \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{x}_i, y_i, \alpha_i).$$

Also, we denote the objective function of optimization problem (1.1) as $\mathcal{T}_{\frac{m}{m-1}} : \mathcal{B}^{\frac{m}{m-1}} \to \mathbb{R}$. Thus,

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

where $\circ$ denotes composition. Since L is nonnegative, it is clear that for any $f \in \mathcal{B}^{\frac{m}{m-1}}$,

$$\mathcal{T}_{\frac{m}{m-1}}(f) = (F \circ \boldsymbol{\delta})(f) + \lambda \|f\|^{\frac{m}{m-1}}_{\mathcal{B}^{\frac{m}{m-1}}} \geq \lambda \|f\|^{\frac{m}{m-1}}_{\mathcal{B}^{\frac{m}{m-1}}} \geq 0.$$

Therefore, $\mathcal{T}_{\frac{m}{m-1}}(f)$ tends to ∞ as $\|f\|_{\mathcal{B}^{\frac{m}{m-1}}}$ tends to ∞. Since $\mathcal{B}^{\frac{m}{m-1}}$ is reflexive, [10, Example 1.14] shows that $\mathcal{T}_{\frac{m}{m-1}}$ is sequentially coercive in the weak topology of $\mathcal{B}^{\frac{m}{m-1}}$. Moreover, since L is a lower semi-continuous loss function, F is lower semi-continuous on $\mathbb{R}^N$. Also, the reproducing property of $\mathcal{B}^{\frac{m}{m-1}}$ assures that

$$\|\boldsymbol{\delta}(f)\| = \|(f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_N))^{\mathrm{T}}\| \leq \|f\|_{\mathcal{B}^{\frac{m}{m-1}}} \|(\|K(\boldsymbol{x}_1, \cdot)\|_{\mathcal{B}^m}, \cdots, \|K(\boldsymbol{x}_N, \cdot)\|_{\mathcal{B}^m})^{\mathrm{T}}\|,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^N$. Thus, $\boldsymbol{\delta}$ is a continuous linear mapping on $\mathcal{B}^{\frac{m}{m-1}}$, which ensures that $\boldsymbol{\delta}$ is weakly continuous on $\mathcal{B}^{\frac{m}{m-1}}$ by [18, Proposition 2.5.3]. In conclusion, $F \circ \boldsymbol{\delta}$ is lower semi-continuous and weakly lower semi-continuous on $\mathcal{B}^{\frac{m}{m-1}}$. On the other hand, since $f \mapsto \|f\|_{\mathcal{B}^{\frac{m}{m-1}}}$ is continuous and weakly lower semi-continuous on $\mathcal{B}^{\frac{m}{m-1}}$ by [18, Theorem 2.5.21] and $t \mapsto \lambda t^{\frac{m}{m-1}}$ is continuous on $[0, \infty)$, $f \mapsto \lambda \|f\|^{\frac{m}{m-1}}_{\mathcal{B}^{\frac{m}{m-1}}}$ is continuous and weakly lower semi-continuous on $\mathcal{B}^{\frac{m}{m-1}}$. In conclusion, $\mathcal{T}_{\frac{m}{m-1}}$ is lower semi-continuous and weakly lower semi-continuous on $\mathcal{B}^{\frac{m}{m-1}}$.

In the rest of this paper, we usually discuss the optimal conditions of lower semi-continuous functions. We need the concept of limiting subdifferential. For a lower semi-continuous function $\mathcal{T}_{\frac{m}{m-1}}$, we use $f^k \overset{\mathcal{T}_{\frac{m}{m-1}}}{\longrightarrow} f$ to denote $f^k \to f$ and $\mathcal{T}_{\frac{m}{m-1}}(f^k) \to \mathcal{T}_{\frac{m}{m-1}}(f)$. The regular subdifferential of $\mathcal{T}_{\frac{m}{m-1}}$ at $f \in \mathcal{B}^{\frac{m}{m-1}}$ is given by

$$\widehat{\partial}\mathcal{T}_{\frac{m}{m-1}}(f) := \Big\{ g \in \mathcal{B}^m : \liminf_{h \to f,\ h \neq f} \frac{\mathcal{T}_{\frac{m}{m-1}}(h) - \mathcal{T}_{\frac{m}{m-1}}(f) - \langle h - f, g \rangle_{\mathcal{B}}}{\|h - f\|_{\mathcal{B}^{\frac{m}{m-1}}}} \geq 0 \Big\}.$$

Also, the limiting subdifferential of $\mathcal{T}_{\frac{m}{m-1}}$ at f is given by

$$\partial\mathcal{T}_{\frac{m}{m-1}}(f) := \{ g \in \mathcal{B}^m : \exists f^k \overset{\mathcal{T}_{\frac{m}{m-1}}}{\longrightarrow} f,\ g^k \to g \text{ with } g^k \in \widehat{\partial}\mathcal{T}_{\frac{m}{m-1}}(f^k) \text{ for each } k \}.$$

Similarly, we can define the limiting subdifferential of $F \circ \boldsymbol{\delta}$ at $f \in \mathcal{B}^{\frac{m}{m-1}}$ and F at $\boldsymbol{\alpha} \in \mathbb{R}^N$.

Now we discuss the limiting subdifferential of $\mathcal{T}_{\frac{m}{m-1}}$. For any $f \neq 0$, the chain rule and (2.1) assure that for any $f \neq 0$,

$$d_G(\lambda \|\cdot\|^{\frac{m}{m-1}}_{\mathcal{B}^{\frac{m}{m-1}}})(f) = \frac{m\lambda}{m-1}(\mathcal{J}_m)^{-1}(f).$$

Thus, [20, Proposition 1.107] shows that

$$\partial \mathcal{T}_{\frac{m}{m-1}}(f) = \partial(F \circ \boldsymbol{\delta})(f) + d_G(\lambda \|\cdot\|^{\frac{m}{m-1}}_{\mathcal{B}^{\frac{m}{m-1}}})(f) = \partial(F \circ \boldsymbol{\delta})(f) + \frac{m\lambda}{m-1}(\mathcal{J}_m)^{-1}(f). \tag{2.2}$$

Moreover, for any $h \in \mathcal{B}^{\frac{m}{m-1}}$ and $t \in \mathbb{R}$, since $\boldsymbol{\delta}$ is a linear mapping, we have that

$$\begin{aligned} d_G\boldsymbol{\delta}(f)(h) &\overset{\text{(a)}}{=} \lim_{t\to 0} \frac{\boldsymbol{\delta}(f+th) - \boldsymbol{\delta}(f)}{t} \overset{\text{(b)}}{=} \boldsymbol{\delta}(h) \overset{\text{(c)}}{=} (h(\boldsymbol{x}_1), \cdots, h(\boldsymbol{x}_N))^{\mathrm{T}} \\ &\overset{\text{(d)}}{=} (\langle h, K(\boldsymbol{x}_1, \cdot)\rangle_{\mathcal{B}^{\frac{m}{m-1}}}, \cdots, \langle h, K(\boldsymbol{x}_N, \cdot)\rangle_{\mathcal{B}^{\frac{m}{m-1}}})^{\mathrm{T}}, \end{aligned}$$

where (a) holds because of the definition of Gateaux derivative, (b) follows from the linearity of $\boldsymbol{\delta}$, (c) holds thanks to the definition of $\boldsymbol{\delta}$ and (d) holds because of the reproducing property of $\mathcal{B}^{\frac{m}{m-1}}$. Following the isometric isomorphism from $\mathcal{B}^m$ onto $(\mathcal{B}^{\frac{m}{m-1}})'$ mentioned in Subsection 2.1, it follows that

$$d_G\boldsymbol{\delta}(f) = (K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot))^{\mathrm{T}}. \tag{2.3}$$

By the chain rule [22, Proposition 6.17], it follows that

$$(d_G\boldsymbol{\delta}(f))^{\mathrm{T}}\partial F(\boldsymbol{\delta}(f)) = (K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot))\partial F(\boldsymbol{\delta}(f)) \subseteq \partial(F \circ \boldsymbol{\delta})(f). \tag{2.4}$$

Combining (2.2) with (2.4), we have that for any $f \neq 0$,

$$(K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot))\partial F(\boldsymbol{\delta}(f)) + \frac{m\lambda}{m-1}(\mathcal{J}_m)^{-1}(f) \subseteq \partial \mathcal{T}_{\frac{m}{m-1}}(f). \tag{2.5}$$

A basic question for optimization problem (1.1) is whether the minimizer exists. Next, we show the existence of the minimizer of optimization problem (1.1) and provide the space where the minimizer is located.

Lemma 2.1 *For any even integer $m \geq 4$ and lower semi-continuous loss function L, optimization problem* (1.1) *has a minimizer f_D. Moreover, for any minimizer f_D of optimization problem* (1.1)*, it follows that*

$$f_D \in \mathcal{J}_m(\mathrm{span}\{K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot)\}).$$

Proof First, we show the existence of the minimizer. As $\mathcal{T}_{\frac{m}{m-1}}$ is weakly lower semi-continuous on $\mathcal{B}^{\frac{m}{m-1}}$ and sequentially coercive in the weak topology of $\mathcal{B}^{\frac{m}{m-1}}$, [10, Theorem 1.15 (a)] shows that optimization problem (1.1) has a minimizer f_D.

If $f_D = 0$, then the proof is straightforward. If $f_D \neq 0$, we discuss the optimization problem

$$\begin{array}{ll} \min\limits_{f\in\mathcal{B}^{\frac{m}{m-1}}} & \|f\|_{\mathcal{B}^{\frac{m}{m-1}}}, \\ \text{s.t.} & \boldsymbol{\delta}(f) = \boldsymbol{\delta}(f_D). \end{array} \tag{2.6}$$

It is clear that f_D is in the feasible set. Since f_D is a minimizer of $\mathcal{T}_{\frac{m}{m-1}}$ on $\mathcal{B}^{\frac{m}{m-1}}$, for any f in the feasible set, we have that $\mathcal{T}_{\frac{m}{m-1}}(f_D) \leq \mathcal{T}_{\frac{m}{m-1}}(f)$. Since for any f in the feasible set, $(F\circ\boldsymbol{\delta})(f_D) = (F\circ\boldsymbol{\delta})(f)$ and $\lambda > 0$, $1 < \frac{m}{m-1} \leq \frac{4}{3}$, it follows that $\|f_D\|_{\mathcal{B}^{\frac{m}{m-1}}} \leq \|f\|_{\mathcal{B}^{\frac{m}{m-1}}}$. Thus,

f_D is a minimizer of optimization problem (2.6). From the optimal condition of optimization problem (2.6), there exists $\boldsymbol{\theta}_D \in \mathbb{R}^N$ such that

$$0 = d_G(\|\cdot\|_{\mathcal{B}^{\frac{m}{m-1}}})(f_D) + d_G\boldsymbol{\delta}(f_D)^{\mathrm{T}}\boldsymbol{\theta}_D. \tag{2.7}$$

Since $f_D \neq 0$, from (2.1), (2.3) and (2.7), we have that

$$\frac{(\mathcal{J}_m)^{-1}(f_D)}{\|f_D\|_{\mathcal{B}^{\frac{m}{m-1}}}^{\frac{1}{m-1}}} \in \mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\} \subseteq \mathcal{B}^m.$$

As $\|f_D\|_{\mathcal{B}^{\frac{m}{m-1}}} \neq 0$, we have that

$$f_D \in \mathcal{J}_m(\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}).$$

This completes the proof.

Lemma 2.1 shows the existence of a minimizer of optimization problem (1.1). The minimizer may not be unique, but all minimizers are contained in $\mathcal{J}_m(\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\})$. In the next subsection, we discuss how to equivalently transfer optimization problem (1.1) to a finite-dimensional tensor optimization problem.

2.3 Tensor optimization problem

This subsection discusses how to equivalently transfer infinite-dimensional optimization problem (1.1) to a finite-dimensional tensor optimization problem. First, we review a tensor, which is an extension of the Gram matrix (see [33, Section 3]). For the convenience of readers, the notations and operations of tensors are defined as in [23]. For a given $\mathcal{B}^{\frac{m}{m-1}}$ and the training data D, where $m \geq 4$ is an even integer, we denote

$$\boldsymbol{\Phi}_n := (\phi_n(\boldsymbol{x}_1), \phi_n(\boldsymbol{x}_2), \cdots, \phi_n(\boldsymbol{x}_N))^{\mathrm{T}} \in \mathbb{R}^N, \quad n \in \mathbb{N}.$$

Also, we define the following m-th order N-dimensional real tensor

$$\mathcal{A}_m := \Big(\sum_{n\in\mathbb{N}} \phi_n(\boldsymbol{x}_{i_1})\phi_n(\boldsymbol{x}_{i_2})\cdots\phi_n(\boldsymbol{x}_{i_m})\Big)_{i_1,i_2,\cdots,i_m=1}^{N,N,\cdots,N} = \sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n)^{\otimes m},$$

where $\otimes$ denotes the tensor outer product. Next, we introduce some operations with $\mathcal{A}_m$. For any $\boldsymbol{c} \in \mathbb{R}^N$, we denote

$$\begin{aligned}
\mathcal{A}_m\boldsymbol{c}^m &:= \Big(\sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n)^{\otimes m}\Big)\cdot\boldsymbol{c}^{\otimes m} = \sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^m \geq 0,\\
\mathcal{A}_m\boldsymbol{c}^{m-1} &:= \Big(\sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n)^{\otimes m}\Big)\cdot\boldsymbol{c}^{\otimes m-1} = \sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-1}\boldsymbol{\Phi}_n \in \mathbb{R}^N,\\
\mathcal{A}_m\boldsymbol{c}^{m-2} &:= \Big(\sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n)^{\otimes m}\Big)\cdot\boldsymbol{c}^{\otimes m-2} = \sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-2}\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}} \in \mathbb{R}^{N\times N}.
\end{aligned}$$

By the definition of $\mathcal{A}_m\boldsymbol{c}^m$, $\mathcal{A}_m\boldsymbol{c}^{m-1}$ and $\mathcal{A}_m\boldsymbol{c}^{m-2}$, we show that

$$\mathcal{A}_m\boldsymbol{c}^m = (\mathcal{A}_m\boldsymbol{c}^{m-1})^{\mathrm{T}}\boldsymbol{c}, \quad \mathcal{A}_m\boldsymbol{c}^{m-1} = \mathcal{A}_m\boldsymbol{c}^{m-2}\cdot\boldsymbol{c}. \tag{2.8}$$

Moreover, by the definition of $\mathcal{A}_m\boldsymbol{c}^{m-2}$, it is clear that $\mathcal{A}_m\boldsymbol{c}^{m-2}$ is symmetric. For any $\boldsymbol{d}\in\mathbb{R}^N$, we have that

$$\boldsymbol{d}^{\mathrm{T}}(\mathcal{A}_m\boldsymbol{c}^{m-2})\boldsymbol{d}=\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-2}\Big(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{d}\Big)^2\geq 0.$$

Thus, $\mathcal{A}_m\boldsymbol{c}^{m-2}$ is positive semi-definite. In conclusion, $\mathcal{A}_m\boldsymbol{c}^{m-2}$ is symmetric and positive semi-definite for any $\boldsymbol{c}\in\mathbb{R}^N$. Also, $\boldsymbol{c}\mapsto\mathcal{A}_m\boldsymbol{c}^m$ can be seen as a function from $\mathbb{R}^N$ to $\mathbb{R}$. By the derivative rule, we finally note that $\boldsymbol{c}\mapsto\mathcal{A}_m\boldsymbol{c}^m$ is a twice-differentiable function from $\mathbb{R}^N$ to $\mathbb{R}$ and for any $\boldsymbol{c}\in\mathbb{R}^N$,

$$\nabla(\mathcal{A}_m(\cdot)^m)(\boldsymbol{c})=m\mathcal{A}_m\boldsymbol{c}^{m-1},\quad \nabla^2(\mathcal{A}_m(\cdot)^m)(\boldsymbol{c})=m(m-1)\mathcal{A}_m\boldsymbol{c}^{m-2},\tag{2.9}$$

where ∇ represents the gradient and ∇^2 represents the Hessian matrix, respectively.

