Computing D-Stationary Points of ρ-Margin Loss SVM

Lai Tian
The Chinese University of Hong Kong
tianlai@se.cuhk.edu.hk

Anthony Man-Cho So
The Chinese University of Hong Kong
manchoso@se.cuhk.edu.hk

Abstract

This paper is concerned with the algorithmic aspects of sharper stationarity of a nonconvex, nonsmooth, Clarke irregular machine learning model. We study the SVM problem with a ρ-margin loss function, which is the margin theory generalization bound of SVM introduced in the learning theory textbook by Mohri et al. [2018], and has been extensively studied in operations research, statistics, and machine learning communities. However, due to its nonconvex, nonsmooth, and irregular nature, none of the existing optimization methods can efficiently compute a d(irectional)-stationary point, which turns out to be also a local minimum, for the ρ-margin loss SVM problem. After a detailed discussion of various nonsmooth stationarity notions, we propose a highly efficient nonconvex semi-proximal ADMM-based scheme that provably computes d-stationary points and enjoys a local linear convergence rate. We report concrete examples to demonstrate the necessity of our assumptions. Numerical results verify the effectiveness of the new algorithm and complement our theoretical results.

1 INTRODUCTION

Nonconvex nonsmooth models are ubiquitous in modern statistical and machine learning tasks. Though holding strong expressive power, these "non"-problems pose serious computational challenges as most of the tools from smooth and convex analysis are inapplicable in this "non"-scenario. To handle their algorithmic design and theoretical analysis without sacrificing rigor, we borrow ideas and tools from variational analysis [Rockafellar and Wets, 2009], in which concepts that we are familiar with branch into many and calculus rules that we take for granted may no longer hold. Even choosing a proper first-order stationarity concept for these "non"-problems becomes a highly nontrivial thing (see Definition 2.7, and [Li et al., 2020]). These obstacles could be alleviated if the optimization problem is Clarke regular (see Definition 2.4), which contains smooth, convex, or more generally, weakly convex problems as special cases.

In this paper, we study the SVM model with a ρ-margin loss, which is a nonconvex, nonsmooth, and irregular problem in the sense of Clarke. Specifically, for a given set of data points $\{(\boldsymbol{x}_i, y_i) : i \in [n]\} \subseteq \mathbb{R}^d \times \{+1, -1\}$, we aim to solve

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{c}{2}\|\boldsymbol{\theta}\|^2 + \sum_{i=1}^{n} \phi_\rho\Big(y_i \cdot (\boldsymbol{x}_i^\top \boldsymbol{\theta} + b)\Big), \tag{1.1}$$

where the regularization parameter $c > 0$, and $\phi_\rho : \mathbb{R} \to \mathbb{R}$ is the ρ-margin loss defined by

$$\phi_\rho(u) := \min\left(1, \max\left(0, 1 - \frac{u}{\rho}\right)\right).$$

We notice that Problem (1.1) (particularly its $\rho = 1$ version, which is usually termed ramp loss) has been widely recognized in operations research [Brooks, 2011, Carrizosa et al., 2014, Wang et al., 2021], statistics [Shen et al., 2003, Wu and Liu, 2007, Liu et al., 2005], and machine learning [Huang et al., 2014, Keshet and McAllester, 2011, Collobert et al., 2006b,a, Ertekin et al., 2010, Suzumura et al., 2017, Maibing and Igel, 2015] communities as it provides better robustness against data outliers than the vanilla SVM. The ρ-margin loss (see Figure 1) we used in this paper can also be seen from the learning theory textbook by Mohri et al. [2018, Corollary 5.11], in which Problem (1.1) serves as the ρ-margin generalization bound for linear hypothesis set.

However, due to its "non"-properties, especially the irregularity, Problem (1.1) poses serious computational

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

challenges in the optimization aspect. Indeed, Maibing and Igel [2015] showed that finding a global minimizer for Problem (1.1) is NP-hard. Thus, we must be satisfied with computing a local minimizer instead, which is still a highly nontrivial task even for a quadratic function. Recently, Ahmadi and Zhang [2022] showed that finding a local minimizer is NP-hard generally, while the NP-hardness of detecting local optimality was proven decades ago by Murty and Kabadi [1987].

1.1 Prior Arts

In the literature concerning Problem (1.1), DC (Difference of Convex) Algorithms are probably the most popular strategies [Huang et al., 2014, Keshet and McAllester, 2011, Collobert et al., 2006b, Shen et al., 2003, Wu and Liu, 2007, Liu et al., 2005, Collobert et al., 2006a, Ertekin et al., 2010] as the nonsmooth part ϕ_ρ admits the following DC-decomposition:

$$\phi_\rho(u) = \max\left(1 - \frac{u}{\rho}, 0\right) - \max\left(-\frac{u}{\rho}, 0\right).$$

However, as pointed out by Nouiehed et al. [2019], these DCA-type algorithms only compute so-called DC-critical points (see Definition 2.7), which is a fairly weak notion of stationarity (see Definition 2.7 and remark afterwards) and depend on the DC-decomposition rather than variational properties of the problem. Indeed, it is easy to construct an example that is DC-critical but not stationary in any conventional sense (see Figure 2). By considering a polyhedral reformulation of Problem (1.1), Suzumura et al. [2017] proposed a homotopy-type algorithmic framework. However, in Section 4.3, we will show that their argument cannot guarantee sharper stationarity convergence. Recently, Wang et al. [2021] reported optimality conditions for Problem (1.1) (specialized to $\rho = 1$) with so-called P-stationarity and carefully computed the closed-form solution of the ramp loss proximal mapping. However, no known algorithm can efficiently compute points satisfying such conditions.

On the other front, for max-structured DC programs, Pang et al. [2017] and Cui et al. [2018] proposed enhanced DCA schemes for computing d-stationary points (see Definition 2.7), which is arguably the sharpest stationarity type for certain structured DC programs [Nouiehed et al., 2019]. Nevertheless, as mentioned by Pang et al. [2017, Section 7], these enhanced DCAs require solving an exponentially large number of subproblems in a single iteration step, which is not even suitable for hundreds of data points. Then, here comes the main question that this paper aims to answer:

Can we compute d-stationary points efficiently for SVM with ρ-margin loss?

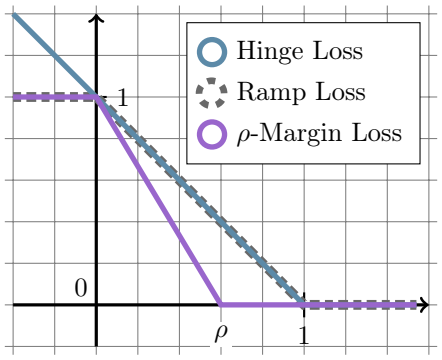

Figure 1: Hinge loss, ramp loss, and ρ-margin loss.

1.2 Contributions

We highlight the main contributions of this paper as follows:

- We propose a highly efficient *nonconvex* semi-proximal ADMM procedure, which provably computes a d-stationary point (which is also local minimal, see Proposition 3.5), of Problem (1.1).
- We show that the distance between the nonconvex ADMM generated sequence and the set of d-stationary points enjoys a local linear convergence rate to zero.
- We provide a detailed discussion on existing computing strategies and report concrete examples to demonstrate their incapability in sharp stationarity computing.

Notations. Most of the notations used in this paper are standard. Throughout this paper, scalars, vectors, and matrices are denoted by lowercase letters, boldface lowercase letters, and boldface uppercase letters, respectively; $\|\boldsymbol{x}\|$ is the Euclidean norm of $\boldsymbol{x}$; $\|\boldsymbol{A}\|$ is the operator norm of $\boldsymbol{A}$; for a set $S \subseteq \mathbb{R}^n$ and point $\boldsymbol{x}$, $\operatorname{dist}(\boldsymbol{x}, S) := \inf_{\boldsymbol{v}\in S} \|\boldsymbol{v} - \boldsymbol{x}\|$; $A \oplus B$ denotes the direct sum of A and B; $\boldsymbol{A} \odot \boldsymbol{B}$ is the Hadamard product of $\boldsymbol{A}$ and $\boldsymbol{B}$; we use $\operatorname{prox} : \mathbb{R}^n \to \mathbb{R}^n$ for the proximal operator, and use $\operatorname{Prox} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ for set-valued proximal mapping (similarly, arg and Arg); $\operatorname{cl} S$ denotes the closure of set S; $\mathbb{B}_\epsilon(\boldsymbol{x}) := \{\boldsymbol{v} : \|\boldsymbol{v} - \boldsymbol{x}\| \le \epsilon\}$; $\operatorname{Co} S$ is the convex hull of set S; $\sigma_i(\boldsymbol{A})$ is the i-th largest singular value; $\boldsymbol{e}_i$ is the i-th column of identity matrix.