We now introduce how to equivalently transfer optimization problem (1.1) to a tensor optimization problem. By Lemma 2.1, for any $f\in\mathcal{J}_m(\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\})$, there exists $\boldsymbol{c}\in\mathbb{R}^N$ such that f has the representation

$$f=\mathcal{J}_m((K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot))\boldsymbol{c})=\mathcal{J}_m\Big(\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})\phi_n\Big)=\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-1}\phi_n,$$

which ensures that

$$\boldsymbol{\delta}(f)=(f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_N))^{\mathrm{T}}=\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-1}\boldsymbol{\Phi}_n=\mathcal{A}_m\boldsymbol{c}^{m-1}.\tag{2.10}$$

On the other hand, since m is an even integer and $m\geq 4$, by the definition of $\|\cdot\|_{\mathcal{B}^{\frac{m}{m-1}}}$, we compute

$$\lambda\|f\|_{\mathcal{B}^{\frac{m}{m-1}}}^{\frac{m}{m-1}}=\lambda\sum_{n\in\mathbb{N}}|(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-1}|^{\frac{m}{m-1}}=\lambda\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^m=\lambda\mathcal{A}_m\boldsymbol{c}^m.$$

Thus, optimization problem (1.1) can be equivalently transferred to the following tensor optimization problem

$$\min_{\boldsymbol{c}\in\mathbb{R}^N}\ F(\mathcal{A}_m\boldsymbol{c}^{m-1})+\lambda\mathcal{A}_m\boldsymbol{c}^m.$$

Let $\boldsymbol{\alpha}=\mathcal{A}_m\boldsymbol{c}^{m-1}$. Then we reformulate the optimization problem above as

$$\begin{array}{ll}\min\limits_{(\boldsymbol{\alpha},\boldsymbol{c})\in\mathbb{R}^{2N}} & F(\boldsymbol{\alpha})+\lambda\mathcal{A}_m\boldsymbol{c}^m,\\ \text{s.t.} & \boldsymbol{\alpha}=\mathcal{A}_m\boldsymbol{c}^{m-1}.\end{array}\tag{2.11}$$

Since $m\geq 4$, the constraint is nonlinear.

Example 2.1 Suppose that $N=1$ and $\mathcal{A}_m$ is an m-th order 1-dimensional identity tensor. Thus, optimization problem (2.11) can be written as

$$\begin{array}{ll}\min\limits_{a,c} & L(\boldsymbol{x}_1,y_1,\alpha)+\lambda c^m,\\ \text{s.t.} & \alpha=c^{m-1}.\end{array}$$

Next, we introduce some properties of optimization problem (2.11). Since L is a lower semi-continuous loss function, F is lower semi-continuous on $\mathbb{R}^N$. On the other hand, by (2.9), we have that $\boldsymbol{c}\mapsto\lambda\mathcal{A}_m\boldsymbol{c}^m$ is twice-differentiable on $\mathbb{R}^N$, and for any $\boldsymbol{c}\in\mathbb{R}^N$,

$$\nabla^2(\lambda\mathcal{A}_m(\cdot)^m)(\boldsymbol{c})=m(m-1)\lambda\mathcal{A}_m\boldsymbol{c}^{m-2}.$$

Since $\mathcal{A}_m \boldsymbol{c}^{m-2}$ is symmetric and positive semi-definite for any $\boldsymbol{c} \in \mathbb{R}^N$, [2, Proposition 1.1.10 (ii)] shows that $\boldsymbol{c} \mapsto \lambda \mathcal{A}_m \boldsymbol{c}^m$ is convex on $\mathbb{R}^N$. In conclusion, optimization problem (2.11) is lower semi-continuous on $\mathbb{R}^{2N}$ with nonlinear constraint. In the next section, we discuss how to solve optimization problem (1.1) based on optimization problem (2.11).

3 Splitting Method for Support Vector Machine in $\frac{m}{m-1}$-norm Reproducing Kernel Banach Space

In this section, we find an algorithm based on optimization problem (2.11) for optimization problem (1.1). Currently, we mainly use the subgradient method, Lagrangian multipliers method, and sequential minimal optimization (SMO for short) for SVM. These classical algorithms are suitable for convex and smooth optimization problem. We would like to find algorithms for lower semi-continuous optimization problem (2.11).

The ADMM algorithm, as one of the splitting techniques, can even be used for nonsmooth and nonconvex optimization problem with linear constraint. For general lower semi-continuous loss functions, we observe that the subproblems in ADMM for optimization problem (2.11) are then easier to handle. Hence, we study how to obtain the minimizer of optimization problem (1.1) by splitting method based on ADMM. Although optimization problem (2.11) has a nonlinear constraint, we still follow the idea of ADMM to write the steps of the splitting method. Recall that the augmented Lagrangian function for optimization problem (2.11) is defined as:

$$\mathcal{L}_\beta(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) := F(\boldsymbol{\alpha}) + \lambda \mathcal{A}_m \boldsymbol{c}^m + \boldsymbol{\gamma}^{\mathrm{T}}(\boldsymbol{\alpha} - \mathcal{A}_m \boldsymbol{c}^{m-1}) + \frac{\beta}{2}\|\boldsymbol{\alpha} - \mathcal{A}_m \boldsymbol{c}^{m-1}\|^2,$$

where $\beta > 0$. Suppose that the splitting method based on ADMM for optimization problem (1.1) is initialized at $(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0, g_0, s_0)$, where $g_0 = (K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot))\boldsymbol{c}_0$ and $s_0 = \mathcal{J}_m(g_0)$, its iterative scheme is

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

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

$$\boldsymbol{\gamma}_{k+1} := \boldsymbol{\gamma}_k + \beta(\boldsymbol{\alpha}_{k+1} - \mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}), \tag{S-3}$$

$$g_{k+1} := (K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot))\boldsymbol{c}_{k+1}, \tag{S-4}$$

$$s_{k+1} := \mathcal{J}_m(g_{k+1}), \tag{S-5}$$

where k represents the number of iteration. Next, we discuss subproblems (S-1) and (S-2).

Remark 3.1 It is worth mentioning that our main goal is to use $\{s_k\}$ to approximate the minimizer of optimization problem (1.1) rather than solving optimization problem (2.11). Hence, we mainly focus on $\{s_k\}$.

As for (S-1), combining the linear with quadratic terms of $\mathcal{L}_\beta(\boldsymbol{\alpha}, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)$, we have that

$$\mathcal{L}_\beta(\boldsymbol{\alpha}, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) = F(\boldsymbol{\alpha}) + \frac{\beta}{2}\left\|\boldsymbol{\alpha} - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_k\right\|^2 + \lambda \mathcal{A}_m(\boldsymbol{c}_k)^m - \frac{1}{2\beta}\|\boldsymbol{\gamma}_k\|^2.$$

Since (S-1) only depends on $\boldsymbol{\alpha}$, and F is lower semi-continuous and nonnegative on $\mathbb{R}^N$, it is easy to check that (S-1) is lower semi-continuous, nonnegative and coercive. Hence, Weierstrass

Theorem (see [1, Theorem 2.14]) assures that (S-1) has a minimizer. Moreover, since

$$F(\boldsymbol{\alpha})+\frac{\beta}{2}\Big\|\boldsymbol{\alpha}-\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}+\frac{1}{\beta}\boldsymbol{\gamma}_k\Big\|^2=\sum_{i=1}^{N}\frac{L(\boldsymbol{x}_i,y_i,\alpha_i)}{N}+\frac{\beta}{2}\Big(\alpha_i-(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1})_i+\frac{1}{\beta}(\boldsymbol{\gamma}_k)_i\Big)^2$$

and $\frac{L(\boldsymbol{x}_i,y_i,\alpha_i)}{N}+\frac{\beta}{2}\big(\alpha_i-(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1})_i+\frac{1}{\beta}(\boldsymbol{\gamma}_k)_i\big)^2\geq 0,\ \ i=1,2,\cdots,N$, we equivalently transfer (S-1) in $\mathbb{R}^N$ to some optimization problems in $\mathbb{R}$, that is, for $i=1,2,\cdots,N$,

$$(\boldsymbol{\alpha}_{k+1})_i\in\underset{\alpha_i\in\mathbb{R}}{\operatorname{argmin}}\ \frac{L(\boldsymbol{x}_i,y_i,\alpha_i)}{N}+\frac{\beta}{2}\Big(\alpha_i-(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1})_i+\frac{1}{\beta}(\boldsymbol{\gamma}_k)_i\Big)^2. \tag{S-1$'$}$$

For a general lower semi-continuous loss function, (S-1$'$) may have more than one minimizer. In this case, we choose one of the minimizers as $(\boldsymbol{\alpha}_{k+1})_i,\ i=1,2,\cdots,N$.

As for (S-2), combining the linear with quadratic terms of $\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c},\boldsymbol{\gamma}_k)$, it follows that for any $\boldsymbol{c}\in\mathbb{R}^N$,

$$\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c},\boldsymbol{\gamma}_k)=\lambda\mathcal{A}_m\boldsymbol{c}^m+\frac{\beta}{2}\Big\|\boldsymbol{\alpha}_{k+1}-\mathcal{A}_m\boldsymbol{c}^{m-1}+\frac{1}{\beta}\boldsymbol{\gamma}_k\Big\|^2+F(\boldsymbol{\alpha}_{k+1})-\frac{1}{2\beta}\|\boldsymbol{\gamma}_k\|^2.$$

By the derivative rule, it is easy to see that (S-2) is differentiable on $\mathbb{R}^N$ and

$$\begin{aligned}\nabla\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\cdot,\boldsymbol{\gamma}_k)(\boldsymbol{c})&\overset{\text{(a)}}{=}m\lambda\mathcal{A}_m\boldsymbol{c}^{m-1}-\beta(m-1)\mathcal{A}_m\boldsymbol{c}^{m-2}\cdot\Big(\boldsymbol{\alpha}_{k+1}-\mathcal{A}_m\boldsymbol{c}^{m-1}+\frac{1}{\beta}\boldsymbol{\gamma}_k\Big)\\&\overset{\text{(b)}}{=}\beta(m-1)\mathcal{A}_m\boldsymbol{c}^{m-2}\cdot\Big(\mathcal{A}_m\boldsymbol{c}^{m-1}+\frac{m\lambda}{(m-1)\beta}\boldsymbol{c}-\Big(\boldsymbol{\alpha}_{k+1}+\frac{1}{\beta}\boldsymbol{\gamma}_k\Big)\Big),\end{aligned}$$

where (a) holds thanks to derivative rule and (2.9), (b) follows from (2.8) and rearranging terms. Next, we show that the following tensor equation

$$\mathcal{A}_m\boldsymbol{c}^{m-1}+\frac{m\lambda}{(m-1)\beta}\boldsymbol{c}=\boldsymbol{\alpha}_{k+1}+\frac{1}{\beta}\boldsymbol{\gamma}_k \tag{S-2$'$}$$

has a unique solution $\boldsymbol{d}_{k+1}$, and $\boldsymbol{d}_{k+1}$ is a minimizer of (S-2). To this end, let

$$H(\boldsymbol{c}):=\frac{1}{m}\mathcal{A}_m\boldsymbol{c}^m+\frac{m\lambda}{2(m-1)\beta}\|\boldsymbol{c}\|^2-\Big(\boldsymbol{\alpha}_{k+1}+\frac{1}{\beta}\boldsymbol{\gamma}_k\Big)^{\mathrm{T}}\boldsymbol{c}.$$

By the derivative rule and (2.9), for any $\boldsymbol{c}\in\mathbb{R}^N$, we have

$$\begin{aligned}\nabla H(\boldsymbol{c})&=\mathcal{A}_m\boldsymbol{c}^{m-1}+\frac{m\lambda}{(m-1)\beta}\boldsymbol{c}-\Big(\boldsymbol{\alpha}_{k+1}+\frac{1}{\beta}\boldsymbol{\gamma}_k\Big),\\\nabla^2H(\boldsymbol{c})&=(m-1)\mathcal{A}_m\boldsymbol{c}^{m-2}+\frac{m\lambda}{(m-1)\beta}I,\end{aligned}$$

where I denotes the identity matrix of N-dimensional. Hence, $\boldsymbol{c}$ is a stationary point of H if and only if $\boldsymbol{c}$ is a solution of tensor equation (S-2$'$). Moreover, since $\mathcal{A}_m\boldsymbol{c}^{m-2}$ is symmetric and positive semi-definite for any $\boldsymbol{c}\in\mathbb{R}^N$, it follows that $\nabla^2H(\boldsymbol{c})$ is positive definite. Therefore, [2, Proposition 1.1.10 (ii)] guarantees that H is strictly convex on $\mathbb{R}^N$. Since H is also coercive, H has a unique minimizer $\boldsymbol{d}_{k+1}$ by Weierstrass Theorem, which means that H has a unique stationary point $\boldsymbol{d}_{k+1}$. In conclusion, tensor equation (S-2$'$) has a unique solution $\boldsymbol{d}_{k+1}$.

Now we show that $\boldsymbol{d}_{k+1}$ is a minimizer of optimization problem (S-2). To this end, for any $\boldsymbol{c} \in \mathbb{R}^N$, we have that

$$\begin{aligned}
&\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}, \boldsymbol{\gamma}_k) - \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{d}_{k+1}, \boldsymbol{\gamma}_k) \\
=&\lambda \mathcal{A}_m \boldsymbol{c}^m + \frac{\beta}{2}\Big\|\boldsymbol{\alpha}_{k+1} - \mathcal{A}_m \boldsymbol{c}^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_k\Big\|^2 \\
&- \lambda \mathcal{A}_m (\boldsymbol{d}_{k+1})^m - \frac{\beta}{2}\Big\|\boldsymbol{\alpha}_{k+1} - \mathcal{A}_m (\boldsymbol{d}_{k+1})^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_k\Big\|^2 \\
\overset{\text{(a)}}{=}&\lambda \mathcal{A}_m \boldsymbol{c}^m - \lambda \mathcal{A}_m (\boldsymbol{d}_{k+1})^m + \frac{\beta}{2}\Big\|\mathcal{A}_m (\boldsymbol{d}_{k+1})^{m-1} - \mathcal{A}_m \boldsymbol{c}^{m-1} + \frac{m\lambda}{(m-1)\beta}\boldsymbol{d}_{k+1}\Big\|^2 \\
&- \frac{\beta}{2}\Big\|\frac{m\lambda}{(m-1)\beta}\boldsymbol{d}_{k+1}\Big\|^2 \\
\overset{\text{(b)}}{=}&\lambda \mathcal{A}_m \boldsymbol{c}^m - \lambda \mathcal{A}_m (\boldsymbol{d}_{k+1})^m + \Big(\frac{m\lambda}{m-1}\boldsymbol{d}_{k+1}\Big)^{\mathrm{T}} (\mathcal{A}_m (\boldsymbol{d}_{k+1})^{m-1} - \mathcal{A}_m \boldsymbol{c}^{m-1}) \\
&+ \frac{\beta}{2}\|\mathcal{A}_m (\boldsymbol{d}_{k+1})^{m-1} - \mathcal{A}_m \boldsymbol{c}^{m-1}\|^2 \\
\overset{\text{(c)}}{=}&\frac{\lambda}{m-1}\mathcal{A}_m (\boldsymbol{d}_{k+1})^m - \frac{\lambda}{m-1}\mathcal{A}_m \boldsymbol{c}^m - \Big(\frac{m\lambda}{m-1}\mathcal{A}_m \boldsymbol{c}^{m-1}\Big)^{\mathrm{T}} (\boldsymbol{d}_{k+1} - \boldsymbol{c}) \\
&+ \frac{\beta}{2}\|\mathcal{A}_m (\boldsymbol{d}_{k+1})^{m-1} - \mathcal{A}_m \boldsymbol{c}^{m-1}\|^2,
\end{aligned}$$

where (a) follows from tensor equation (S-2′), (b) holds as rearranging terms, and (c) holds thanks to (2.8) and rearranging terms. Since $\boldsymbol{c} \mapsto \frac{\lambda}{m-1}\mathcal{A}_m \boldsymbol{c}^m$ is differentiable and convex on $\mathbb{R}^N$, and $\nabla\big(\frac{\lambda}{m-1}\mathcal{A}_m(\cdot)^m\big)(\boldsymbol{c}) = \frac{m\lambda}{m-1}\mathcal{A}_m \boldsymbol{c}^{m-1}$ for any $\boldsymbol{c} \in \mathbb{R}^N$, [2, Proposition 1.1.7 (a)] shows that

$$\frac{\lambda}{m-1}\mathcal{A}_m (\boldsymbol{d}_{k+1})^m \geq \frac{\lambda}{m-1}\mathcal{A}_m \boldsymbol{c}^m + \Big(\frac{m\lambda}{m-1}\mathcal{A}_m \boldsymbol{c}^{m-1}\Big)^{\mathrm{T}} (\boldsymbol{d}_{k+1} - \boldsymbol{c}). \tag{3.1}$$

As $\frac{\beta}{2}\|\mathcal{A}_m (\boldsymbol{d}_{k+1})^{m-1} - \mathcal{A}_m \boldsymbol{c}^{m-1}\|^2 \geq 0$, it follows that for any $\boldsymbol{c} \in \mathbb{R}^N$,

$$\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}, \boldsymbol{\gamma}_k) \geq \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{d}_{k+1}, \boldsymbol{\gamma}_k).$$

Therefore, $\boldsymbol{d}_{k+1}$ is a minimizer of optimization problem (S-2). In conclusion, we take $\boldsymbol{c}_{k+1} = \boldsymbol{d}_{k+1}$. We consider using the Newton method for tensor equation (S-2′), whose convergence can be guaranteed by [21, Newton Attraction Theorem 10.2.2].

When $\boldsymbol{\alpha}_{k+1}$ and $\boldsymbol{c}_{k+1}$ are acquired, we can obtain $\boldsymbol{\gamma}_{k+1}$ by (S-3). However, substituting (S-3) into the tensor equation (S-2′), that is,

$$\begin{cases}
\mathcal{A}_m (\boldsymbol{c}_{k+1})^{m-1} + \dfrac{m\lambda}{(m-1)\beta}\boldsymbol{c}_{k+1} = \boldsymbol{\alpha}_{k+1} + \dfrac{1}{\beta}\boldsymbol{\gamma}_k, \\
\boldsymbol{\gamma}_{k+1} = \boldsymbol{\gamma}_k + \beta(\boldsymbol{\alpha}_{k+1} - \mathcal{A}_m (\boldsymbol{c}_{k+1})^{m-1}),
\end{cases}$$

we have that

$$\boldsymbol{\gamma}_{k+1} = \frac{m\lambda}{m-1}\boldsymbol{c}_{k+1}. \tag{S-3′}$$

In conclusion, the splitting method is well-defined, and a sequence $\{(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k, g_k, s_k)\}$ is generated. Also, $\{s_k\}$ can be seen as an infinite iterative sequence to approximate f_D. Next, we present the splitting method for optimization problem (1.1).

Algorithm 1 Splitting method for optimization problem (1.1)

Require: initial value $(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0, g_0, s_0)$, the training data D, lower semi-continuous loss function L, the RKBS $\mathcal{B}^{\frac{m}{m-1}}$, $\lambda > 0$ and $\beta > 0$.

Step 1: Solve optimization problems (S-1′) and take the output as $\boldsymbol{\alpha}_{k+1}$.

Step 2: Solve the tensor equation (S-2′) by Newton method and take the output as $\boldsymbol{c}_{k+1}$, where the initial value of Newton method is $\boldsymbol{c}_k$.

Step 3: Set $\boldsymbol{\gamma}_{k+1} \leftarrow \frac{m\lambda}{m-1}\boldsymbol{c}_{k+1}$.

Step 4: Set $g_{k+1} \leftarrow (K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot))\boldsymbol{c}_{k+1}$.

Step 5: Set $s_{k+1} \leftarrow \mathcal{J}_m(g_{k+1})$.

Update $k \leftarrow k+1$ and go to Step 1.

Example 3.1 As for the optimization problem in Example 2.1, the corresponding iterative scheme of the splitting method can be written as

$$\begin{aligned}
&\alpha_{k+1} \in \operatorname*{argmin}_{\alpha\in\mathbb{R}} \ L(\boldsymbol{x}_1, y_1, \alpha) + \frac{\beta}{2}\Big(\alpha - (c_k)^{m-1} + \frac{1}{\beta}\gamma_k\Big)^2,\\
&c_{k+1} \in \operatorname*{argmin}_{c\in\mathbb{R}} \ \lambda c^m + \frac{\beta}{2}\Big(\alpha_{k+1} - c^{m-1} + \frac{1}{\beta}\gamma_k\Big)^2,\\
&\gamma_{k+1} = \gamma_k + \beta(\alpha_{k+1} - (c_{k+1})^{m-1}),\\
&g_{k+1} := K(\boldsymbol{x}_1, \cdot)c_{k+1},\\
&s_{k+1} := \mathcal{J}_m(g_{k+1}).
\end{aligned}$$

Usually, we assume that α_{k+1} is easy to compute. Also, we solve the nonlinear equation

$$c^{m-1} + \frac{m\lambda}{(m-1)\beta}c = \alpha_{k+1} + \frac{1}{\beta}\gamma_k$$

to obtain c_{k+1}. At last, we obtain γ_{k+1} by $\frac{m\lambda}{m-1}c_{k+1}$.

In the next section, we verify that under some assumptions, $\{s_k\}$ is globally convergent to a stationary point of optimization problem (1.1). Hence, it is better to solve optimization problem (1.1) repeatedly by selecting some initial values randomly and choosing the minimizer of these outputs as the approximate solution $s_D : X \to \mathbb{R}$. Finally, we construct $\mathcal{R}_D$ by s_D according to the task requirement.

4 Convergence Analysis

In this section, we investigate the convergence of $\{s_k\}$ inspired by the papers [13, 15, 19] and use a similar line of arguments therein.

4.1 Assumption

In Sections 2–3, we assume that L is a lower semi-continuous loss function. Also, from Subsection 2.3, we show that for any even integer $m \geq 4$, $\mathcal{A}_m\boldsymbol{c}^{m-2}$ is symmetric and positive

semi-definite for any $\boldsymbol{c} \in \mathbb{R}^N$. To ensure the convergence of $\{s_k\}$, we need some additional conditions which we describe below.

Assumption 4.1 For any even integer $m \geq 4$, the following conditions for optimization problem (1.1) hold

(i) L is a lower semi-continuous and subanalytic loss function.

(ii) There exist $\xi_1, \xi_2 > 0$ such that for any $\boldsymbol{x} \in X$, $y \in Y$ and $t \in [-\xi_1, \xi_1]$, $\partial L(\boldsymbol{x}, y, t) \cap [-\xi_2, \xi_2] = \emptyset$.

(iii) $\mathcal{A}_m \boldsymbol{c}^{m-2}$ is a symmetric and positive definite matrix for any $\boldsymbol{c} \neq \boldsymbol{0}$.

Now we give the sufficient conditions of Assumption 4.1. Subanalytic functions are quite wide, including semi-algebraic, analytic and semi-analytic functions (see [11, 6.6 Analytic Problems]). More precisely, polynomial functions and piecewise polynomial functions are subanalytic functions. However, subanalyticity does not even imply continuity. Specially, some margin-based loss functions (see [28, Section 2.3]) satisfy Assumption 4.1 (i)–(ii), such as the least square loss, the Hinge loss, the truncated least squares loss, logistic loss, and so on. As for Assumption 4.1 (iii), we need the following concept. We denote the space consisting of all real symmetric matrices of N-dimensional as $\mathbb{S}^{N\times N}$. It is easy to check that $\mathbb{S}^{N\times N}$ is an $\frac{N(N+1)}{2}$-dimensional space. For any $n \in \mathbb{N}$, it follows that

$$\boldsymbol{\Phi}_n \boldsymbol{\Phi}_n^{\mathrm{T}} = (\phi_n(\boldsymbol{x}_{i_1})\phi_n(\boldsymbol{x}_{i_2}))_{i_1,i_2=1}^{N,N} \in \mathbb{S}^{N\times N}.$$

If $\{\boldsymbol{\Phi}_n \boldsymbol{\Phi}_n^{\mathrm{T}} : n \in \mathbb{N}\}$ has full-rank in $\mathbb{S}^{N\times N}$, then we show that Assumption 4.1 (iii) holds. Since $\mathcal{A}_m \boldsymbol{c}^{m-2}$ is symmetric and positive semi-definite for any $\boldsymbol{c} \neq \boldsymbol{0}$, it suffices to prove that for any $\boldsymbol{d} \in \mathbb{R}^N$, if $\boldsymbol{d}^{\mathrm{T}}(\mathcal{A}_m \boldsymbol{c}^{m-2})\boldsymbol{d} = 0$, then $\boldsymbol{d} = \boldsymbol{0}$. To this end, by the definition of $\mathcal{A}_m \boldsymbol{c}^{m-2}$, it follows that

$$\boldsymbol{d}^{\mathrm{T}}(\mathcal{A}_m \boldsymbol{c}^{m-2})\boldsymbol{d} = \sum_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{c})^{m-2} (\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{d})^2 = 0.$$

Therefore, we assure that

$$(\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{c})(\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{d}) = (\boldsymbol{\Phi}_n \boldsymbol{\Phi}_n^{\mathrm{T}}) \cdot (\boldsymbol{c}\boldsymbol{d}^{\mathrm{T}}) = 0, \quad n \in \mathbb{N}, \tag{4.1}$$

where "$\cdot$" denotes the inner product of the matrix. Let $E := (e_{i_1 i_2})_{i_1,i_2=1}^{N,N}$ and

$$e_{i_1 i_2} := \begin{cases} c_{i_1} d_{i_1}, & i_1 = i_2, \\ \frac{1}{2} c_{i_1} d_{i_2} + \frac{1}{2} c_{i_2} d_{i_1}, & i_1 \neq i_2. \end{cases}$$

Then $E \in \mathbb{S}^{N\times N}$, and for any $n \in \mathbb{N}$,

$$\begin{aligned} &\sum_{i_1,i_2=1}^{N,N} \phi_n(\boldsymbol{x}_{i_1})\phi_n(\boldsymbol{x}_{i_2}) c_{i_1} d_{i_2} \\ &= \sum_{i_1=i_2}^{N,N} \phi_n(\boldsymbol{x}_{i_1})\phi_n(\boldsymbol{x}_{i_1}) c_{i_1} d_{i_1} + \sum_{i_1\neq i_2}^{N,N} \phi_n(\boldsymbol{x}_{i_1})\phi_n(\boldsymbol{x}_{i_2}) \Big(\frac{1}{2} c_{i_1} d_{i_2} + \frac{1}{2} c_{i_2} d_{i_1}\Big), \end{aligned}$$

which ensures that

$$(\boldsymbol{\Phi}_n \boldsymbol{\Phi}_n^{\mathrm{T}}) \cdot (\boldsymbol{c}\boldsymbol{d}^{\mathrm{T}}) = (\boldsymbol{\Phi}_n \boldsymbol{\Phi}_n^{\mathrm{T}}) \cdot E, \quad n \in \mathbb{N}. \tag{4.2}$$

From (4.1)–(4.2), we obtain the following system

$$(\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1^{\mathrm{T}},\cdots,\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}},\cdots)^{\mathrm{T}}\cdot(\boldsymbol{c}\boldsymbol{d}^{\mathrm{T}})=(\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1^{\mathrm{T}},\cdots,\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}},\cdots)^{\mathrm{T}}\cdot E=(0,\cdots,0,\cdots)^{\mathrm{T}}.$$

Since $\{\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}}:n\in\mathbb{N}\}$ has full-rank, the system above only has zero solution, that is, $E=O$, where O denotes zero matrix in $\mathbb{S}^{N\times N}$. As $\boldsymbol{c}\neq\boldsymbol{0}$, there exists $c_{i_1}\neq 0$ such that

$$\begin{cases}e_{i_1i_1}=c_{i_1}d_{i_1}=0,\\ e_{i_1i_2}=\dfrac{1}{2}c_{i_1}d_{i_2}+\dfrac{1}{2}c_{i_1}d_{i_2}=0,\quad i_2=1,2,\cdots,N,\ i_1\neq i_2.\end{cases}$$

Thus, we check that $d_1=\cdots=d_N=0$, that is, $\boldsymbol{d}=\boldsymbol{0}$. In conclusion, $\mathcal{A}_m\boldsymbol{c}^{m-2}$ is symmetric and positive definite when $\boldsymbol{c}\neq\boldsymbol{0}$.