2 PRELIMINARIES

In this section, we will introduce the necessary background on variational analysis for our development. As the problem we are interested in is nonconvex and nonsmooth, conventional tools from smooth or convex analysis will be inapplicable.

Table 1: Algorithms for Problem (1.1), stationarity, computational complexity, and convergence rate.

Category	Algorithms	Stationarity	Per-Iter. Compl.	Rate
DCA (incl. CCCP)	[Huang et al., 2014] [Wu and Liu, 2007] [Brooks, 2011]	DC-critical	poly(n,d)	—
Nonmonotone MM	[Cui et al., 2018]	d-stationary	$O(2^n)$	—
Subgradient Method	[Davis et al., 2020] [Majewski et al., 2018]	C-stationary	poly(n,d)	—
Homotopy Algorithm	[Suzumura et al., 2014, 2017]	—[1]	poly(n,d)	—
Semi-proximal ADMM	Algorithm 1	d-stationary	poly(n,d)	local linear

[1] See Section 4.3.

2.1 Generalized Differentiation Theory

To tell the subtle difference between various stationary concepts and appreciate the benefit of computing sharper stationary points, we recall several notions from nonsmooth analysis in this section. We will confine the discussion to locally Lipschitz and directional differentiable functions as our Problem (3.1) is exactly in that class. To begin, let us recall the directional derivative and Clarke generalized subderivative of a locally Lipschitz functon $f:\mathbb{R}^n\to\mathbb{R}$, which will play fundamental roles in the analyses and definitions of other notions.

Definition 2.1 (directional derivative)**.** *Given a point $\boldsymbol{x}$ and direction $\boldsymbol{d}$, the directional derivative of f at $\boldsymbol{x}$ in the direction $\boldsymbol{d}$ is defined by*

$$f'(\boldsymbol{x};\boldsymbol{d}) \coloneqq \lim_{t\searrow 0}\frac{f(\boldsymbol{x}+t\boldsymbol{d})-f(\boldsymbol{x})}{t}.$$

Definition 2.2 (Clarke subderivative)**.** *Given a point $\boldsymbol{x}$ and direction $\boldsymbol{d}$, the Clarke directional derivative of f at $\boldsymbol{x}$ in the direction $\boldsymbol{d}$ is defined by*

$$f^\circ(\boldsymbol{x};\boldsymbol{d}) \coloneqq \limsup_{\substack{\boldsymbol{x}'\to\boldsymbol{x}\\ t\searrow 0}}\frac{f(\boldsymbol{x}'+t\boldsymbol{d})-f(\boldsymbol{x}')}{t}.$$

Using Rademacher's Theorem [Rockafellar and Wets, 2009, Theorem 9.60], the Clarke subdifferential of a locally Lipschitz function can be defined as follows [Rockafellar and Wets, 2009, Theorem 9.61]:

Definition 2.3 (Clarke subdifferential)**.** *Given a point $\boldsymbol{x}$, the Clarke subdifferential of f at $\boldsymbol{x}$ is defined by*

$$\partial_C f(\boldsymbol{x}) \coloneqq \mathrm{Co}\left\{\boldsymbol{s}:\exists \boldsymbol{x}'\to\boldsymbol{x}, \nabla f(\boldsymbol{x}') \text{ exists}, \nabla f(\boldsymbol{x}')\to\boldsymbol{s}\right\}.$$

We remark that $\partial_C f(\boldsymbol{x})$ is convex and can be equivalently defined as the closed convex set whose support function is $f^\circ(\boldsymbol{x};\cdot)$. Now, we introduce the notion of Clarke regularity with Clarke subderivative.

Definition 2.4 (Clarke regularity [Clarke, 1990, Definition 2.3.4])**.** *A locally Lipschitz function $f:\mathbb{R}^n\to\mathbb{R}$ is regular at $\boldsymbol{x}\in\mathbb{R}^n$ if for every $\boldsymbol{d}\in\mathbb{R}^n$, the ordinary directional derivative $f'(\boldsymbol{x};\boldsymbol{d})$ exists and*

$$f^\circ(\boldsymbol{x};\boldsymbol{d}) = f'(\boldsymbol{x};\boldsymbol{d}).$$

An important implication of regularity is the validity of various subdifferential calculus rules, which allows sharper characterization of first-order stationarity [Rockafellar and Wets, 2009, Theorem 10.6, Corollary 10.9]; see [Li et al., 2020] for a quick overview.

To characterize the local behavior of f from a variational perspective, we need the notion of Fréchet subdifferential [Rockafellar and Wets, 2009, Exercise 8.4], which has a close relation with d-stationarity (see Definition 2.7 and Proposition 3.2) and is defined as follows:

Definition 2.5 (Fréchet subdifferential)**.** *Given a point $\boldsymbol{x}$, the Fréchet subdifferential of a locally Lipschitz and directional differentiable function f at $\boldsymbol{x}$ is defined by*

$$\widehat{\partial} f(\boldsymbol{x}) \coloneqq \left\{\boldsymbol{s}:\boldsymbol{s}^\top\boldsymbol{d}\le f'(\boldsymbol{x};\boldsymbol{d}) \text{ for all } \boldsymbol{d}\right\}.$$

Though enjoying great variational properties, Fréchet subdifferential is not a robust notion as its graph is not closed (or, $\widehat{\partial} f$ is not outer semi-continuous, see Definition 2.9). Therefore, the following limiting version generalized subdifferential [Rockafellar and Wets, 2009, Theorem 8.3(b)] is sometimes more convenient:

Definition 2.6 (limiting subdifferential)**.** *Given a point $\boldsymbol{x}$, the limiting subdifferential of a locally Lipschitz and directional differentiable function f at $\boldsymbol{x}$ is defined by*

$$\partial f(\boldsymbol{x}) \coloneqq \limsup_{\boldsymbol{x}'\to\boldsymbol{x}} \widehat{\partial} f(\boldsymbol{x}'),$$

where the outer limit is taken in the set-valued mapping sense (see [Rockafellar and Wets, 2009, 5(1)]).

2.2 Stationarity Notions

Now, let us introduce various stationarity notions for a nonconvex nonsmooth problem.

Definition 2.7. *For a locally Lipschitz and directionally differentiable DC function $f = h - g$, a point $\bar{\boldsymbol{x}}$ is*

(a) ***local minimal*** *if there exists $\epsilon > 0$:*

$$f(\boldsymbol{y}) \geq f(\bar{\boldsymbol{x}}), \quad \forall \boldsymbol{y} \in \mathbb{B}_\epsilon(\bar{\boldsymbol{x}});$$

(b) ***d(irectional)-stationary*** *if:*

$$f'(\bar{\boldsymbol{x}}; \boldsymbol{d}) \geq 0, \ \forall \boldsymbol{d} \in \mathbb{R}^n;$$

(c) ***l(imiting)-stationary*** *if:*

$$0 \in \partial f(\bar{\boldsymbol{x}});$$

(d) ***C(larke)-stationary*** *if:*

$$0 \in \partial_C f(\bar{\boldsymbol{x}}) \quad \Leftrightarrow \quad f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{d}) \geq 0, \ \forall \boldsymbol{d} \in \mathbb{R}^n;$$

(e) ***DC-critical*** *if:*

$$\partial h(\bar{\boldsymbol{x}}) \cap \partial g(\bar{\boldsymbol{x}}) \neq \emptyset.$$

By [Rockafellar and Wets, 2009, Theorem 8.6, 8.15, 8.49], the following relation holds (see also [Li et al., 2020]):

$$\{(a)\} \subseteq \{(b)\} \subseteq \{(c)\} \subseteq \{(d)\} \subseteq \{(e)\}.$$

If the problem is Clarke regular, we have the following tighter relation:

$$\{(a)\} \subseteq \{(b)\} = \{(c)\} = \{(d)\} \subseteq \{(e)\}.$$

Remark 2.8. *Note that the DC-criticality in Definition 2.7 is actually dependent on the DC-decomposition $f = h - g$ of the function. How to find the best DC-decomposition is still an important open problem [Migdalas and Pardalos, 2018]. An illustration of various stationarity notions is provided in Figure 2. See its caption for more information. Besides, it is notable here that for Problem (1.1), we actually have the sharper stationarity relation $\{(a)\} = \{(b)\}$, see Proposition 3.5.*

2.3 Other Useful Notions

Below we record some other notions that will be useful for our reference and development. These definitions are taken from [Rockafellar and Wets, 2009, Definition 5.4] and [Rockafellar and Wets, 2009, Definition 1.23].

Definition 2.9 (outer semi-continuity)**.** *A set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is outer semi-continuous (osc) at $\bar{\boldsymbol{x}}$ if*

$$\limsup_{\boldsymbol{x} \to \bar{\boldsymbol{x}}} S(\boldsymbol{x}) \subset S(\bar{\boldsymbol{x}}).$$

Definition 2.10 (prox-boundedness)**.** *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* prox-bounded *if there exists $\lambda > 0$ such that $f\big(\mathrm{Prox}_{\lambda f}(\boldsymbol{x})\big) > -\infty$ for some $\boldsymbol{x} \in \mathbb{R}^n$.*

3 MAIN RESULTS

In this section, we reformulate Problem (1.1) into a compact form and consider the computation of d-stationary points (see Definition 2.3) of the following nonconvex, nonsmooth, and irregular problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} F(\boldsymbol{w}) := f(\boldsymbol{w}) + g(\boldsymbol{Z}\boldsymbol{w}), \tag{3.1}$$

where we define $\boldsymbol{w} = \begin{bmatrix} \boldsymbol{\theta} \\ b \end{bmatrix} \in \mathbb{R}^{d+1}$,

$$f(\boldsymbol{w}) := \frac{c}{2}\|\boldsymbol{w}\|_{\boldsymbol{K}}^2 = \frac{c}{2}\boldsymbol{w}^\top \boldsymbol{K}\boldsymbol{w}, \quad g(\boldsymbol{q}) := \sum_{i=1}^n \phi_\rho(q_i),$$

and the i-th row of $\boldsymbol{Z} \in \mathbb{R}^{n \times (d+1)}$ as

$$\boldsymbol{z}_i = \begin{bmatrix} y_i \boldsymbol{x}_i \\ y_i \end{bmatrix} \in \mathbb{R}^{d+1}, \boldsymbol{K} = \begin{bmatrix} \mathbb{I}_{d \times d} & \\ & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

3.1 Algorithm: Semi-proximal ADMM

To begin, we make the following surjectivity assumption on $\boldsymbol{Z}$, which is crucial for the dual characterization of d-stationarity and convergence of the ADMM scheme. We will elaborate on that assumption in Section 4.4.

Assumption 3.1 (surjective)**.** *The matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times (d+1)}$ in Problem (3.1) is surjective and $\sigma := \sigma_n(\boldsymbol{Z}) > 0$.*

The following characterization of a d-stationary point is convenient for both analysis and computation.

Proposition 3.2 (dual d-stationarity characterization)**.** *If Assumption 3.1 holds, then a point $\boldsymbol{w}$ is d-stationary of F if and only if*

$$0 \in \nabla f(\boldsymbol{w}) + \boldsymbol{A}^\top \widehat{\partial} g(\boldsymbol{A}\boldsymbol{w}).$$

Proof. By [Rockafellar and Wets, 2009, Exercise 8.4], $\boldsymbol{w}$ is d-stationary if and only if $0 \in \widehat{\partial} F(\boldsymbol{w})$. Then, the relation is immediate from [Rockafellar and Wets, 2009, Corollary 8.8(c), Exercise 10.7]. □

The main contribution of this section is algorithmic ideas that provably computes a point $\boldsymbol{w}$ satisfying the dual characterization of d-stationarity (see Proposition 3.2). Before the new development, we highlight the key difficulties below:

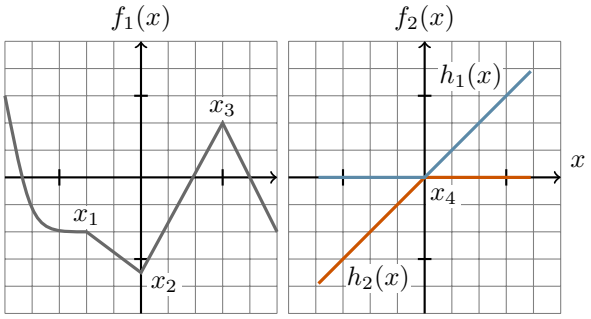

Figure 2: Illustration of various stationarities. In the left subfigure, {local min.} = $\{x_2\}$, {d-stat.} = $\{x_2\}$, {l-stat.} = $\{x_1, x_2\}$, and {C-stat.} = $\{x_1, x_2, x_3\}$. In the right subfigure, we consider a simple DC function $f_2(x) \coloneqq h_1(x) + h_2(x) = \max\{x, 0\} - \max\{-x, 0\} = x$. Then, $x_4 = 0$ is DC-critical while it is obviously not stationary in any conventional (d-, l-, or even C-) sense.

- The first difficulty is the irregularity of the term g in the sense of Clarke. The lack of Clarke regularity may lead to invalidation of subdifferential calculus rules and thus difficulties in computing and characterizing sharper stationarity.
- The second difficulty, which is a consequence of the first one but specific to d-stationarity, is the lack of outer semi-continuity of the Fréchet subdifferential mapping. In other words, the graph of the set-valued mapping $\widehat{\partial} F(\cdot)$ is not closed, which implies that even if we generate a sequence satisfying $\boldsymbol{w}^k \to \boldsymbol{w}^*, \boldsymbol{v}^k \in \widehat{\partial} F(\boldsymbol{w}^k), \boldsymbol{v}^k \to 0$, we cannot say $0 \in \widehat{\partial} F(\boldsymbol{w}^*)$, which invalidates the conventional convergence analysis that relies on the outer semi-continuity of a subdifferential mapping. An illustration of that issue with Figure 2 is that considering $x^k \nearrow x_1$, we have $v^k \in \widehat{\partial} f(x^k), v^k \to 0$ as $k \to +\infty$. So $0 \in \partial f(x_1)$. However, it holds that $0 \notin \widehat{\partial} f(x_1) = \emptyset$.

To proceed, with the operator splitting technique, we introduce the following reformulation of Problem (3.1):

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^{d+1}, \boldsymbol{q}\in\mathbb{R}^n} \quad & Q(\boldsymbol{w}, \boldsymbol{q}) \coloneqq f(\boldsymbol{w}) + g(\boldsymbol{q}) \\ \text{s.t.} \quad & \boldsymbol{Z}\boldsymbol{w} = \boldsymbol{q}. \end{aligned} \tag{MSVM}$$

Then, we tackle the above reformulation with a nonconvex semi-proximal ADMM scheme. The augmented Lagrangian for Problem (MSVM) can be written as:

$$L_\beta(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda}) \coloneqq f(\boldsymbol{w}) + g(\boldsymbol{q}) + \langle \boldsymbol{\lambda}, \boldsymbol{Z}\boldsymbol{w} - \boldsymbol{q}\rangle + \frac{\beta}{2}\|\boldsymbol{Z}\boldsymbol{w} - \boldsymbol{q}\|^2,$$

where $\beta \geq 0$ is the dual step size, $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the Lagrange multiplier. To ensure convergence in such a nonconvex scenario, we need a cautiously chosen dual step size β. The overall iteration scheme can be summarized into the following diagram:

Algorithm 1 Nonconvex Semi-proximal ADMM

Input: $\boldsymbol{Z} \in \mathbb{R}^{(d+1)\times n}$, choose $\gamma > 0, c > 0$ and $\beta > 1 + \max\left\{ \frac{2\left(1+\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\right\|^2\right)}{c\sigma^2}, \frac{2}{n\sigma^2}, \frac{4}{cn\sigma^4} \right\}$.

for all $k \in \{0, 1, 2, \dots\}$ **do**

$$\begin{aligned} \boldsymbol{q}^{k+1} &\in \operatorname*{Arg\,min}_{\boldsymbol{q}} L_\beta(\boldsymbol{w}^k, \boldsymbol{q}, \boldsymbol{\lambda}^k) + \frac{\gamma}{2}\|\boldsymbol{q} - \boldsymbol{q}^k\|^2, \\ \boldsymbol{w}^{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{w}} L_\beta(\boldsymbol{w}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^k), \\ \boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \beta(\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1}). \end{aligned}$$

end for

In the sequel, we will detail the practical updating computation of each variable.