In the rest of this subsection, we discuss what can be drawn under Assumption 4.1. First, we show that $\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}$ is linearly independent in $\mathcal{B}^m$ under Assumption 4.1 (iii). For any $\boldsymbol{c}\in\mathbb{R}^N$ such that

$$(K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot))\boldsymbol{c}=\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})\phi_n=0,$$

we see that

$$\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^2=\boldsymbol{c}^{\mathrm{T}}\Big(\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})\boldsymbol{\Phi}_n\Big)=\boldsymbol{c}^{\mathrm{T}}\boldsymbol{0}=0.$$

Therefore, it follows that $\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c}=0,\ n\in\mathbb{N}$ and

$$\boldsymbol{c}^{\mathrm{T}}(\mathcal{A}_m\boldsymbol{c}^{m-2})\boldsymbol{c}=\mathcal{A}_m\boldsymbol{c}^m=\sum_{n\in\mathbb{N}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^m=0.$$

By Assumption 4.1 (iii), if $\boldsymbol{c}\neq\boldsymbol{0}$, then $\boldsymbol{c}^{\mathrm{T}}(\mathcal{A}_m\boldsymbol{c}^{m-2})\boldsymbol{c}>0$. Hence, we have that $\boldsymbol{c}=\boldsymbol{0}$, which means that $\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}$ is linearly independent. Also, $\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}$ is N-dimensional linear space. Thus, for any $g\in\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}\subseteq\mathcal{B}^m$, there exists a unique $\boldsymbol{c}_g\in\mathbb{R}^N$ such that

$$g=(K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot))\boldsymbol{c}_g.$$

Also, we check that $g\mapsto\boldsymbol{c}_g$ is a linear mapping from $\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}$ onto $\mathbb{R}^N$. As for $g\mapsto\boldsymbol{c}_g$, we have the following lemma.

Lemma 4.1 *If Assumption* 4.1 (iii) *holds, then $g\mapsto\boldsymbol{c}_g$ is an isomorphism. Moreover, there exist $0<w_1\leq w_2$ such that*

$$w_1\|g_1-g_2\|_{\mathcal{B}^m}\leq\|\boldsymbol{c}_{g_1}-\boldsymbol{c}_{g_2}\|\leq w_2\|g_1-g_2\|_{\mathcal{B}^m}\quad \textit{for } g_1,g_2\in\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}.$$

Proof Since $g\mapsto\boldsymbol{c}_g$ is a linear mapping from $\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}$ onto $\mathbb{R}^N$, [18, Theorem 1.4.15] shows that $g\mapsto\boldsymbol{c}_g$ is an isomorphism. Moreover, [18, Proposition 1.4.14(a)] assures the inequality above. This proof is completed.

Next, we discuss what can be drawn under Assumption 4.1 and $\beta>0$. Since $F(\boldsymbol{\alpha})=\frac{1}{N}\sum\limits_{i=1}^{N}L(\boldsymbol{x}_i,y_i,\alpha_i)$, [24, Proposition 10.5] and [24, D. Rescaling] assure that

$$\partial F(\boldsymbol{\alpha})=\Big(\frac{1}{N}\partial L(\boldsymbol{x}_1,y_1,\alpha_1)\Big)\times\cdots\times\Big(\frac{1}{N}\partial L(\boldsymbol{x}_N,y_N,\alpha_N)\Big).$$

Thus, by Assumption 4.1 (ii), for any $\boldsymbol{x} \in X$, $y \in Y$ and $\boldsymbol{\alpha} \in [-\xi_1, \xi_1]^N$,

$$\partial F(\boldsymbol{\alpha}) \cap [-\xi_2, \xi_2]^N = \emptyset. \tag{4.3}$$

Also, by simple algebra, it is easy to check that $t \mapsto \frac{2m\lambda}{(m-1)\beta}t + \|\mathcal{A}_m\|_F t^{m-1}$ and $t \mapsto \frac{m\lambda}{m-1}t + 2\beta\|\mathcal{A}_m\|_F t^{m-1}$ are strictly increasing on $[0, \infty)$, where $\|\cdot\|_F$ denotes the Frobenius norm of tensor (see [23, Section 1.1]). Clearly, $\frac{2m\lambda}{(m-1)\beta}0 + \|\mathcal{A}_m\|_F 0^{m-1} = 0$ and $\frac{m\lambda}{m-1}0 + 2\beta\|\mathcal{A}_m\|_F 0^{m-1} = 0$. Hence, for $\xi_1, \xi_2 > 0$, we denote

$$\varepsilon_\beta := \max\Big\{t \geq 0 : \frac{2m\lambda}{(m-1)\beta}t + \|\mathcal{A}_m\|_F t^{m-1} \leq \xi_1,\ \frac{m\lambda}{m-1}t + 2\beta\|\mathcal{A}_m\|_F t^{m-1} \leq \xi_2\Big\}.$$

Then $\varepsilon_\beta > 0$. ε_β not only depends on Assumption 4.1, but also depends on $\beta > 0$. Next, we show that ε_β is a lower bound of $\{\|(\boldsymbol{c}_k, \boldsymbol{c}_{k+1})\|\}$.

Lemma 4.2 *Suppose that Assumption* 4.1 (ii) *holds and* $\beta > 0$. *If Algorithm* 1 *is initialized at* $(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0, g_0, s_0)$, *then for any* $k \in \mathbb{N}$, $\|(\boldsymbol{c}_k, \boldsymbol{c}_{k+1})\| > \varepsilon_\beta$.

Proof To finish the proof, we assume that there exists $k_0 \in \mathbb{N}$ such that $\|(\boldsymbol{c}_{k_0}, \boldsymbol{c}_{k_0+1})\| \leq \varepsilon_\beta$. From the optimality condition of (S-1), the iterates generated satisfy

$$\mathbf{0} \in \partial F(\boldsymbol{\alpha}_{k_0+1}) + \boldsymbol{\gamma}_{k_0} + \beta(\boldsymbol{\alpha}_{k_0+1} - \mathcal{A}_m(\boldsymbol{c}_{k_0})^{m-1}). \tag{4.4}$$

On the other hand, as $\|(\boldsymbol{c}_{k_0}, \boldsymbol{c}_{k_0+1})\| \leq \varepsilon_\beta$, it follows that $\|\boldsymbol{c}_{k_0}\| \leq \varepsilon_\beta$ and $\|\boldsymbol{c}_{k_0+1}\| \leq \varepsilon_\beta$. Hence, we have that

$$\begin{aligned}
\|\boldsymbol{\alpha}_{k_0+1}\| &\overset{\text{(a)}}{=} \Big\|\frac{1}{\beta}(\boldsymbol{\gamma}_{k_0+1} - \boldsymbol{\gamma}_{k_0}) + \mathcal{A}_m(\boldsymbol{c}_{k_0+1})^{m-1}\Big\| \leq \frac{1}{\beta}\|\boldsymbol{\gamma}_{k_0+1}\| + \frac{1}{\beta}\|\boldsymbol{\gamma}_{k_0}\| + \|\mathcal{A}_m(\boldsymbol{c}_{k_0+1})^{m-1}\| \\
&\overset{\text{(b)}}{\leq} \frac{m\lambda}{(m-1)\beta}(\|\boldsymbol{c}_{k_0+1}\| + \|\boldsymbol{c}_{k_0}\|) + \|\mathcal{A}_m\|_F\|\boldsymbol{c}_{k_0+1}\|^{m-1} \\
&\leq \frac{2m\lambda}{(m-1)\beta}\varepsilon_\beta + \|\mathcal{A}_m\|_F(\varepsilon_\beta)^{m-1},
\end{aligned}$$

where (a) holds thanks to (S-3) and (b) follows from (S-3′) and [23, Lemma 2.2]. Also, we have

$$\begin{aligned}
\|\boldsymbol{\gamma}_{k_0} + \beta(\boldsymbol{\alpha}_{k_0+1} - \mathcal{A}_m(\boldsymbol{c}_{k_0})^{m-1})\| &\overset{\text{(a)}}{=} \|\boldsymbol{\gamma}_{k_0+1} + \beta(\mathcal{A}_m(\boldsymbol{c}_{k_0+1})^{m-1} - \mathcal{A}_m(\boldsymbol{c}_{k_0})^{m-1})\| \\
&\leq \|\boldsymbol{\gamma}_{k_0+1}\| + \beta\|\mathcal{A}_m(\boldsymbol{c}_{k_0+1})^{m-1}\| + \beta\|\mathcal{A}_m(\boldsymbol{c}_{k_0})^{m-1}\| \\
&\overset{\text{(b)}}{\leq} \frac{m\lambda}{m-1}\|\boldsymbol{c}_{k_0+1}\| + \beta\|\mathcal{A}_m\|_F(\|\boldsymbol{c}_{k_0+1}\|^{m-1} + \|\boldsymbol{c}_{k_0}\|^{m-1}) \\
&\leq \frac{m\lambda}{m-1}\varepsilon_\beta + 2\beta\|\mathcal{A}_m\|_F(\varepsilon_\beta)^{m-1},
\end{aligned}$$

where (a) also holds thanks to (S-3), (b) follows from (S-3′) and [23, Lemma 2.2]. Hence, $\boldsymbol{\alpha}_{k_0+1} \in [-\xi_1, \xi_1]^N$ and $\boldsymbol{\gamma}_{k_0} + \beta(\boldsymbol{\alpha}_{k_0+1} - \mathcal{A}_m(\boldsymbol{c}_{k_0})^{m-1}) \in [-\xi_2, \xi_2]^N$. By (4.3), it follows that

$$\mathbf{0} \notin \partial F(\boldsymbol{\alpha}_{k_0+1}) + \boldsymbol{\gamma}_{k_0} + \beta(\boldsymbol{\alpha}_{k_0+1} - \mathcal{A}_m(\boldsymbol{c}_{k_0})^{m-1}). \tag{4.5}$$

Clearly, (4.4) and (4.5) are contradictions. Thus, $\|(\boldsymbol{c}_k, \boldsymbol{c}_{k+1})\| > \varepsilon_\beta$ for any $k \in \mathbb{N}$. This proof is completed.

Now we verify another important inequality by Assumption 4.1 (ii)–(iii) and $\beta > 0$. We need to the following concept. For any even integer $m \geq 4$, we define $\chi_m : \mathbb{R}^{2N} \to \mathbb{R}$,

$$\chi_m(\boldsymbol{c}, \boldsymbol{d}) := \lambda_{\min}\Big(\frac{1}{2}\mathcal{A}_m \boldsymbol{c}^{m-2} + \frac{1}{2}\mathcal{A}_m \boldsymbol{d}^{m-2}\Big), \quad (\boldsymbol{c}, \boldsymbol{d}) \in \mathbb{R}^{2N},$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of matrix. Since $m \geq 4$, it follows that $m-2 \geq 2$ and $\mathcal{A}_m \mathbf{0}^{m-2} = \sum\limits_{n\in\mathbb{N}} (\boldsymbol{\Phi}_n^{\mathrm{T}} \mathbf{0})^{m-2} \boldsymbol{\Phi}_n \boldsymbol{\Phi}_n^{\mathrm{T}} = O$. Therefore, $\chi_m(\mathbf{0}, \mathbf{0}) = 0$. Since $\mathcal{A}_m \boldsymbol{c}^{m-2}$ is symmetric and positive definite for any $\boldsymbol{c} \neq \mathbf{0}$ by Assumption 4.1 (iii), we have that $\chi_m(\boldsymbol{c}, \boldsymbol{d}) > 0$ when $(\boldsymbol{c}, \boldsymbol{d}) \neq (\mathbf{0}, \mathbf{0})$. Moreover, by the definition of $\mathcal{A}_m \boldsymbol{c}^{m-2}$, it is easy to check that for any $t > 0$ and $(\boldsymbol{c}, \boldsymbol{d}) \in \mathbb{R}^{2N}$,

$$\frac{1}{2}\mathcal{A}_m (t\boldsymbol{c})^{m-2} + \frac{1}{2}\mathcal{A}_m (t\boldsymbol{d})^{m-2} = t^{m-2}\Big(\frac{1}{2}\mathcal{A}_m \boldsymbol{c}^{m-2} + \frac{1}{2}\mathcal{A}_m \boldsymbol{d}^{m-2}\Big).$$

Thus

$$\chi_m(t\boldsymbol{c}, t\boldsymbol{d}) = t^{m-2}\chi_m(\boldsymbol{c}, \boldsymbol{d}),$$

which ensures that χ_m is coercive on $\mathbb{R}^{2N}$. Hence, [34, Exercise 1.15] shows that there exists $\mu > 0$ such that when $\|(\boldsymbol{c}, \boldsymbol{d})\| > \mu$, it follows that $\chi_m(\boldsymbol{c}, \boldsymbol{d}) > 1$. We denote

$$\nu_\beta := \Big(\frac{\varepsilon_\beta}{\mu}\Big)^{m-2}.$$

Then $\nu_\beta > 0$. Based on $\varepsilon_\beta > 0$, $\nu_\beta > 0$ and χ_m, we have the following inequality.