3.1.1 $\boldsymbol{q}^{k+1}$ Updating Step

To compute $\boldsymbol{q}^{k+1}$, we note that the $\boldsymbol{q}$-subproblem can be rewritten as the following nonconvex proximal problem:

$$\boldsymbol{q}^{k+1} \in \operatorname{Prox}_{\frac{1}{\beta+\gamma}g}\left(\frac{\beta}{\beta+\gamma}\left(\boldsymbol{Z}\boldsymbol{w}^k + \frac{1}{\beta}\boldsymbol{\lambda}^k\right) + \frac{\gamma}{\beta+\gamma}\boldsymbol{q}^k\right),$$

where $\operatorname{Prox}_{\frac{1}{\beta+\gamma}g} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued mapping due to the nonconvexity of g. In general, evaluation of a nonconvex proximal operator can be extremely difficult. Note that $g(\boldsymbol{v}) = \sum_{i=1}^n \phi_\rho(v_i)$ is separable and

$$\operatorname{Prox}_{\lambda g}(\boldsymbol{v}) = \bigoplus_{i=1}^n \operatorname{Prox}_{\lambda\phi_\rho}(v_i).$$

We provide the following closed-form solution for the one-dimensional separated term $\operatorname{Prox}_{\lambda\phi_\rho}$:

Lemma 3.3 (proximal operator)**.** *For $0 < \lambda < 2\rho^2$, it holds*

$$\operatorname{Prox}_{\lambda\phi_\rho}(v) = \begin{cases} \{v\} & \text{for} \quad v < -\frac{\lambda}{2\rho}, \\ \left\{v, v + \frac{\lambda}{\rho}\right\} & \text{for} \quad v = -\frac{\lambda}{2\rho}, \\ \left\{v + \frac{\lambda}{\rho}\right\} & \text{for} \quad -\frac{\lambda}{2\rho} < v \leq \rho - \frac{\lambda}{\rho}, \\ \{\rho\} & \text{for} \quad \rho - \frac{\lambda}{\rho} < v \leq \rho, \\ \{v\} & \text{for} \quad v > \rho. \end{cases}$$

For $0 < 2\rho^2 \leq \lambda$, we have

$$\operatorname{Prox}_{\lambda\phi_\rho}(v) = \begin{cases} \{v\} & \text{for} \quad v < \rho - \sqrt{2\lambda}, \\ \{v, \rho\} & \text{for} \quad v = \rho - \sqrt{2\lambda}, \\ \{\rho\} & \text{for} \quad \rho - \sqrt{2\lambda} < v < \rho, \\ \{v\} & \text{for} \quad v \geq \rho. \end{cases}$$

3.1.2 $\boldsymbol{w}^{k+1}$ Updating Step

In the $\boldsymbol{w}$-subproblem, to compute $\boldsymbol{w}^{k+1}$, we need to solve the following linear system:

$$\left(c\boldsymbol{K}+\beta\boldsymbol{Z}^\top\boldsymbol{Z}\right)\boldsymbol{w}^{k+1}=\boldsymbol{Z}^\top\left(\beta\boldsymbol{q}^{k+1}-\boldsymbol{\lambda}^k\right).$$

The following proposition certifies the existence and uniqueness of solution $\boldsymbol{w}^{k+1}$:

Proposition 3.4. *Let $p=n\beta\cdot\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\|^2$. For any $c>0,\beta>0$, we have*

$$\left(c\boldsymbol{K}+\beta\boldsymbol{Z}^\top\boldsymbol{Z}\right)\succcurlyeq\mu\cdot\boldsymbol{I},$$

where $\mu>0$ and defined by

$$\mu\coloneqq\frac{1}{2}\left(c+n\beta+p-\sqrt{(c-n\beta)^2+2p(c+n\beta)+p^2)}\right),$$

which reduces to $\mu=\min\{c,\beta n\}$ if the data points are centered and thus $p=0$.

We pre-compute $(c\boldsymbol{K}+\beta\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}\boldsymbol{Z}^\top\in\mathbb{R}^{n\times(d+1)}$, which by careful matrix partition will cost $O(n^2d)$. But in every iteration, we only need to do matrix-vector product computation and never re-compute the matrix inverse again, which only costs $O(nd)$.

3.2 Convergence Analysis

Following the idea of [Cui et al., 2020, Proposition 4.1], for our specific MSVM problem, we claim that d-stationarity is necessary and sufficient for local optimality.

Proposition 3.5. *A point $\boldsymbol{w}$ is local minimal for Problem (3.1) if and only if $\boldsymbol{w}$ is d-stationary.*

It is well-known that the first-order condition, e.g., d-stationarity, is only necessary for local optimality (see [Rockafellar and Wets, 2009, Theorem 10.1]). The sufficient optimality conditions usually require some sort of convexity [Rockafellar and Wets, 2009, Theorem 8.15] or second-order information [Rockafellar and Wets, 2009, Theorem 13.24]. Luckily, as shown in Proposition 3.5, for the ρ-margin loss SVM problem, d-stationarity is sufficient for local minimality, which is the tightest stationarity that we can hope for without knowing more global information.

We notice that recently [Suzumura et al., 2017, Theorem 4] report a KKT-type necessary and sufficient condition for local minimality of Problem (3.1). However, it turns out that their argument is flawed. We will provide a detailed discussion on that issue and a concrete counterexample to [Suzumura et al., 2017, Theorem 4] in Section 4.3.

In view of the above, an efficient algorithm that can provably generate d-stationary points for Problem (3.1) is highly desirable. We will show that the new nonconvex ADMM scheme in Algorithm 1 is capable of doing so with a local linear convergence rate.

Theorem 3.6. *Let $\gamma>0,c>0$, and*

$$\beta>1+\max\left\{\frac{2\left(1+\left\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\right\|^2\right)}{c\sigma^2},\frac{2}{n\sigma^2},\frac{4}{cn\sigma^4}\right\}.$$

For the sequence $\{(\boldsymbol{w}^k,\boldsymbol{q}^k,\boldsymbol{\lambda}^k)\}_k$ generated by Algorithm 1, the following holds:

- *When $k\to\infty$, we have sequential convergence $(\boldsymbol{w}^k,\boldsymbol{q}^k,\boldsymbol{\lambda}^k)\to(\boldsymbol{w}^*,\boldsymbol{q}^*,\boldsymbol{\lambda}^*)$ and $\boldsymbol{w}^*$ is a d-stationary point of Problem (3.1).*
- *For any $0\le T<\infty$, we have*

$$\min_{k\in[T]}\operatorname{dist}\left(0,\partial L_\beta(\boldsymbol{w}^k,\boldsymbol{q}^k,\boldsymbol{\lambda}^k)\right)\le\frac{C_1}{\sqrt{T}},$$

  where $C_1\coloneqq\tau_2\cdot\sqrt{\frac{L_\beta(\boldsymbol{w}^0,\boldsymbol{q}^0,\boldsymbol{\lambda}^0)-F^*}{\tau_1}}$. See Section 5 for definitions of τ_1,τ_2.
- *Let $\mathcal{S}_d$ be the set of d-stationary points of Problem (3.1). There exist $C_2<\infty$, $\rho\in[0,1)$ and finite $\bar{k}$ such that for all $k\ge\bar{k}$:*

$$\operatorname{dist}\left(\boldsymbol{w}^k,\mathcal{S}_d\right)\le C_2\rho^k.$$

The constants C_2 and ρ can be determined from [Han et al., 2018, Theorem 2] and Hoffman's error bound for linear system [Dontchev and Rockafellar, 2009, Lemma 3C.4]. We will sketch the main idea of proof in Section 5 and defer the formal argument to Appendix C.

4 DISCUSSION

In this section, we will first construct concrete examples that lead to failure of computing sharp stationary points by the subgradient, DCA, and homotopy methods. Then, we will explain the subtleties of the technical Assumption 3.1 and its consequences on optimality conditions and algorithmic convergence.

4.1 Subgradient Method

We consider the subgradient method with step size $t_k<\frac{1}{c}$ and iteration scheme:

$$\boldsymbol{w}^{k+1}\coloneqq\boldsymbol{w}^k-t_k\boldsymbol{v}_k,\qquad\boldsymbol{v}^k\in\partial F(\boldsymbol{w}^k).$$

Then, we introduce the following set:

$$S\coloneqq\left\{\boldsymbol{w}\in\mathbb{R}^{d+1}:\boldsymbol{Z}\boldsymbol{K}\boldsymbol{w}<0,w_{d+1}=0\right\}.$$

Assuming $\boldsymbol{w}_k \in S$, we have $\partial F(\boldsymbol{w}^k) = c \cdot \boldsymbol{K}\boldsymbol{w}^k$, which implies that

$$\boldsymbol{Z}\boldsymbol{K}\boldsymbol{w}^{k+1} = \boldsymbol{Z}(1 - t_k \cdot c)\boldsymbol{K}\boldsymbol{w}^k < 0.$$

Thus, the sequence generated by the subgradient method cannot escape from the set S in finite time. Considering any accumulation point $\boldsymbol{w}^*$ of the sequence $\{\boldsymbol{w}^k\}_k$, by $\boldsymbol{w}^* \in \operatorname{cl} S$, we will show that if $0 \in \partial F(\boldsymbol{w}^*)$ then $\boldsymbol{K}\boldsymbol{w}^* = 0$ and $0 \notin \widehat{\partial} F(\boldsymbol{w}^*)$. So, any accumulation point $\boldsymbol{w}^*$ cannot be locally minimal due to [Rockafellar and Wets, 2009, Theorem 10.1]. To this end, note that

$$0 \in (\boldsymbol{w}^*)^\top \partial F(\boldsymbol{w}^*) \subseteq c \cdot \|\boldsymbol{w}^*\|_{\boldsymbol{K}}^2 + (\boldsymbol{Z}\boldsymbol{w}^*)^\top \partial g(\boldsymbol{Z}\boldsymbol{w}^*).$$

By $\boldsymbol{w}^* \in \operatorname{cl} S$, it holds that $\boldsymbol{Z}\boldsymbol{w}^* \leq 0$ and thus $\partial g(\boldsymbol{Z}\boldsymbol{w}^*) \subseteq \{0, -1\}^n$. This gives $-c \cdot \|\boldsymbol{w}^*\|_{\boldsymbol{K}}^2 \geq 0$ and thus $\|\boldsymbol{w}^*\|_{\boldsymbol{K}} = 0$ and $\boldsymbol{w}^* = 0$. Hence, $0 \notin \widehat{\partial} F(\boldsymbol{w}^*) = \boldsymbol{Z}^\top \widehat{\partial} g(\boldsymbol{0}) = \emptyset$.

4.2 DCA-type Algorithms

For DCA-type algorithms, the failure case is similar to that for the subgradient method. Consider the DC decomposition of $\phi_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}) = \phi_{\rho,1}(\boldsymbol{z}_i^\top \boldsymbol{w}) - \phi_{\rho,2}(\boldsymbol{z}_i^\top \boldsymbol{w})$, where $\phi_{\rho,1}(\boldsymbol{z}_i^\top \boldsymbol{w}) = \max\left(1 - \frac{\boldsymbol{z}_i^\top \boldsymbol{w}}{\rho}, 0\right)$, $\phi_{\rho,2}(\boldsymbol{z}_i^\top \boldsymbol{w}) = \max\left(-\frac{\boldsymbol{z}_i^\top \boldsymbol{w}}{\rho}, 0\right)$. Then, for $\boldsymbol{w} = \boldsymbol{0}$, we have

$$0 \in \partial \phi_{\rho,1}(0) \cap \phi_{\rho,2}(0),$$

which, by summing up, implies that $\boldsymbol{w} = 0$ is DC-critical for Problem (3.1). However, as shown in Section 4.1, $\boldsymbol{w} = 0$ cannot be a d-stationary point.

4.3 Homotopy Algorithm

Recently, Suzumura et al. [2017, Theorem 4] reported a KKT-type necessary and sufficient condition for local minimality of Problem (3.1). However, the following example shows that their KKT-type condition is in fact not necessary:

Example 4.1. *Let $d = 1, n = 6, c = 1, \rho = 1$. We consider the following data points $\{(x_i, y_i) : i \in [n]\} \coloneqq \{(1, +1), (1, +1), (-1, -1), (-1, -1), (0, +1), (0, -1)\} \subseteq \mathbb{R}^d \times \{+1, -1\}$. We claim that $\boldsymbol{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is a global minimizer for Problem (3.1) on $\{(x_i, y_i) : i \in [n]\}$. However, $\boldsymbol{w}$ does not satisfy the condition in [Suzumura et al., 2017, Theorem 4]. The proof of that claim is deferred to Appendix D due to space limitation.*

The reason for the invalidation of [Suzumura et al., 2017, Theorem 4] is that, in the proof of [Suzumura et al., 2017, Theorem 3], the failure of a primal-dual pair $(f_{\mathcal{P}}^*, \boldsymbol{\alpha})$ being a KKT point does not imply the sub-optimality of primal solution $f_{\mathcal{P}}^*$.

4.4 On Assumption 3.1

The surjectivity assumption is crucial in our subdifferential d-stationary characterization (Proposition 3.2), without which efficient computation of d-stationary points would be quite hard if not impossible. An exception is the nonmonotone MM algorithm [Pang et al., 2017, Cui et al., 2018], which computes d-stationary points using the primal directional derivative characterization Definition 2.7(b) without Assumption 3.1. However, as pointed out in [Pang et al., 2017, Section 7], these DCA-type algorithms may need to solve an exponentially large number of subproblems in a single iteration step. We do not notice any algorithm that can efficiently compute d-stationary points for Problem (1.1) without the surjectivity assumption.

On the other front, the surjectivity assumption is necessary for a nonconvex ADMM scheme to converge in general. Concrete examples can be constructed [Li and Pong, 2015, Example 7] in theory, and oscillating behavior can be observed in numerical experiments. Besides, we notice that in the nonconvex ADMM literature, the surjectivity assumption on mapping $\boldsymbol{Z}$ seems pervasive and necessary [Jiang et al., 2019, Wang et al., 2019, Li and Pong, 2015].

5 PROOF SKETCH

In this section, we lay out a sketch for the proof of the main result Theorem 3.6. The formal argument is deferred to Appendix C.

5.1 Limiting Stationary Convergence

The first step is to prove the sequential convergence of $\{\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k\}_k$. To this end, we will use the augmented Lagrangian $L_\beta(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda})$ as a Lyapunov function. We first show that the sequence $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ is bounded and $L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) > -\infty$ for any $k \in \mathbb{N}$.

Lemma 5.1 (bounded and proper). *For $\beta > \frac{2}{c\sigma^2}, \gamma > 0, c > 0$, and $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ generated by Algorithm 1, there exists an $R > 0$ such that $\|(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\| \leq R$ and $L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) > -\infty$ for any $k \in \mathbb{N}$.*

Then, we show that the update rules in Algorithm 1 satisfy the following *sufficient decrease* and *limiting safeguard* properties:

Lemma 5.2 (sufficient decrease). *For $\gamma > 0, c > 0, \beta > 1 + \max\left\{\frac{2\left(1 + \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\right\|^2\right)}{c\sigma^2}, \frac{2}{n\sigma^2}, \frac{4}{cn\sigma^4}\right\}$, and $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ generated by Algorithm 1, there exists a constant $\tau_1 > 0$ only depending on $\{\sigma, c, n, \gamma, \beta\}$, such*

that

$$L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)$$
$$\leq -\tau_1 \|(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - (\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\|^2.$$

Lemma 5.3 (limiting safeguard)**.** *There exists a $\tau_2 > 0$ such that*

$$\operatorname{dist}\left(0, \partial L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1})\right)$$
$$\leq \tau_2 \|(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - (\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\|.$$

Finally, by showing that the Lyapunov function satisfies the KŁ property (see Definition C.1 in Appendix C), and with the general convergence result in [Attouch et al., 2013, Theorem 2.9], we establish the sequential convergence of Algorithm 1:

Lemma 5.4 (limiting stationarity convergence)**.** *Let $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ be the sequence generated by Algorithm 1. As $k \to +\infty$, we have $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \to (\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ with $\boldsymbol{w}^*$ is limiting-stationary, that is,*

$$0 \in c\boldsymbol{K}\boldsymbol{w}^* + \boldsymbol{Z}^\top \partial g(\boldsymbol{Z}\boldsymbol{w}^*).$$

5.2 Stationarity Refinement

For the second step, we will refine the stationary characterization of the limiting point $(\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ by showing that $\boldsymbol{w}^*$ satisfies the following sharper condition:

$$0 \in c\boldsymbol{K}\boldsymbol{w}^* + \boldsymbol{Z}^\top \widehat{\partial} g(\boldsymbol{Z}\boldsymbol{w}^*),$$

which by Proposition 3.2 indicates that $\boldsymbol{w}^*$ is d-stationary. The crux of the proof is to bypass the lack of outer semi-continuity of the subdifferential mapping $\widehat{\partial} g$ in the optimality condition. Our strategy is to wrap $\widehat{\partial} g : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ up with the nonconvex proximal mapping $\operatorname{Prox}_{\lambda g} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, whose outer semi-continuity can be obtained from the prox-boundedness of g (see Definition 2.10). See Appendix C for details.

5.3 Local Linear Convergence

Third, we will establish the local linear convergence rate of Algorithm 1 with a manifold identification argument. The key observation from the analytic solution of $\operatorname{Prox}_{\lambda\phi_\rho}$ (see Lemma 3.3) is that for any $\lambda > 0, \rho > 0$, $q \in \mathbb{R}$, and $q \in \operatorname{Prox}_{\lambda\phi_\rho}(v)$, where $v \in \mathbb{R}$ is arbitrary, we have $|q| \geq \min\left\{\frac{\lambda}{2\rho}, \rho\right\}$. This uniform lower bound allows us to find a $\bar{k}$ such that for any $k \geq \bar{k}$, the iteration $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)$ from Algorithm 1 can be viewed as running Algorithm 1 on a special convex piecewise-linear quadratic (PLQ) problem [Rockafellar and Wets, 2009, Definition 10.20]. Moreover, the KKT mapping of this convex PLQ problem is polyhedral, which implies that its inverse mapping satisfies the so-called calmness condition (see [Han et al., 2018, Definition 2]). By [Han et al., 2018, Theorem 2], the sequence $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)$ converges linearly to the set of KKT points of the convex problem. After showing that all these KKT points are indeed d-stationary for Problem (3.1), the proof is complete.