Lemma 4.3 *Suppose that Assumption* 4.1 (ii)–(iii) *holds and* $\beta > 0$. *If* $\|(\boldsymbol{c}, \boldsymbol{d})\| > \varepsilon_\beta$, *then*

$$(\boldsymbol{c} - \boldsymbol{d})^{\mathrm{T}}(\mathcal{A}_m \boldsymbol{c}^{m-1} - \mathcal{A}_m \boldsymbol{d}^{m-1}) \geq \nu_\beta \|\boldsymbol{c} - \boldsymbol{d}\|^2.$$

Proof Since $m \geq 4$, we have that

$$\begin{aligned}
&(\boldsymbol{c} - \boldsymbol{d})^{\mathrm{T}}(\mathcal{A}_m \boldsymbol{c}^{m-1} - \mathcal{A}_m \boldsymbol{d}^{m-1}) \\
&\overset{\text{(a)}}{=} \sum_{n\in\mathbb{N}} ((\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{c})^{m-1} - (\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{d})^{m-1})(\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{c} - \boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{d}) \\
&\overset{\text{(b)}}{\geq} \sum_{n\in\mathbb{N}} \Big(\sum_{j=0}^{m-2} (\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{c})^{m-2-j} (\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{d})^j\Big)(\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{c} - \boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{d})^2 \\
&\overset{\text{(c)}}{=} (\boldsymbol{c} - \boldsymbol{d})^{\mathrm{T}}\Big(\sum_{n\in\mathbb{N}} \Big(\frac{1}{2}(\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{c})^{m-2} + \frac{1}{2}(\boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{d})^{m-2}\Big)\boldsymbol{\Phi}_n \boldsymbol{\Phi}_n^{\mathrm{T}}\Big)(\boldsymbol{c} - \boldsymbol{d}) \\
&\overset{\text{(d)}}{=} (\boldsymbol{c} - \boldsymbol{d})^{\mathrm{T}}\Big(\frac{1}{2}\mathcal{A}_m \boldsymbol{c}^{m-2} + \frac{1}{2}\mathcal{A}_m \boldsymbol{d}^{m-2}\Big)(\boldsymbol{c} - \boldsymbol{d}),
\end{aligned} \tag{4.6}$$

where (a) follows from the definition of $\mathcal{A}_m \boldsymbol{c}^{m-1}$, (b) holds because of the fact that $(a-b)(a^{m-1} - b^{m-1}) \geq \frac{1}{2}(a^{m-2} + b^{m-2})(a-b)^2$ for $a, b \in \mathbb{R}$, (c) holds thanks to rearranging terms and (d) follows from the definition of $\mathcal{A}_m \boldsymbol{c}^{m-2}$. When $\|(\boldsymbol{c}, \boldsymbol{d})\| > \varepsilon_\beta$, it follows that $\big\|\big(\frac{\mu}{\varepsilon_\beta}\boldsymbol{c}, \frac{\mu}{\varepsilon_\beta}\boldsymbol{d}\big)\big\| > \mu$ and $\chi_m\big(\frac{\mu}{\varepsilon_\beta}\boldsymbol{c}, \frac{\mu}{\varepsilon_\beta}\boldsymbol{d}\big) > 1$. Hence

$$\chi_m(\boldsymbol{c}, \boldsymbol{d}) = \Big(\frac{\varepsilon_\beta}{\mu}\Big)^{m-2} \chi_m\Big(\frac{\mu}{\varepsilon_\beta}\boldsymbol{c}, \frac{\mu}{\varepsilon_\beta}\boldsymbol{d}\Big) > \Big(\frac{\varepsilon_\beta}{\mu}\Big)^{m-2} = \nu_\beta,$$

that is, $\frac{1}{2}\mathcal{A}_m\boldsymbol{c}^{m-2}+\frac{1}{2}\mathcal{A}_m\boldsymbol{d}^{m-2}-\nu_\beta I$ is symmetric and semi-positive definite, which ensures that

$$(\boldsymbol{c}-\boldsymbol{d})^{\mathrm{T}}\Big(\frac{1}{2}\mathcal{A}_m\boldsymbol{c}^{m-2}+\frac{1}{2}\mathcal{A}_m\boldsymbol{d}^{m-2}\Big)(\boldsymbol{c}-\boldsymbol{d})\geq\nu_\beta\|\boldsymbol{c}-\boldsymbol{d}\|^2. \tag{4.7}$$

Combining (4.6) with (4.7), we have that

$$(\boldsymbol{c}-\boldsymbol{d})^{\mathrm{T}}(\mathcal{A}_m\boldsymbol{c}^{m-1}-\mathcal{A}_m\boldsymbol{d}^{m-1})\geq\nu_\beta\|\boldsymbol{c}-\boldsymbol{d}\|^2.$$

This proof is completed.

Next, we give the descent inequality by Lemmas 4.1–4.3.

Lemma 4.4 *Suppose that Assumption* 4.1 (ii)–(iii) *holds and* $\beta\nu_\beta>\frac{2m\lambda}{m-1}$. *If Algorithm* 1 *is initialized at* $(\boldsymbol{\alpha}_0,\boldsymbol{c}_0,\boldsymbol{\gamma}_0,g_0,s_0)$, *then there exists* $(\zeta_1)_\beta>0$ *such that for any* $k\in\mathbb{N}$,

$$\zeta_1\|g_k-g_{k+1}\|^2_{\mathcal{B}^m}\leq\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)-\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c}_{k+1},\boldsymbol{\gamma}_{k+1}).$$

Moreover, we show that $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)\}$ *is monotonically decreasing and lower bounded. Furthermore,* $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)\}$ *is convergent.*

Proof First, we show that $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)\}$ is monotonically decreasing. From (S-1), we know that $\boldsymbol{\alpha}_{k+1}$ is the minimizer of $\boldsymbol{\alpha}\mapsto\mathcal{L}_\beta(\boldsymbol{\alpha},\boldsymbol{c}_k,\boldsymbol{\gamma}_k)$. Thus

$$0\leq\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)-\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c}_k,\boldsymbol{\gamma}_k). \tag{I-1}$$

Moreover, from the definition of $\mathcal{L}_\beta$, we use $\boldsymbol{c}_k$ and $\boldsymbol{c}_{k+1}$ to replace $\boldsymbol{\alpha}_{k+1}$, $\boldsymbol{\gamma}_k$ and $\boldsymbol{\gamma}_{k+1}$ in $\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c}_k,\boldsymbol{\gamma}_k)-\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c}_{k+1},\boldsymbol{\gamma}_{k+1})$, that is,

$$\begin{aligned}
&\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c}_k,\boldsymbol{\gamma}_k)-\mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1},\boldsymbol{c}_{k+1},\boldsymbol{\gamma}_k)\\
&\overset{\text{(a)}}{=}\lambda\mathcal{A}_m(\boldsymbol{c}_k)^m-\lambda\mathcal{A}_m(\boldsymbol{c}_{k+1})^m-(\boldsymbol{\gamma}_k)^{\mathrm{T}}(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1})\\
&\quad+\frac{\beta}{2}\|\boldsymbol{\alpha}_{k+1}-\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}\|^2-\frac{\beta}{2}\|\boldsymbol{\alpha}_{k+1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\|^2\\
&\overset{\text{(b)}}{=}\lambda\mathcal{A}_m(\boldsymbol{c}_k)^m-\lambda\mathcal{A}_m(\boldsymbol{c}_{k+1})^m-(\boldsymbol{\gamma}_k)^{\mathrm{T}}(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1})\\
&\quad+\frac{\beta}{2}\Big\|\frac{1}{\beta}(\boldsymbol{\gamma}^{k+1}-\boldsymbol{\gamma}^k)+\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}-\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}\Big\|^2-\frac{\beta}{2}\Big\|\frac{1}{\beta}(\boldsymbol{\gamma}^{k+1}-\boldsymbol{\gamma}^k)\Big\|^2\\
&\overset{\text{(c)}}{=}\lambda\mathcal{A}_m(\boldsymbol{c}_k)^m-\lambda\mathcal{A}_m(\boldsymbol{c}_{k+1})^m-(\boldsymbol{\gamma}_{k+1})^{\mathrm{T}}(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1})\\
&\quad+\frac{\beta}{2}\|\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\|^2\\
&\overset{\text{(d)}}{=}\lambda\mathcal{A}_m(\boldsymbol{c}_k)^m-\lambda\mathcal{A}_m(\boldsymbol{c}_{k+1})^m-\Big(\frac{m\lambda}{m-1}\boldsymbol{c}_{k+1}\Big)^{\mathrm{T}}(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1})\\
&\quad+\frac{\beta}{2}\|\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\|^2\\
&\overset{\text{(e)}}{=}\frac{\lambda}{m-1}\mathcal{A}_m(\boldsymbol{c}_{k+1})^m-\frac{\lambda}{m-1}\mathcal{A}_m(\boldsymbol{c}_k)^m-\Big(\frac{m\lambda}{m-1}\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}\Big)^{\mathrm{T}}(\boldsymbol{c}_{k+1}-\boldsymbol{c}_k)\\
&\quad+\frac{\beta}{2}\|\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\|^2,
\end{aligned}\tag{4.8}$$

where (a) holds because of rearranging terms, (b) follows from (S-3), (c) holds thanks to rearranging terms, (d) follows from (S-3′) and (e) follows from (2.9). From (3.1), we take $\boldsymbol{d}_{k+1} = \boldsymbol{c}_{k+1}$ and $\boldsymbol{c} = \boldsymbol{c}_k$, it follows that

$$\frac{\lambda}{m-1}\mathcal{A}_m(\boldsymbol{c}_{k+1})^m - \frac{\lambda}{m-1}\mathcal{A}_m(\boldsymbol{c}_k)^m - \Big(\frac{m\lambda}{m-1}\mathcal{A}_m(\boldsymbol{c}_k)^{m-1}\Big)^{\mathrm{T}}(\boldsymbol{c}_{k+1} - \boldsymbol{c}_k) \geq 0. \tag{4.9}$$

On the other hand, since Lemma 4.2 assures that $\|(\boldsymbol{c}_k, \boldsymbol{c}_{k+1})\| > \varepsilon_\beta$ for any $k \in \mathbb{N}$, the Cauchy-Schwartz inequality and Lemma 4.3 show that

$$\begin{aligned}
&\|\boldsymbol{c}_k - \boldsymbol{c}_{k+1}\|\|\mathcal{A}_m(\boldsymbol{c}_k)^{m-1} - \mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\| \\
\geq&(\boldsymbol{c}_k - \boldsymbol{c}_{k+1})^{\mathrm{T}}(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1} - \mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}) \\
\geq&\nu_\beta\|\boldsymbol{c}_k - \boldsymbol{c}_{k+1}\|^2,
\end{aligned}$$

which ensures that

$$\|\mathcal{A}_m(\boldsymbol{c}_k)^{m-1} - \mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\| \geq \nu_\beta\|\boldsymbol{c}_k - \boldsymbol{c}_{k+1}\|. \tag{4.10}$$

Since $g_k, g_{k+1} \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot)\}$, Lemma 4.1 and (4.10) show that

$$\frac{\beta}{2}\|\mathcal{A}_m(\boldsymbol{c}_k)^{m-1} - \mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\|^2 \geq \frac{\beta(\nu_\beta)^2}{2}\|\boldsymbol{c}_k - \boldsymbol{c}_{k+1}\|^2 \geq \frac{\beta(\nu_\beta w_1)^2}{2}\|g_k - g_{k+1}\|_{\mathcal{B}^m}^2. \tag{4.11}$$

From (4.8)–(4.9) and (4.11), we have that

$$\frac{\beta(\nu_\beta w_1)^2}{2}\|g_k - g_{k+1}\|_{\mathcal{B}^m}^2 \leq \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) - \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}_{k+1}, \boldsymbol{\gamma}_k). \tag{I-2}$$

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

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

Since $g_k, g_{k+1} \in \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot)\}$, combining Lemma 4.1 with (S-3′), we see that

$$-\frac{m^2\lambda^2(w_2)^2}{(m-1)^2\beta}\|g_k - g_{k+1}\|_{\mathcal{B}^m}^2 \leq -\frac{m^2\lambda^2}{(m-1)^2\beta}\|\boldsymbol{c}_k - \boldsymbol{c}_{k+1}\|^2 = -\frac{1}{\beta}\|\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_{k+1}\|^2.$$

From the two inequalities above, we have that

$$-\frac{m^2\lambda^2(w_2)^2}{(m-1)^2\beta}\|g_k - g_{k+1}\|_{\mathcal{B}^m}^2 \leq \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}_{k+1}, \boldsymbol{\gamma}_k) - \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}_{k+1}, \boldsymbol{\gamma}_{k+1}). \tag{I-3}$$

In conclusion, (I-1), (I-2) and (I-3) assure that

$$\Big(\frac{\beta(\nu_\beta w_1)^2}{2} - \frac{m^2\lambda^2(w_2)^2}{(m-1)^2\beta}\Big)\|g_k - g_{k+1}\|_{\mathcal{B}^m}^2 \leq \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) - \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}_{k+1}, \boldsymbol{\gamma}_{k+1}).$$

Since $\beta\nu_\beta > \frac{2m\lambda}{m-1}$ and $w_2 \geq w_1$, we have

$$\frac{\beta(\nu_\beta w_1)^2}{2} - \frac{m^2\lambda^2(w_2)^2}{(m-1)^2\beta} > \frac{2m^2\lambda^2(w_1)^2}{(m-1)^2\beta} - \frac{m^2\lambda^2(w_1)^2}{(m-1)^2\beta} = \frac{m^2\lambda^2(w_1)^2}{(m-1)^2\beta}.$$

We denote $(\zeta_1)_\beta := \frac{m^2\lambda^2(w_1)^2}{(m-1)^2\beta}$. Then $(\zeta_1)_\beta > 0$. Thus, we finally have that for any $k \in \mathbb{N}$,

$$(\zeta_1)_\beta \|g_k - g_{k+1}\|^2_{\mathcal{B}^m} \leq \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) - \mathcal{L}_\beta(\boldsymbol{\alpha}_{k+1}, \boldsymbol{c}_{k+1}, \boldsymbol{\gamma}_{k+1}), \tag{4.12}$$

which ensures that $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ is monotonically decreasing.

Next, we show that $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ is lower bounded. For any $k \in \mathbb{N}$, we see that

$$\begin{aligned}
\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) &\overset{(a)}{=} F(\boldsymbol{\alpha}_k) + \lambda\mathcal{A}_m(\boldsymbol{c}_k)^m + (\boldsymbol{\gamma}_k)^{\mathrm{T}}(\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}) + \frac{\beta}{2}\|\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}\|^2 \\
&\overset{(b)}{=} F(\boldsymbol{\alpha}_k) + \lambda\mathcal{A}_m(\boldsymbol{c}_k)^m - \frac{1}{2\beta}\|\boldsymbol{\gamma}_k\|^2 + \frac{\beta}{2}\Big\|\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_k\Big\|^2 \\
&\overset{(c)}{=} F(\boldsymbol{\alpha}_k) + \lambda\mathcal{A}_m(\boldsymbol{c}_k)^m - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}_k\|^2 + \frac{\beta}{2}\Big\|\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_k\Big\|^2 \\
&\overset{(d)}{\geq} \lambda\mathcal{A}_m(\boldsymbol{c}_k)^m - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}_k\|^2,
\end{aligned} \tag{4.13}$$

where (a) holds because of the definition of $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$, (b) follows from rearranging terms, (c) holds thanks to (S-3′) and (d) follows from $F(\boldsymbol{\alpha}_k) \geq 0$ and $\frac{\beta}{2}\|\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_k\|^2 \geq 0$. To find the lower bound of $\lambda\mathcal{A}_m(\boldsymbol{c}_k)^m - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}_k\|^2$, we need to consider the following two cases.

(I) As $\boldsymbol{c} \mapsto \lambda\mathcal{A}_m\boldsymbol{c}^m - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}\|^2$ is continuous on the closed ball $\{\boldsymbol{c} \in \mathbb{R}^N : \|\boldsymbol{c}\| \leq \varepsilon_\beta\}$, there exists $\tau_\beta \in \mathbb{R}$ such that when $\|\boldsymbol{c}\| \leq \varepsilon_\beta$, it follows that $\lambda\mathcal{A}_m\boldsymbol{c}^m - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}\|^2 \geq \tau_\beta$. Thus, if $\|\boldsymbol{c}_k\| \leq \varepsilon_\beta$, then

$$\lambda\mathcal{A}_m(\boldsymbol{c}_k)^m - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}_k\|^2 \geq \tau_\beta.$$

(II) If $\|\boldsymbol{c}_k\| > \varepsilon_\beta$, then $\|(\boldsymbol{c}_k, \boldsymbol{0})\| > \varepsilon_\beta$. Thus, it follows that

$$\begin{aligned}
\lambda\mathcal{A}_m(\boldsymbol{c}_k)^m - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}_k\|^2 &\overset{(a)}{=} \lambda(\boldsymbol{c}_k - \boldsymbol{0})^{\mathrm{T}}(\mathcal{A}_m(\boldsymbol{c}_k)^{m-1} - \mathcal{A}_m\boldsymbol{0}^{m-1}) - \frac{m^2\lambda^2}{2(m-1)^2\beta}\|\boldsymbol{c}_k\|^2 \\
&\overset{(b)}{\geq} \Big(\lambda\nu_\beta - \frac{m^2\lambda^2}{2(m-1)^2\beta}\Big)\|\boldsymbol{c}_k\|^2 \\
&\overset{(c)}{>} \frac{\lambda\nu_\beta}{2}\|\boldsymbol{c}_k\|^2 \geq 0,
\end{aligned} \tag{4.14}$$

where (a) holds because of (2.8), (b) follows from Lemma 4.3, and (c) holds thanks to the fact that $-\frac{m^2\lambda^2}{2(m-1)^2\beta} \geq -\frac{m\lambda}{(m-1)\beta} > -\frac{\nu_\beta}{2}$ by $\beta\nu_\beta > \frac{2m\lambda}{m-1}$ and $2 > \frac{4}{3} \geq \frac{m}{m-1}$. From (4.14) and (I)–(II), we show that for any $k \in \mathbb{N}$,

$$\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) \geq \min\{\tau_\beta, 0\}.$$

Thus, $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ is lower bounded. Since $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ is monotonically decreasing and lower bounded, [25, Theorem 3.24] shows the convergence of $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$. This proof is completed.

By Lemma 4.4, we can define the residual of $\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)$,

$$r_k := \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) - \lim_{k\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k).$$

By the definition of $\{r_k\}$ and Lemma 4.4, we have that $\{r_k\}$ is monotonically decreasing and $\lim_{k\to\infty} r_k = 0$. Next, we show the boundedness of $\{(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ by Lemma 4.4.

Lemma 4.5 *Suppose that Assumption* 4.1 (ii)–(iii) *holds and* $\beta\nu_\beta > \frac{2m\lambda}{m-1}$. *If Algorithm* 1 *is initialized at* $(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0, g_0, s_0)$, *then* $\{(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ *is bounded.*

Proof Since $\{\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ is monotonically decreasing by Lemma 4.4, we have that

$$\mathcal{L}_\beta(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0) \geq \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) \quad \text{for } k \in \mathbb{N}. \tag{4.15}$$

If $\|\boldsymbol{c}_k\| > \varepsilon_\beta$, then (4.13)–(4.15) assure that

$$\|\boldsymbol{c}_k\| \leq \sqrt{\frac{2}{\lambda\nu_\beta}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)} \leq \sqrt{\frac{2}{\lambda\nu_\beta}\mathcal{L}_\beta(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0)}.$$

Hence

$$\|\boldsymbol{c}_k\| \leq \max\Big\{\sqrt{\frac{2}{\lambda\nu_\beta}\mathcal{L}_\beta(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0)}, \varepsilon_\beta\Big\}, \quad \text{for } k \in \mathbb{N},$$

that is, $\{\boldsymbol{c}_k\}$ is bounded. Moreover, (S-3′) shows that $\{\boldsymbol{\gamma}_k\}$ is also bounded. Furthermore, by [23, Lemma 2.2] and (S-3), we have that $\|\boldsymbol{\alpha}_k\| \leq \frac{1}{\beta}(\|\boldsymbol{\gamma}_k\| + \|\boldsymbol{\gamma}_{k-1}\|) + \|\mathcal{A}_m\|_F\|\boldsymbol{c}_k\|^{m-1}$. Thus, $\{(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$ is bounded. This proof is completed.

Let S be the set of subsequential limits of $\{(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)\}$. Then [25, Theorems 3.6–3.7] show that if Assumption 4.1 (ii)–(iii) holds and $\beta\nu_\beta > \frac{2m\lambda}{m-1}$, then S is nonempty compact, and

$$\lim_{k\to\infty} \operatorname{dist}((\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k), S) = 0, \tag{4.16}$$

where $\operatorname{dist}(\cdot,\cdot)$ denotes the Euclidean distance. Next, we verify some properties of $\mathcal{L}_\beta$ on S.

By Assumption 4.1 (i), since L is a lower semi-continuous and subanalytic loss function, [27, (I.2.1.9)] shows that F is nonnegative, lower semi-continuous and subanalytic. Also, $\boldsymbol{c} \mapsto \lambda\mathcal{A}_m\boldsymbol{c}^m$ is nonnegative, continuous and semi-algebraic. Moreover, $(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) \mapsto \boldsymbol{\gamma}^{\mathrm{T}}(\boldsymbol{\alpha} - \mathcal{A}_m\boldsymbol{c}^{m-1})$ is continuous and semi-algebraic. By Cauchy-Schwartz inequality and [23, Lemma 2.2], we have that

$$|\boldsymbol{\gamma}^{\mathrm{T}}(\boldsymbol{\alpha} - \mathcal{A}_m\boldsymbol{c}^{m-1})| \leq \|\boldsymbol{\gamma}\|\|\boldsymbol{\alpha} - \mathcal{A}_m\boldsymbol{c}^{m-1}\| \leq \|\boldsymbol{\gamma}\|(\|\boldsymbol{\alpha}\| + \|\mathcal{A}_m\|_F\|\boldsymbol{c}\|^{m-1}).$$

Thus, $(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) \mapsto \boldsymbol{\gamma}^{\mathrm{T}}(\boldsymbol{\alpha} - \mathcal{A}_m\boldsymbol{c}^{m-1})$ is bounded for any bounded set in $\mathbb{R}^{3N}$. Finally, $(\boldsymbol{\alpha}, \boldsymbol{c}) \mapsto \frac{\beta}{2}\|\boldsymbol{\alpha} - \mathcal{A}_m\boldsymbol{c}^{m-1}\|^2$ is nonnegative, continuous and semi-algebraic. Since semi-algebraic function is a subanalytic function, [27, (I.2.1.9)] shows that $\mathcal{L}_\beta$ is subanalytic. By [3, 4, 32], it follows that $\mathcal{L}_\beta$ is a KL function on $\mathbb{R}^{3N}$, which ensures that $\mathcal{L}_\beta$ is a KL function on S.

Moreover, to show that $\mathcal{L}_\beta$ has uniformized KL property on S, we verify that $\mathcal{L}_\beta$ 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 $(\boldsymbol{\alpha}_*, \boldsymbol{c}_*, \boldsymbol{\gamma}_*)$. Hence, the lower semi-continuity of $\mathcal{L}_\beta$ at $(\boldsymbol{\alpha}_*, \boldsymbol{c}_*, \boldsymbol{\gamma}_*)$ and Lemma 4.4 show that

$$\mathcal{L}_\beta(\boldsymbol{\alpha}_*, \boldsymbol{c}_*, \boldsymbol{\gamma}_*) \leq \liminf_{j\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_{k_j}, \boldsymbol{c}_{k_j}, \boldsymbol{\gamma}_{k_j}) = \lim_{k\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k). \tag{4.17}$$

Conversely, since $\boldsymbol{\alpha}^{k_j+1}$ minimizes $\boldsymbol{\alpha} \mapsto \mathcal{L}_\beta(\boldsymbol{\alpha}, \boldsymbol{c}^{k_j}, \boldsymbol{\gamma}^{k_j})$, (I-1)–(I-3) show that

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

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

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

On the other hand, Lemma 4.4 shows that

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

Also, passing to the limit along $\{(\boldsymbol{\alpha}_{k_j}, \boldsymbol{c}_{k_j}, \boldsymbol{\gamma}_{k_j})\}$, [25, Theorem 3.19] assures that

$$\mathcal{L}_\beta(\boldsymbol{\alpha}_*, \boldsymbol{c}_*, \boldsymbol{\gamma}_*) \geq \lim_{k\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k). \tag{4.18}$$

Finally, (4.17)–(4.18) assure that

$$\mathcal{L}_\beta(\boldsymbol{\alpha}_*, \boldsymbol{c}_*, \boldsymbol{\gamma}_*) = \lim_{k\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k). \tag{4.19}$$

Hence, $\mathcal{L}_\beta$ is constant on S. In conclusion, $\mathcal{L}_\beta$ has uniformized KL property on S (see [5, Lemma 3.6]), that is, there exist $\varepsilon > 0$, $\eta > 0$ and a continuous concave function $\varphi : [0, \eta) \to (0, \infty)$ such that

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

(ii) For any $(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) \in \mathbb{R}^{3N}$ such that $\operatorname{dist}((\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}), S) < \varepsilon$ and $\lim_{k\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) < \mathcal{L}_\beta(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) < \lim_{k\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) + \eta$, it follows that

$$\varphi'(\mathcal{L}_\beta(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}) - \lim_{k\to\infty} \mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)) \operatorname{dist}(0, \partial\mathcal{L}_\beta(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma})) \geq 1. \tag{4.20}$$

Based on the uniformized KL property of $\mathcal{L}_\beta$ on S, we derive an important inequality.

Lemma 4.6 *Suppose that Assumption* 4.1 *holds and* $\beta\nu_\beta > \frac{2m\lambda}{m-1}$. *If Algorithm* 1 *is initialized at* $(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0, g_0, s_0)$ *and for any* $k \in \mathbb{N}$, $r_k > 0$, *then there exists* $k_1 \in \mathbb{N}$ *such that*

$$\varphi'(r_k) \inf_{\substack{(\boldsymbol{\alpha},\boldsymbol{c},\boldsymbol{\gamma})\in \\ \partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)}} \|(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma})\| \geq 1 \quad \text{for } k > k_1.$$

Proof From Lemma 4.4 and (4.16), there exists $k_1 \in \mathbb{N}$ such that if $k > k_1$, then we have that $\operatorname{dist}((\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k), S) < \varepsilon$ and $r_k < \eta$. Since for any $k \in \mathbb{N}$, $r_k > 0$, (4.20) assures that

$$\varphi'(r_k) \operatorname{dist}(0, \partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)) = \varphi'(r_k) \inf_{\substack{(\boldsymbol{\alpha},\boldsymbol{c},\boldsymbol{\gamma})\in \\ \partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)}} \|(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma})\| \geq 1 \quad \text{for } k > k_1.$$

This proof is completed.

In this subsection, we discuss what conditions satisfy Assumption 4.1 and derive some facts of optimization problem (1.1) under Assumption 4.1. In the next subsection, we use Lemmas 4.1–4.6 to discuss the convergence of $\{s_k\}$ under Assumption 4.1.

4.2 Convergence of splitting method

In this subsection, we discuss the convergence of $\{s_k\}$ under Assumption 4.1.

Theorem 4.1 *Suppose that Assumption* 4.1 *holds and* $\beta\nu_\beta > \frac{2m\lambda}{m-1}$. *If Algorithm* 1 *is initialized at* $(\boldsymbol{\alpha}_0, \boldsymbol{c}_0, \boldsymbol{\gamma}_0, g_0, s_0)$, *then* $\{s_k\}$ *converges to a stationary point* s_* *of optimization problem* (1.1), *that is,*

$$\lim_{k\to\infty} \|s_k - s_*\|_{\mathcal{B}^{\frac{m}{m-1}}} = 0, \quad 0 \in \partial\mathcal{T}_{\frac{m}{m-1}}(s_*).$$

The main idea for proving Theorem 4.1 is to find a lower bound and an upper bound of $\{r_k - r_{k+1}\}$. Combining this upper bound with this lower bound, we verify the convergence of $\{g_k\}$. Since $\mathcal{J}_m$ is a homeomorphism, we verify the convergence of $\{s_k\}$. Finally, we show that $\{s_*\}$ is a stationary point of optimization problem (1.1).

Before presenting our main result in this section, we introduce two useful inequalities that play crucial roles in estimating the error bounds of $\{g_k\}$. First, by the definition of $\{r_k\}$ and Lemma 4.4, (4.12) can be rewritten and we find a lower bound of $\{r_k - r_{k+1}\}$.

Lemma 4.7 *Suppose that the conditions in Theorem* 4.1 *hold. Then there exists* $(\zeta_1)_\beta > 0$ *such that*

$$(\zeta_1)_\beta \|g_k - g_{k+1}\|^2_{\mathcal{B}^m} \le r_k - r_{k+1} \quad \textit{for } k \in \mathbb{N}.$$

On the other hand, we find an upper bound of $\{r_k - r_{k+1}\}$ by Lemma 4.6.

Lemma 4.8 *Suppose that the conditions in Theorem* 4.1 *hold and for any* $k \in \mathbb{N}$, $r_k > 0$. *Then there exists* $(\zeta_2)_\beta > 0$ *such that*

$$r_k - r_{k+1} \le (\zeta_2)_\beta \|g_{k-1} - g_k\|_{\mathcal{B}^m}(\varphi(r_k) - \varphi(r_{k+1})) \quad \textit{for } k > k_1.$$

Proof From the concavity of φ, we get that

$$\varphi'(r_k)(r_k - r_{k+1}) \le \varphi(r_k) - \varphi(r_{k+1}).$$

Combining Lemma 4.6 and the inequality above, we obtain that

$$r_k - r_{k+1} \le \inf_{\substack{(\boldsymbol{\alpha},\boldsymbol{c},\boldsymbol{\gamma}) \\ \in\partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)}} \|(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma})\|(\varphi(r_k) - \varphi(r_{k+1})) \quad \text{for } k > k_1. \tag{4.21}$$

Next, we find an element in $\partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)$. To this end, by [24, Exercise 8.8(c), Proposition 10.5] and (S-3′), it follows that

$$\partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) = \partial_{\boldsymbol{\alpha}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) \times \nabla_{\boldsymbol{c}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) \times \nabla_{\boldsymbol{\gamma}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k), \tag{4.22}$$

where

$$\begin{aligned}
&\partial_{\boldsymbol{\alpha}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) = \partial F(\boldsymbol{\alpha}_k) + \boldsymbol{\gamma}_k + \beta(\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}),\\
&\nabla_{\boldsymbol{c}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) = -(m-1)\mathcal{A}_m(\boldsymbol{c}_k)^{m-2} \cdot \beta(\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}),\\
&\nabla_{\boldsymbol{\gamma}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k) = \boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}.
\end{aligned}$$

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

$$-\beta\Big(\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_{k-1})^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_{k-1}\Big) \in \partial F(\boldsymbol{\alpha}_k). \tag{4.23}$$

From (S-3), (S-3′) and (4.22)–(4.23), we obtain further that

$$\begin{aligned}
\boldsymbol{\alpha}_k^{\#} &:= \beta(\mathcal{A}_m(\boldsymbol{c}_{k-1})^{m-1} - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}) - \frac{m\lambda}{m-1}(\boldsymbol{c}_{k-1} - \boldsymbol{c}_k) \in \partial_{\boldsymbol{\alpha}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k),\\
\boldsymbol{c}_k^{\#} &:= m\lambda\mathcal{A}_m(\boldsymbol{c}_k)^{m-2} \cdot (\boldsymbol{c}_{k-1} - \boldsymbol{c}_k) = \nabla_{\boldsymbol{c}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k),\\
\boldsymbol{\gamma}_k^{\#} &:= -\frac{m\lambda}{(m-1)\beta}(\boldsymbol{c}_{k-1} - \boldsymbol{c}_k) = \nabla_{\boldsymbol{\gamma}}\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k).
\end{aligned}$$

Hence, $(\boldsymbol{\alpha}_k^{\#}, \boldsymbol{c}_k^{\#}, \boldsymbol{\gamma}_k^{\#}) \in \partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k, \boldsymbol{c}_k, \boldsymbol{\gamma}_k)$. It means that

$$\inf_{\substack{(\boldsymbol{\alpha},\boldsymbol{c},\boldsymbol{\gamma})\\ \in\partial\mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)}} \|(\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma})\| \le \|(\boldsymbol{\alpha}_k^{\#}, \boldsymbol{c}_k^{\#}, \boldsymbol{\gamma}_k^{\#})\| \le \|\boldsymbol{\alpha}_k^{\#}\| + \|\boldsymbol{c}_k^{\#}\| + \|\boldsymbol{\gamma}_k^{\#}\|. \tag{4.24}$$

To finish the proof, we need to find upper bounds of $\|\boldsymbol{\alpha}_k^{\#}\|$, $\|\boldsymbol{c}_k^{\#}\|$ and $\|\boldsymbol{\gamma}_k^{\#}\|$.