6 NUMERICAL RESULTS

In this section, we present numerical results to demonstrate the effectiveness of the new algorithmic scheme. All simulations are implemented[2] with MATLAB R2021a on a Windows 10 PC with Intel Core i7-11700K, and 32 GB RAM. We will show the numerical performance of the subgradient method, DCA, SpADMM in Algorithm 1, and a hybrid method (Algorithm 3) that runs DCA until convergence and then starts SpADMM from that DC-critical point (see Appendix E).

We first consider a toy setting with multi-dimensional Gaussian inputs and random binary labels. Let $\boldsymbol{x}_i \sim N(0, \boldsymbol{I}_d)$ and $y_i \sim 2 \cdot \text{Bernoulli}(\frac{1}{2}) - 1$ for any $i \in [n]$, where $n = 50, d = 100$. We run subGD (subgradient), DCA, SpADMM (Algorithm 1) and Hybrid Algorithm 3 with the same Gaussian initial point $\boldsymbol{w}^0$ for 150 iteration steps and record the objective function value and the augmented Lagrangian function value of Algorithm 1. We report the difference between the current objective function value and the best objective value among all algorithms in Figure 3. It is notable that *the reflection-like curve for augmented Lagrangian* of Algorithm 1 is due to the fact that the augmented Lagrangian value could be smaller than the best objective function value. From Figure 3, we highlight two interesting observations:

- It is easy to see in Figure 3 that both Algorithm 1 and Algorithm 3 outperform subGD and DCA, which is because the semi-proximal ADMM scheme escapes from non-locally minimal stationary points and will eventually converge only to a locally minimal one.
- When running the Hybrid Algorithm 3, we can see it decreases the objective value even further from the DC-critical point that DCA converges to. In other words, SpADMM escaped from the DC-critical point that is not locally minimal. This provides empirical evidence that computing a sharper kind of stationarity does not only have advantages in theory but also in numerical performance.

We also conduct simulations on real-world datasets, which are all from the LIBSVM datasets collection.

[2] See https://github.com/icety3/rho-marginSVM

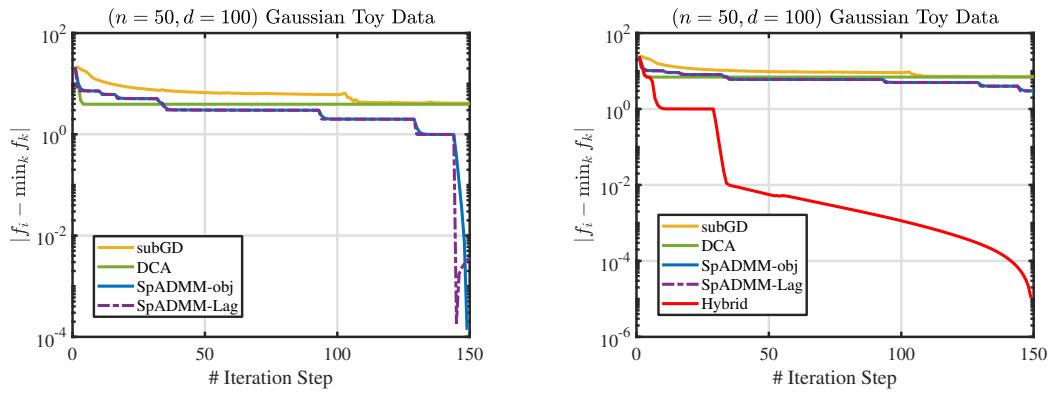

Figure 3: Performance of subGD (subgradient), DCA, SpADMM (Algorithm 1), and Hybrid Algorithm 3 on Gaussian Toy Data. The min in log-scaled y-axis is taken over all aforementioned algorithms.

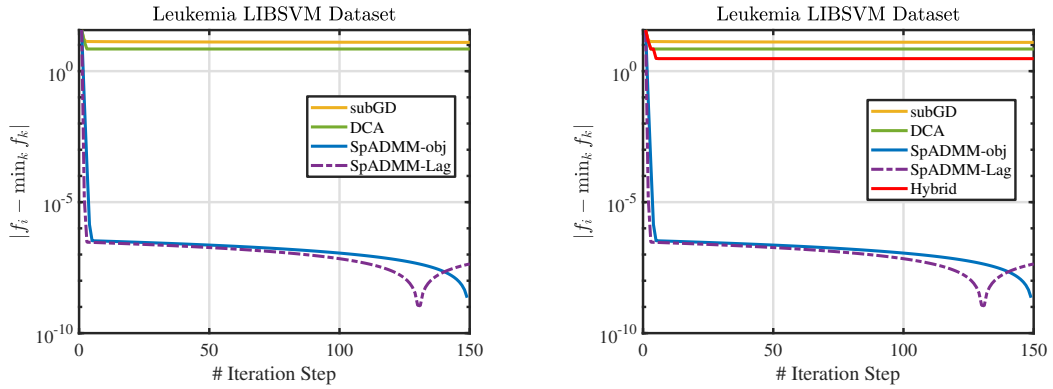

Figure 4: Performance of subGD (subgradient), DCA, SpADMM (Algorithm 1), and Hybrid Algorithm 3 on Real-world Data. The min in log-scaled y-axis is taken over all aforementioned algorithms.

The results are reported in Figure 4 (see also Figure 5 in Appendix F.1). It is easy to see that our observations from toy data remain true for real-world data. Besides, we note that the Hybrid Algorithm 3 does not necessarily outperform SpADMM in Algorithm 1 (see the results for `Leukemia` in Figure 4 and `Duke breast-cancer` in Appendix F.1, Figure 5).

7 LIMITATIONS & REMARKS

In this paper, we proposed a highly efficient nonconvex semi-proximal ADMM-based scheme that provably computes d-stationary points for the ρ-margin loss SVM problem, which is nonconvex, nonsmooth, and Clarke irregular. We further show that the local convergence rate of the nonconvex ADMM is linear. Our development certainly has its limitations and might inspire further investigation for new tools and ideas for nonconvex nonsmooth models in machine learning.

- One limitation is of course the surjectivity Assumption 3.1 in our analysis. Without this assumption, we are not aware of any algorithm that can efficiently compute d-stationary points for Problem (1.1). One might guess that Assumption 3.1 can be relaxed with a better analysis of the ADMM scheme. But we think a more intriguing direction is to find a better computable d-stationary characterization (e.g., Proposition 3.2). In [Pang et al., 2017], their enhanced DCA algorithm works on a primal directional derivative condition, which needs less regularity conditions to hold but causes serious computational trouble in the updating steps. Our dual subdifferential condition decomposes the nonsmooth parts into separable pieces and leads to efficient computing with extra cost on more assumptions.

- The other limitation is that our scheme is required to solve a nonconvex subproblem (i.e., to compute $\mathrm{Prox}_{\lambda g}$) to optimality, which is crucial to do a limiting argument without osc in $\widehat{\partial} g$. For Problem (1.1), we do these computations by hand and the optimal solution is carefully computed as a closed-form (see Lemma 3.3). However, for a general nonconvex problem, it is unrealistic to compute its proximal mapping. Finding a computational cheaper mechanism to do limiting argument for $\widehat{\partial} g$ would be interesting.

References

Amir Ali Ahmadi and Jeffrey Zhang. On the complexity of finding a local minimizer of a quadratic function over a polytope. *Mathematical Programming*, 2022.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

J. Paul Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59(2):467–479, 2011.

Emilio Carrizosa, Amaya Nogales-Gómez, and Dolores Romero Morales. Heuristic approaches for support vector machines with the ramp loss. *Optimization Letters*, 8(3):1125–1135, 2014.

Augustin-Louis Cauchy. *Exercices de Mathématiques*, volume 3. De Bure frères, 1828.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *International Conference on Machine Learning*, pages 201–208. PMLR, 2006a.

Ronan Collobert, Fabian Sinz, Jason Weston, Léon Bottou, and Thorsten Joachims. Large scale transductive SVMs. *Journal of Machine Learning Research*, 7(8), 2006b.

Ying Cui, Jong-Shi Pang, and Bodhisattva Sen. Composite difference-max programs for modern statistical estimation problems. *SIAM Journal on Optimization*, 28(4):3344–3374, 2018.

Ying Cui, Tsung-Hui Chang, Mingyi Hong, and Jong-Shi Pang. A study of piecewise linear-quadratic programs. *Journal of Optimization Theory and Applications*, 186(2):523–553, 2020.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings*, volume 543. Springer, 2009.

Seyda Ertekin, Leon Bottou, and C. Lee Giles. Nonconvex online support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):368–381, 2010.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

Deren Han, Defeng Sun, and Liwei Zhang. Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Mathematics of Operations Research*, 43(2):622–637, 2018.