From the fundamental theorem of calculus and (2.9), we have that

$$\mathcal{A}_m(\boldsymbol{c}_{k-1})^{m-1} - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1} = \int_0^1 (m-1)\mathcal{A}_m(\boldsymbol{c}_k + t(\boldsymbol{c}_{k-1} - \boldsymbol{c}_k))^{m-2}\mathrm{d}t \cdot (\boldsymbol{c}_{k-1} - \boldsymbol{c}_k).$$

By Lemma 4.5, it follows that $\{\boldsymbol{c}_k\}$ is bounded, that is, there exists $M > 0$ such that $\|\boldsymbol{c}_k\| \le M$ for any $k \in \mathbb{N}$. Since the closed ball $\{\boldsymbol{c} \in \mathbb{R}^N : \|\boldsymbol{c}\| \le M\}$ is convex, when $t \in [0, 1]$, $\|\boldsymbol{c}_{k-1} + t(\boldsymbol{c}_k - \boldsymbol{c}_{k-1})\| \le M$. From [23, Lemma 2.2], we see that

$$\Big\| \int_0^1 (m-1)\mathcal{A}_m(\boldsymbol{c}_{k-1} + t(\boldsymbol{c}_k - \boldsymbol{c}_{k-1}))^{m-2}\mathrm{d}t \Big\| \le (m-1)\|\mathcal{A}_m\|_F M^{m-2}.$$

Therefore

$$\|\mathcal{A}_m(\boldsymbol{c}_{k-1})^{m-1} - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}\| \le (m-1)\|\mathcal{A}_m\|_F M^{m-2}\|\boldsymbol{c}_{k-1} - \boldsymbol{c}_k\|. \tag{4.25}$$

Since $g_{k-1}, g_k \in \mathrm{span}\{K(\boldsymbol{x}_1, \cdot), \cdots, K(\boldsymbol{x}_N, \cdot)\}$, inserting Lemma 4.1 and (4.25) into $\boldsymbol{\alpha}_k^{\#}$, $\boldsymbol{c}_k^{\#}$ and $\boldsymbol{\gamma}_k^{\#}$, we have

$$\begin{aligned}
\|\boldsymbol{\alpha}_k^{\#}\| &\le \Big((m-1)\beta w_2\|\mathcal{A}_m\|_F M^{m-2} + \frac{m\lambda w_2}{m-1}\Big)\|g_{k-1} - g_k\|_{\mathcal{B}^m},\\
\|\boldsymbol{c}_k^{\#}\| &\le m\lambda w_2\|\mathcal{A}_m\|_F M^{m-2}\|g_{k-1} - g_k\|_{\mathcal{B}^m},\\
\|\boldsymbol{\gamma}_k^{\#}\| &\le \frac{m\lambda w_2}{(m-1)\beta}\|g_{k-1} - g_k\|_{\mathcal{B}^m}.
\end{aligned}$$

Thus, it follows that there exists $(\zeta_2)_\beta > 0$ such that

$$\|\boldsymbol{\alpha}_k^{\#}\| + \|\boldsymbol{c}_k^{\#}\| + \|\boldsymbol{\gamma}_k^{\#}\| \le (\zeta_2)_\beta\|g_{k-1} - g_k\|_{\mathcal{B}^m}. \tag{4.26}$$

From (4.21), (4.24) and (4.26), we finally have that

$$r_k - r_{k+1} \le (\zeta_2)_\beta\|g_{k-1} - g_k\|_{\mathcal{B}^m}(\varphi(r_k) - \varphi(r_{k+1})) \quad \text{for } k > k_1.$$

This proof is completed.

Now we prove Theorem 4.1 by Lemmas 4.7–4.8.

Proof of Theorem 4.1 since $\{r_k\}$ is monotonically decreasing and convergent to 0, we consider the following two cases of $\{r_k\}$.

(I) If for k sufficiently large, $r_k = r_{k+1} = 0$, then Lemma 4.7 shows that $g_k = g_{k+1}$. Since $\mathcal{J}_m$ is a homeomorphism, it follows that $s_k = s_{k+1}$. Hence, $\{s_k\}$ is convergent in $\mathcal{B}^{\frac{m}{m-1}}$.

(II) If $r_k > 0$ for any $k \in \mathbb{N}$, then we show that $\{g_k\}$ is a Cauchy sequence in $\mathcal{B}^m$. To this end, for any $j, l \in \mathbb{N}$ where $j < l$, from the triangle inequality, we have that

$$\|g_j - g_l\|_{\mathcal{B}^m} \le \sum_{k=j}^{l-1} \|g_k - g_{k+1}\|_{\mathcal{B}^m} \le \sum_{k=j}^{\infty} \|g_k - g_{k+1}\|_{\mathcal{B}^m}. \tag{4.27}$$

Since $(\zeta_1)_\beta, (\zeta_2)_\beta > 0$, it follows that

$$\begin{aligned} \|g_k - g_{k+1}\|_{\mathcal{B}^m} &\overset{\text{(a)}}{\le} \sqrt{\frac{(\zeta_2)_\beta}{(\zeta_1)_\beta}\|g_{k-1} - g_k\|_{\mathcal{B}^m}(\varphi(r_k) - \varphi(r_{k+1}))} \\ &\overset{\text{(b)}}{\le} \frac{\|g_{k-1} - g_k\|_{\mathcal{B}^m} + \frac{(\zeta_2)_\beta}{(\zeta_1)_\beta}(\varphi(r_k) - \varphi(r_{k+1}))}{2} \quad \text{for } k > k_1 + 1, \end{aligned}$$

where (a) holds because of Lemmas 4.7–4.8 and (b) follows from the fact $\sqrt{ab} \le \frac{a+b}{2}$, $a, b \in (0, \infty)$. Thus, we see that

$$\|g_k - g_{k+1}\|_{\mathcal{B}^m} \le \|g_{k-1} - g_k\|_{\mathcal{B}^m} - \|g_k - g_{k+1}\|_{\mathcal{B}^m} + \frac{(\zeta_2)_\beta}{(\zeta_1)_\beta}(\varphi(r_k) - \varphi(r_{k+1})) \quad \text{for } k > k_1 + 1.$$

Summing up the above relation from $k = k_1 + 1$ to ∞, we obtain that

$$\begin{aligned} &\sum_{k\in\mathbb{N}} \|g_k - g_{k+1}\|_{\mathcal{B}^m} \\ &= \sum_{k=1}^{k_1} \|g_k - g_{k+1}\|_{\mathcal{B}^m} + \sum_{k=k_1+1}^{\infty} \|g_k - g_{k+1}\|_{\mathcal{B}^m} \\ &\le \sum_{k=1}^{k_1} \|g_k - g_{k+1}\|_{\mathcal{B}^m} + \|g_{k_1} - g_{k_1+1}\|_{\mathcal{B}^m} + \frac{(\zeta_2)_\beta}{(\zeta_1)_\beta}\varphi(r_{k_1+1}) \\ &< \infty. \end{aligned} \tag{4.28}$$

Hence

$$\lim_{j\to\infty} \sum_{k=j}^{\infty} \|g_k - g_{k+1}\|_{\mathcal{B}^m} = 0. \tag{4.29}$$

Combining (4.27) with (4.29), it follows that

$$\lim_{j\to\infty} \|g_j - g_l\|_{\mathcal{B}^m} = 0,$$

that is, $\{g_k\}$ is a Cauchy sequence in $\mathcal{B}^m$. Since $\mathcal{B}^m$ is a Banach space, the convergence of $\{g_k\}$ follows immediately from this. Since $\mathcal{J}_m$ is a homeomorphism, the convergence of $\{g_k\}$ implies the convergence of $\{s_k\}$.

Combining (I) with (II), we conclude that $\{s_k\}$ is convergent in $\mathcal{B}^{\frac{m}{m-1}}$, that is, there exists $s_* \in \mathcal{B}^{\frac{m}{m-1}}$ such that

$$\lim_{k\to\infty} \|s_k - s_*\|_{\mathcal{B}^{\frac{m}{m-1}}} = 0.$$

Now we show that s_* is a stationary point of optimization problem (1.1) by (2.5).

First, we discuss where s_* is located. Since $\mathcal{J}_m$ is a homeomorphism from $\mathcal{B}^m$ onto $\mathcal{B}^{\frac{m}{m-1}}$ and $\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}$ is a closed subset in $\mathcal{B}^m$ by [18, Corollary 1.4.20], $\mathcal{J}_m(\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\})$ is closed in $\mathcal{B}^{\frac{m}{m-1}}$. Since $\{s_k\}$ converges to s_* in $\mathcal{B}^{\frac{m}{m-1}}$ and $\{s_k\}$ is contained in $\mathcal{J}_m(\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\})$, it follows that

$$s_* \in \mathcal{J}_m(\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}).$$

Thus, there exists a unique $\boldsymbol{c}_* \in \mathbb{R}^N$ such that

$$\frac{m\lambda}{m-1}(\mathcal{J}_m)^{-1}(s_*) = (K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot))\Big(\frac{m\lambda}{m-1}\boldsymbol{c}_*\Big). \tag{4.30}$$

Next, we use the definition of limiting subdifferential of F at $\boldsymbol{\delta}(s_*)$ to finish the proof. By the definition of $\mathcal{L}_\beta$, we have that

$$\begin{aligned} F(\boldsymbol{\alpha}_k) = \mathcal{L}_\beta(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k) - \lambda\mathcal{A}_m(\boldsymbol{c}_k)^m - (\boldsymbol{\gamma}_k)^{\mathrm{T}}(\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}) \\ - \frac{\beta}{2}\|\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_k)^{m-1}\|^2. \end{aligned} \tag{4.31}$$

Since $\mathcal{J}_m$ is a homeomorphism from $\mathcal{B}^m$ onto $\mathcal{B}^{\frac{m}{m-1}}$ and $\{s_k\}$ converges to s_* in $\mathcal{B}^{\frac{m}{m-1}}$, we see that $\{g_k\}$ converges to $(\mathcal{J}_m)^{-1}(s_*)$. Moreover, since $g \mapsto \boldsymbol{c}_g$ is an isomorphism from $\mathrm{span}\{K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot)\}$ onto $\mathbb{R}^N$ by Lemma 4.1, (2.10), (S-3) and (S-3′) show that the sequence $\{(\boldsymbol{\alpha}_k,\boldsymbol{c}_k,\boldsymbol{\gamma}_k)\}$ is also convergent and

$$\lim_{k\to\infty}\boldsymbol{c}_k = \boldsymbol{c}_*, \quad \lim_{k\to\infty}\boldsymbol{\gamma}_k = \frac{m\lambda}{m-1}\boldsymbol{c}_*, \quad \lim_{k\to\infty}\boldsymbol{\alpha}_k = \mathcal{A}_m(\boldsymbol{c}_*)^{m-1} = \boldsymbol{\delta}(s_*). \tag{4.32}$$

Thus, from the continuity of $\boldsymbol{c}\mapsto\lambda\mathcal{A}_m\boldsymbol{c}^m$, (4.19) and (4.31)–(4.32), we show that

$$\lim_{k\to\infty}F(\boldsymbol{\alpha}_k) = F(\boldsymbol{\delta}(s_*)). \tag{4.33}$$

Moreover, from the optimality condition of (S-1), it is easy to check that

$$-\beta\Big(\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_{k-1})^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_{k-1}\Big) \in \partial F(\boldsymbol{\alpha}_k) \tag{4.34}$$

and

$$\lim_{k\to\infty} -\beta\Big(\boldsymbol{\alpha}_k - \mathcal{A}_m(\boldsymbol{c}_{k-1})^{m-1} + \frac{1}{\beta}\boldsymbol{\gamma}_{k-1}\Big) = -\frac{m\lambda}{m-1}\boldsymbol{c}_*. \tag{4.35}$$

In conclusion, by the definition of limiting subdifferential of F at $\boldsymbol{\delta}(s_*)$, (4.32)–(4.35), it follows that

$$-\frac{m\lambda}{m-1}\boldsymbol{c}_* \in \partial F(\boldsymbol{\delta}(s_*)). \tag{4.36}$$

From (2.5), (4.30) and (4.36), we finally have that

$$0 \in (K(\boldsymbol{x}_1,\cdot),\cdots,K(\boldsymbol{x}_N,\cdot))\partial F(\boldsymbol{\delta}(s_*)) + \frac{m\lambda}{m-1}(\mathcal{J}_m)^{-1}(s_*) \subseteq \partial\mathcal{T}_{\frac{m}{m-1}}(s_*),$$

that is, s_* is a stationary point of optimization problem (1.1). This proof is completed.

In this section, leveraging Kurdyka-Lojasiewicz property of $\mathcal{L}_\beta$, we finish the convergence analysis of $\{s_k\}$. In the next section, we give several numerical examples to illustrate the effectiveness of Algorithm 1.

5 Numerical Examples

In this section, we test Algorithm 1 by the synthetic data and the real data for binary classification. We choose some training data and testing data, loss functions and RKBSs to test Algorithm 1. For simplicity, let K_1 be the Gaussian kernel, that is,

$$K_1(\boldsymbol{x},\boldsymbol{x}') = \mathrm{e}^{-\sigma^2\|\boldsymbol{x}-\boldsymbol{x}'\|^2} = \sum_{\boldsymbol{n}\in(\mathbb{N}_0)^d} \phi_{\boldsymbol{n}}(\boldsymbol{x})\phi_{\boldsymbol{n}}(\boldsymbol{x}'), \quad \sigma>0 \text{ for } \boldsymbol{x},\boldsymbol{x}'\in X,$$

where $\phi_{\boldsymbol{n}}(\boldsymbol{x}) = \prod_{j=1}^{d} \big(\frac{2^{n_j}}{(n_j)!}\big)^{\frac{1}{2}}(\sigma x_j)^{n_j}\mathrm{e}^{-\sigma^2(x_j)^2}$ for $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)^{\mathrm{T}}\in X$ and $(\mathbb{N}_0)^d$ is the tensor product of natural numbers. Also, let K_2 be the power series kernel, that is,

$$K_2(\boldsymbol{x},\boldsymbol{x}') = \sum_{\boldsymbol{n}\in(\mathbb{N}_0)^d} \phi_{\boldsymbol{n}}(\boldsymbol{x})\phi_{\boldsymbol{n}}(\boldsymbol{x}'), \quad \forall \boldsymbol{x},\boldsymbol{x}'\in X,$$

where $\phi_{\boldsymbol{n}}(\boldsymbol{x}) = \prod_{j=1}^{d} \frac{1}{(n_j)!}(x_j)^{n_j}$, $\forall \boldsymbol{x}=(x_1,x_2,\cdots,x_d)^{\mathrm{T}}\in X$. For any even integer $m\geq 4$, we construct $\mathcal{B}^{\frac{m}{m-1}}$ by the kernel K_1 or K_2 as we mentioned in Section 2. Specially, in [19], we show that the RKHS can be constructed by the kernel K_1 or K_2, and these RKHSs can be seen as $\mathcal{B}^2$. Also, we discuss the splitting method for the SVM in $\mathcal{B}^2$ in [19]. We will compare with the performance of the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ when $m\geq 4$ and the SVM in $\mathcal{B}^2$ with the same loss function.

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

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

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

We see that L_1 is a convex Hinge loss, L_2 is a convex squared Hinge loss, L_3 is a nonconvex piecewise logarithmic loss function and L_4 is a nonconvex linear piecewise loss function. All of these loss functions satisfy Assumption 4.1 (i)–(ii).

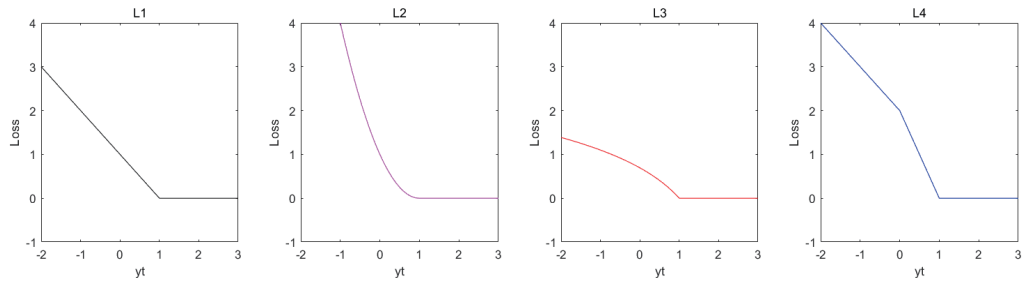


Figure 1 The loss functions L_1–L_4. All are shown as a function of yt rather than t, because of the symmetry between the $y=+1$ and $y=-1$ case.

Now we introduce some numerical techniques of Algorithm 1. It is well-known that storing a high-dimensional tensor and then performing numerical tensor operation is unrealistic. By the definitions of $\mathcal{A}_m\boldsymbol{c}^{m-1}$ and $\mathcal{A}_m\boldsymbol{c}^{m-2}$, we truncate them by $\mathcal{M}$ terms, that is,

$$\mathcal{A}_m\boldsymbol{c}^{m-1} \approx \sum_{n=1}^{\mathcal{M}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-1}\boldsymbol{\Phi}_n \in \mathbb{R}^N, \quad \mathcal{A}_m\boldsymbol{c}^{m-2} \approx \sum_{n=1}^{\mathcal{M}}(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{c})^{m-2}\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}} \in \mathbb{R}^{N\times N}.$$

In other words, we use the sum of vectors and matrices to approximate tensor operations. Moreover, we can show the convergence of the sequence $\{s_k\}$ by (4.28). By the definition of $\{g_k\}$, it follows that $\sum\limits_{k\in\mathbb{N}}\|g_k-g_{k+1}\|_{\mathcal{B}^m} = \sum\limits_{k\in\mathbb{N}}\big(\sum\limits_{n\in\mathbb{N}}|\boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{c}_k-\boldsymbol{c}_{k+1})|^m\big)^{\frac{1}{m}}$. Similarly, we truncate it by $\mathcal{M}$ terms, that is,

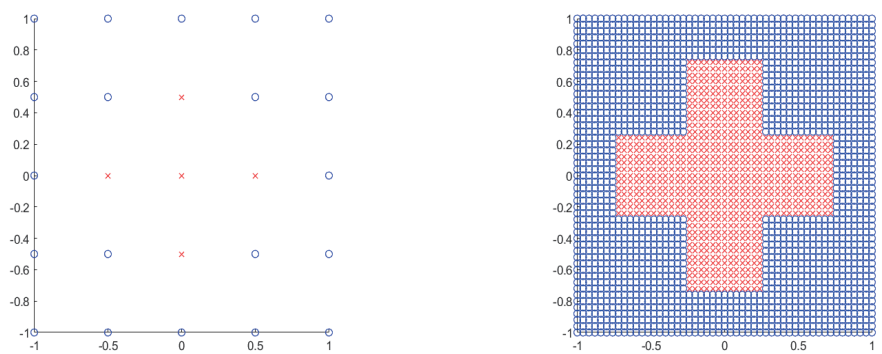
$$\sum_{k\in\mathbb{N}}\|g_k-g_{k+1}\|_{\mathcal{B}^m} \approx \sum_{k\in\mathbb{N}}\Big(\sum_{n=1}^{\mathcal{M}}|\boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{c}_k-\boldsymbol{c}_{k+1})|^m\Big)^{\frac{1}{m}}.$$

On the other hand, we set the terminal criterion as for a given $\varepsilon_0>0$, if $\|\boldsymbol{\alpha}_{k+1}-\mathcal{A}_m(\boldsymbol{c}_{k+1})^{m-1}\|<\varepsilon_0$, then stop. We take s_{k+1} as the output. Since $\{s_k\}$ is globally convergent to a stationary point of optimization problem (1.1), we solve optimization problem (1.1) repeatedly by selecting some initial values randomly and choosing the minimizer of these outputs as the approximate solution $s_D: X\to\mathbb{R}$.

Next, we introduce our numerical results on synthetic data and real data.

5.1 Examples on synthetic data

In this subsection, we introduce our test results on the synthetic data. We use the training set D_{11} with 25 points and the testing set D_{12} with 2601 points to show the effectiveness of Algorithm 1.



(a) Training Set D_{11} (b) Testing Set D_{12}

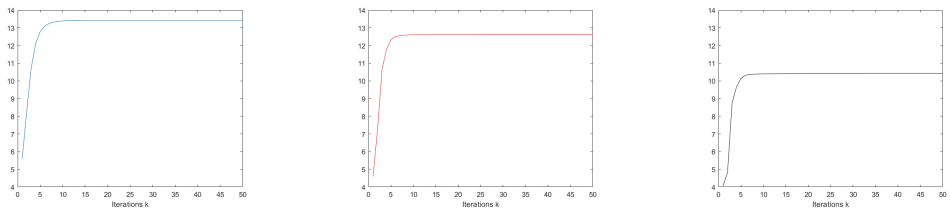
Figure 2 The binary classification $X_1=[-1,1]\times[-1,1]$ and $Y=\{+1,-1\}$: The classes are coded as a binary variable (blue=+1 and red=−1). The left panel represents the training data D_{11} and the right panel represents the testing data D_{12}.

First, we show the convergence of Algorithm 1. The parameters are given below.

- Gaussian kernel K_1, where $\sigma=3.36$.

- Loss function L_3.
- The RKBS $\mathcal{B}^2$, $\mathcal{B}^{\frac{4}{3}}$ and $\mathcal{B}^{\frac{6}{5}}$.
- $N = 25$, $\lambda = 0.01$, $\beta = 1$ and $\varepsilon_0 = 10^{-8}$.
- 20 initial values randomly chosen in $[-1, 1]^N$.

We now conduct experiments to verify the convergence of Algorithm 1.



(a) $m = 2$ (b) $m = 4$ (c) $m = 6$

Figure 3 The convergence of Algorithm 1 in different RKBSs with L_3. The horizontal axis represents iterations and the vertical axis represents $\sum\limits_{k\in\mathbb{N}} \|g_k - g_{k+1}\|_{\mathcal{B}^m}$.

As shown in Figure 3, for the training data D_{11}, loss function L_3 and other parameters above, Algorithm 1 converges in less than 50 iterations. These numerical results show that Algorithm 1 is efficient and stable.

Next, we use loss functions L_1–L_4 and kernel K_1 mentioned above to test Algorithm 1 and obtain the approximate solution s_D. We use s_D to build the following SVM,

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

to predict testing data. In each experiment, we select the training data, a loss function, an RKBS and some other parameters above. Then we build the corresponding SVM to predict the testing data. Here are the results of these experiments.

Table 1 Different testing accuracy on testing set D_{12} with kernel K_1.

RKBS \ L	L_1	L_2	L_3	L_4
$\mathcal{B}^2$	94.31%	94.31%	96.16%	84.16%
$\mathcal{B}^{\frac{4}{3}}$	92.96%	94.00%	96.77%	69.51%
$\mathcal{B}^{\frac{6}{5}}$	93.54%	94.31%	97.85%	69.51%

From Algorithm 5.1, it shows that the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ with a lower semi-continuous loss function by Algorithm 1 is feasible in terms of accuracy. Moreover, it is easy to see that for this training data D_{11} and testing data D_{12}, the SVM in $\mathcal{B}^{\frac{6}{5}}$ with nonconvex loss function L_3 and the kernel K_1 performs better than other cases shown in Table 1. Next, we introduce our experiments on real data.

5.2 Examples on UCI machine learning repository

In this subsection, we use the Parkinson's disease classification dataset in the UCI machine learning repository. There are 752 features about them. We try to train a model to determine whether the patient has Parkinson's disease. Here is the detail of these data (see Table 2).

Table 2 The detail of training data and testing data.

	Positive	Negative	Total
Training Set D_{21}	389	111	500
Testing Set D_{22}	175	81	256
Label	$+1$	-1	

To make features have the same measurement scale, we normalize the data to $[0,1]^{752}$. We use the normalized training data and testing data for binary classification. Next, we introduce some parameters of these experiments as follows.

- Power series kernel K_2.
- Loss functions L_1, L_2, L_3 and L_4.
- The RKBS $\mathcal{B}^2$, $\mathcal{B}^{\frac{4}{3}}$ and $\mathcal{B}^{\frac{6}{5}}$.
- $N = 756$, $\lambda = 10^{-4}$, $\beta = 10^{-4}$ and $\varepsilon_0 = 10^{-8}$.
- 20 initial values randomly chosen in $[0,1]^N$.

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

Table 3 Different testing accuracy of on testing set D_{22} with kernel K_2.

RKBS \ Loss	L_1	L_2	L_3	L_4
$\mathcal{B}^2$	71.09%	74.22%	77.34%	73.44%
$\mathcal{B}^{\frac{4}{3}}$	68.75%	55.86%	80.86%	82.31%
$\mathcal{B}^{\frac{6}{5}}$	68.75%	64.84%	80.47%	77.34%

From Table 3, we check that the SVM in $\mathcal{B}^{\frac{4}{3}}$ with nonconvex loss function L_4 and the kernel K_2 performs better than others in these experiments. It shows that for this training data D_{21} and testing data D_{22}, nonconvex and lower semi-continuous loss function and general RKBS are more suitable than the convex loss function and RKHS, which is our motivation for this paper.

In Section 5, we demonstrate the effectiveness of solving the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ with a lower semi-continuous loss function by splitting method based on ADMM. In addition, we give some examples to show that the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ with a nonconvex and lower semi-continuous loss function performs better than the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ with a convex loss function. Also, we show that the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ performs better than the SVM in $\mathcal{B}^2$ with a lower semi-continuous loss function.

6 Final Remarks

In [16], the second author and the third author propose several numerical tricks in RKBS and discuss the homotopy method for the multikernel-based approximation method. As a continuation of the program, in this paper, we discuss the splitting method based on ADMM for the SVM in $\mathcal{B}^{\frac{m}{m-1}}$ with a lower semi-continuous loss function. Since $\mathcal{B}^p$ $(1 \leq p < \infty)$ are also RKBSs, we will study how to solve the SVM in $\mathcal{B}^p$ $(1 \leq p < \infty)$ with a lower semi-continuous loss function.

Acknowledgements The authors would like to thank the handling editor and the referee for their detailed comments.

Declarations

Conflicts of interest The authors declare no conflicts of interest.

References

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

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

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

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

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

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

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

[8] Cioranescu, I., Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems, Kluwer Academic Publishers, Dordrecht, 1990.

[9] Cortes, C. and Vapnik, C., Support vector network, *Mach. Learn.*, **20**, 1995, 273–297.

[10] Dal Maso, G., An Introduction to Γ-Convergence, Springer Science, Business Media, New York, 1993.

[11] Facchinei, F. and Pang, J. S., Finite-dimensional Variational Inequalities and Complementarity Problems, **I**, Springer-Verlag, Berlin, 2003.

[12] Feng, Y. L., Yang, Y. N., Huang, X., et al., Robust support vector machines for classification with nonconvex and smooth losses, *Neural Comput.*, **28**, 2016, 1217–1247.

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

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

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

[16] Lin, Y., Wei, Y. M. and Ye, Q., A homotopy method for multikernel-based approximation, *J. Nonlinear Var. Anal.*, **6**, 2022, 139–154.

[17] Liu, D. L., Shi, Y., Tian, Y. J. and Huang, X. K., Ramp loss least squares support vector machine, *J. Comput. Sci.*, **14**, 2016, 61–68.

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

[19] Mo, M. Y. and Ye, Q., Splitting method for support vector machines with lower semi-continuous loss, *Pac. J. Optim*, **19**(4), 2023, 689–714.

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

[21] Ortega, J. and Rheinboldt, W., Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, inc, San Diego, 1970.

[22] Penot, J., Calculus Without Derivatives, Springer-Verlag, New York, 2013.

[23] Qi, L. Q., Chen, H. B. and Chen, Y. N., Tensor Eigenvalues and Their Applications, Springer-Verlag, Singapore, 2018.

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

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

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

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

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

[29] Unser, M., A unifying representer theorem for inverse problems and machine learning, *Found. Comput. Math.*, **21**, 2021, 941–960.

[30] Wang, H. J., Shao, Y. H., Zhou, S. L., et al., Support vector machine classifier via $L_{0/1}$ soft-margin loss, *IEEE T. Pattern Anal.*, **44**, 2022, 7253–7265.

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

[32] Xu, Y. Y. and Yin, W. T., 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.

[33] Ye, Q., Positive definite multi-kernels for scattered data interpolations, *Appl. Comput. Harmon. A*, **62**, 2023, 251–260.

[34] Zalinescu, C., Convex Analysis in General Vector Spaces, World Scientific Publishing, River Edge, 2002.