D. Hertz, C.S. Adjiman, and C.A. Floudas. Two results on bounding the roots of interval polynomials. *Computers & Chemical Engineering*, 23(9):1333–1339, 1999.

Xiaolin Huang, Lei Shi, and Johan A.K. Suykens. Ramp loss linear programming support vector machine. *Journal of Machine Learning Research*, 15(1):2185–2211, 2014.

Bo Jiang, Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, 72(1):115–157, 2019.

Joseph Keshet and David McAllester. Generalization bounds and consistency for latent structural probit and ramp loss. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783, 1998.

Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4): 2434–2460, 2015.

Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in non-smooth optimization. *IEEE Signal Processing Magazine*, 37 (5):18–31, 2020.

Yufeng Liu, Xiaotong Shen, and Hani Doss. Multicategory ψ-learning and support vector machine: computational tools. *Journal of Computational and Graphical Statistics*, 14(1):219–236, 2005.

Søren Frejstrup Maibing and Christian Igel. Computational complexity of linear large margin classification with ramp loss. In *Artificial Intelligence and Statistics*, pages 259–267. PMLR, 2015.

Szymon Majewski, Błażej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

Athanasios Migdalas and P.M. Pardalos. A note on open problems and challenges in optimization theory and algorithms. In *Open Problems in Optimization and Data Analysis*, pages 1–8. Springer, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer Nature, second edition, 2018.

Maher Nouiehed, Jong-Shi Pang, and Meisam Razaviyayn. On the pervasiveness of difference-convexity in optimization and statistics. *Mathematical Programming*, 174(1):195–222, 2019.

Jong-Shi Pang, Meisam Razaviyayn, and Alberth Alvarado. Computing B-stationary points of nonsmooth DC programs. *Mathematics of Operations Research*, 42(1):95–118, 2017.

Stephen M Robinson. Some continuity properties of polyhedral multifunctions. In *Mathematical Programming at Oberwolfach*, pages 206–214. Springer, 1981.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

Xiaotong Shen, George C. Tseng, Xuegong Zhang, and Wing Hung Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Shinya Suzumura, Kohei Ogawa, Masashi Sugiyama, and Ichiro Takeuchi. Outlier path: a homotopy algorithm for robust SVM. In *International Conference on Machine Learning*, pages 1098–1106. PMLR, 2014.

Shinya Suzumura, Kohei Ogawa, Masashi Sugiyama, Masayuki Karasuyama, and Ichiro Takeuchi. Homotopy continuation approaches for robust SV classification and regression. *Machine Learning*, 106(7): 1009–1038, 2017.

Huajun Wang, Yuanhai Shao, and Naihua Xiu. Proximal operator and optimality conditions for ramp loss SVM. *Optimization Letters*, pages 1–16, 2021.

Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1): 29–63, 2019.

Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

Supplementary Material: Computing D-Stationary Points of ρ-Margin Loss SVM

A Optimality Condition

For preparation, the optimality conditions for every update iteration of Algorithm 1 can be written as:

$$\begin{cases} 0 \in \widehat{\partial} g(\boldsymbol{q}^{k+1}) + \beta\left(\boldsymbol{q}^{k+1} - \boldsymbol{Z}\boldsymbol{w}^k - \frac{1}{\beta}\boldsymbol{\lambda}^k\right) + \gamma(\boldsymbol{q}^{k+1} - \boldsymbol{q}^k), & \text{(A.1a)} \\ 0 = c\boldsymbol{K}\boldsymbol{w}^{k+1} + \beta\boldsymbol{Z}^\top\left(\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1} + \frac{1}{\beta}\boldsymbol{\lambda}^k\right), & \text{(A.1b)} \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta(\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1}). & \text{(A.1c)} \end{cases}$$

B Technical Lemmas

Lemma B.1. *For any $\boldsymbol{A} \in \mathbb{R}^{n\times(d+1)}$, and any $\boldsymbol{p} \in \mathbb{R}^n$, it holds $\|\boldsymbol{A}^\top\boldsymbol{p}\| \geq \sigma_{\min}(\boldsymbol{A}) \cdot \|\boldsymbol{p}\|$.*

Proof. Let the thin-SVD of $\boldsymbol{A}$ be $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. We compute

$$\|\boldsymbol{A}^\top\boldsymbol{p}\|^2 = \|\boldsymbol{\Sigma}\boldsymbol{U}^\top\boldsymbol{p}\|^2 = \sum_{i=1}^{n} \sigma_i^2(\boldsymbol{A}) \cdot (\boldsymbol{u}_i^\top\boldsymbol{p})^2 \geq \sigma_{\min}^2(\boldsymbol{A}) \cdot \|\boldsymbol{U}^\top\boldsymbol{p}\|^2 = \sigma_{\min}^2(\boldsymbol{A}) \cdot \|\boldsymbol{p}\|^2,$$

which completes the proof. □

Lemma B.2. *For $c_1, c_2, c_3 > 0$, it holds*

$$\frac{1}{2}\left(c_1 + c_2c_3^2 + c_2 - \sqrt{\left(c_1 + c_2c_3^2 + c_2\right)^2 - 4c_1c_2}\right) = \min_{0\leq t\leq 1} Q(t) \coloneqq c_1 \cdot t^2 + c_2 \cdot \left(c_3 \cdot t - \sqrt{1-t^2}\right)^2$$

Proof. Note that

$$Q(t) = \left(c_1 + c_2c_3^2 - c_2\right) \cdot t^2 - 2c_2c_3 \cdot t\sqrt{1-t^2} + c_2.$$

As mapping $t^2 \mapsto t$ is bijective in $[0,1]$, we have

$$\min_{0\leq t\leq 1} Q(t) = \min_{0\leq t\leq 1} P(t) \coloneqq \left(c_1 - c_2 + c_2c_3^3\right) \cdot t - 2c_2c_3 \cdot \sqrt{t} \cdot \sqrt{1-t} + c_2.$$

It is elementary to see $x \mapsto -\sqrt{x} \cdot \sqrt{1-x}$ is convex on $[0,1]$. Thus, $P(t)$ is convex on $[0,1]$. By the first-order optimality condition and the fact $\frac{1}{2}\left(c_4 - \sqrt{1+c_4^2}\right) < \frac{1}{2}\left(c_4 - |c_4|\right) = \min\{c_4, 0\}$, it holds for $c_4 \in \mathbb{R}$ that

$$\min_{0\leq t\leq 1} c_4 \cdot t - \sqrt{t} \cdot \sqrt{1-t} = \frac{1}{2}\left(c_4 - \sqrt{1+c_4^2}\right) \quad \text{with} \quad t^* = \frac{c_4^2 - c_4\sqrt{1+c_4^2} + 1}{2c_4^2 + 2} \in (0,1).$$

The proof completes with simple algebraic manipulation. □

Proposition 3.4. *Let $p = n\beta \cdot \|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\|^2$. For any $c > 0, \beta > 0$, we have*

$$\left(c\boldsymbol{K} + \beta\boldsymbol{Z}^\top\boldsymbol{Z}\right) \succcurlyeq \mu \cdot \boldsymbol{I},$$

where $\mu > 0$ and defined by

$$\mu \coloneqq \frac{1}{2}\left(c + n\beta + p - \sqrt{(c - n\beta)^2 + 2p(c + n\beta) + p^2)}\right),$$

which reduces to $\mu = \min\{c, \beta n\}$ if the data points are centered and thus $p = 0$.

Proof. Let data points $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and labels $\boldsymbol{y} \in \{+1,-1\}^n$. We define $\widetilde{\boldsymbol{X}} \coloneqq \boldsymbol{X} \odot \boldsymbol{y}\mathbb{1}_d^\top, \boldsymbol{v}_1 \in \mathbb{R}^d, v_2 \in \mathbb{R}$, and $\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ v_2 \end{bmatrix} \in \mathbb{R}^{d+1}$. Thus, $\boldsymbol{Z} = \begin{bmatrix} \widetilde{\boldsymbol{X}} & \boldsymbol{y} \end{bmatrix} \in \mathbb{R}^{n\times(d+1)}$. Then, for any $\boldsymbol{v} : \|\boldsymbol{v}\| = 1$, we compute

$$
\begin{aligned}
&\boldsymbol{v}^\top \left(c\boldsymbol{K} + \beta \boldsymbol{Z}^\top \boldsymbol{Z}\right) \boldsymbol{v} \\
&= c\cdot\|\boldsymbol{v}_1\|^2 + \beta\cdot\|\boldsymbol{Z}\boldsymbol{v}\|^2 \\
&= c\cdot\|\boldsymbol{v}_1\|^2 + \beta\cdot\|\widetilde{\boldsymbol{X}}\boldsymbol{v}_1 + v_2\cdot\boldsymbol{y}\|^2 \\
&= c\cdot\|\boldsymbol{v}_1\|^2 + \beta\cdot\left\|\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{y}\boldsymbol{y}^\top\right)\widetilde{\boldsymbol{X}}\boldsymbol{v}_1 + \frac{1}{n}\boldsymbol{y}\boldsymbol{y}^\top\widetilde{\boldsymbol{X}}\boldsymbol{v}_1 + v_2\cdot\boldsymbol{y}\right\|^2 \\
&\geq c\cdot\|\boldsymbol{v}_1\|^2 + \beta\cdot\left\|\frac{1}{n}\boldsymbol{y}\boldsymbol{y}^\top\widetilde{\boldsymbol{X}}\boldsymbol{v}_1 + v_2\cdot\boldsymbol{y}\right\|^2 \qquad \text{(Pythagorean)} \\
&= c\cdot\|\boldsymbol{v}_1\|^2 + n\beta\cdot\left(\left(\frac{1}{n}\mathbb{1}_n^\top\boldsymbol{X}\right)\boldsymbol{v}_1 + v_2\right)^2. \qquad (\natural)
\end{aligned}
$$

Let $\bar{\boldsymbol{x}} \coloneqq \frac{1}{n}\mathbb{1}_n^\top\boldsymbol{X}, \theta \coloneqq \bar{\boldsymbol{x}}^\top\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}, \|\boldsymbol{v}_1\| = t$ with $0 \leq t \leq 1$. Without loss of generality, as $\bar{\boldsymbol{x}}$ is arbitrary, we assume $\theta \geq 0$ and $v_2 = -\sqrt{1-t^2} \leq 0$ (otherwise, the value of Equation ($\natural$) only increases). We continue with

$$
\text{Equation } (\natural) \geq \min_{0\leq t\leq 1} c\cdot t^2 + n\beta\cdot\left(\theta\cdot t - \sqrt{1-t^2}\right)^2.
$$

By Lemma B.2, we have

$$
\text{Equation } (\natural) \geq \frac{1}{2}\left(c + n\beta + n\beta\theta^2 - \sqrt{(c-n\beta)^2 + n\beta\theta^2(2c + 2n\beta + n\beta\theta^2)}\right).
$$

Note that $n\beta\theta^2 \leq n\beta\|\bar{\boldsymbol{x}}\|^2 =: p$ and the right-hand side of above lower bound is monotone decreasing with respect to θ^2. We have Equation ($\natural$) $\geq \mu$. To see $\mu > 0$, note the following identity

$$
\mu = \frac{1}{2}\left((c + n\beta + p) - \sqrt{(c+n\beta+p)^2 - 4cn\beta}\right) > 0,
$$

which completes the proof. □

Lemma 5.2 (sufficient decrease). *For $\gamma > 0, c > 0, \beta > 1 + \max\left\{\frac{2\left(1+\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\right\|^2\right)}{c\sigma^2}, \frac{2}{n\sigma^2}, \frac{4}{cn\sigma^4}\right\}$, and $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ generated by Algorithm 1, there exists a constant $\tau_1 > 0$ only depending on $\{\sigma, c, n, \gamma, \beta\}$, such that*

$$
\begin{aligned}
&L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \\
\leq\ &-\tau_1\|(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - (\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\|^2.
\end{aligned}
$$

Proof. Note that

$$
\begin{aligned}
L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \leq\ &\underbrace{L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^k)}_{T_1} + \\
&\underbrace{L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^k) - L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^k)}_{T_2} + \\
&\underbrace{L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^k) - L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)}_{T_3}.
\end{aligned}
$$

For T_1, a direct computation and Equation (A.1c) gives

$$
T_1 = \langle \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k, \boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1}\rangle = \frac{1}{\beta}\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|^2.
$$

Using Equation (A.1b) and Equation (A.1c), we have

$$
\boldsymbol{Z}^\top\boldsymbol{\lambda}^{k+1} = \boldsymbol{Z}^\top(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k) + \boldsymbol{Z}^\top\boldsymbol{\lambda}^k = \beta\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1}) + \boldsymbol{Z}^\top\boldsymbol{\lambda}^k = -\boldsymbol{K}\boldsymbol{w}^{k+1}.
$$

As $\boldsymbol{Z}$ is surjective, it holds $\sigma \coloneqq \sigma_{\min}(\boldsymbol{Z}) > 0$. Combining above computation and Lemma B.1, we have

$$T_1 = \frac{1}{\beta}\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|^2 \le \frac{1}{\sigma^2\beta}\|\boldsymbol{Z}^\top(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k)\|^2 = \frac{1}{\sigma^2\beta}\|\boldsymbol{K}(\boldsymbol{w}^{k+1} - \boldsymbol{w}^k)\|^2 \le \frac{1}{\sigma^2\beta}\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\|^2.$$

For T_2, we first note the strong convexity of the $\boldsymbol{w}$-subproblem, which is due to the positive definite Hessian matrix $c\boldsymbol{K} + \beta\boldsymbol{Z}^\top\boldsymbol{Z} \succcurlyeq \mu \cdot \boldsymbol{I}$ in Proposition 3.4. Then, by [Nesterov, 2018, Theorem 2.1.8], we have

$$T_2 \le -\mu\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\|^2 \le -\frac{\mu}{2}\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\|^2 - \frac{\sigma^2\mu}{2}\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|^2,$$

where the last inequality is due to the primal-dual control:

$$\sigma \cdot \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\| \le \|\boldsymbol{Z}^\top(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k)\| = \|\boldsymbol{K}(\boldsymbol{w}^{k+1} - \boldsymbol{w}^k)\| \le \|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\|.$$

For T_3, by optimality condition Equation (A.1a), it holds

$$T_3 \le -\frac{\gamma}{2}\|\boldsymbol{q}^{k+1} - \boldsymbol{q}^k\|^2.$$

In summary, we bound

$$T_1 + T_2 + T_3 \le \left(\frac{1}{\sigma^2\beta} - \frac{\mu}{2}\right)\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\|^2 - \frac{\sigma^2\mu}{2}\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|^2 - \frac{\gamma}{2}\|\boldsymbol{q}^{k+1} - \boldsymbol{q}^k\|^2.$$

Finally, we choose parameters as

$$\beta > \frac{2}{\sigma^2\mu}, \quad \gamma > 0, \quad \text{and } \tau_1 = \min\left\{\left(\frac{1}{\sigma^2\beta} - \frac{\mu}{2}\right), \frac{\sigma^2\mu}{2}, \frac{\gamma}{2}\right\} > 0.$$

To get an explicit lower estimation for β, we plug in μ in Proposition 3.4. After algebraic simplification on condition $\beta > \frac{2}{\sigma^2\mu}$, we have

$$\beta^2\left(\theta^2 n\sigma^2 + n\sigma^2\right) + \beta c\sigma^2 - 4 > \beta\sigma^2\sqrt{(c + \beta\theta^2 n + \beta n)^2 - 4\beta cn}.$$

Then, it suffices to ensure $\beta^2\left(\theta^2 n\sigma^2 + n\sigma^2\right) + \beta c\sigma^2 - 4 > 0$ and

$$\left(\beta^2\left(\theta^2 n\sigma^2 + n\sigma^2\right) + \beta c\sigma^2 - 4\right)^2 - \left(\beta\sigma^2\sqrt{(c + \beta\theta^2 n + \beta n)^2 - 4\beta cn}\right)^2 > 0.$$

Solving the quadratic inequality and simplifying the cubic one, we get $\beta > \max\{\mathrm{Rt}_3(\sharp), \nu\}$, where

$$\nu \coloneqq \frac{1}{2}\sqrt{\frac{c^2}{n^2\left(\theta^2+1\right)^2} + \frac{16}{n\left(\theta^2+1\right)\sigma^2}} - \frac{c}{2n\left(\theta^2+1\right)},$$

and $\mathrm{Rt}_3(\sharp)$ is the largest positive root of the following cubic equation with respect to t:

$$cn\sigma^4 t^3 - 2n\sigma^2\left(1 + \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\right\|^2\right)t^2 - 2\sigma^2 ct + 4 = 0. \tag{$\sharp$}$$

Then, we use the Cauchy's bound Cauchy [1828] (see also [Hertz et al., 1999, Theorem 2.1]) for real zeros of polynomials to get

$$\mathrm{Rt}_3(\sharp) \le 1 + \max\left\{\frac{2\left(1 + \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\right\|^2\right)}{c\sigma^2}, \frac{2}{n\sigma^2}, \frac{4}{cn\sigma^4}\right\}.$$

Finally, note that $\nu \le \mathrm{Rt}_3(\sharp)$ and the proof completes. □

Lemma 5.1 (bounded and proper). *For $\beta > \frac{2}{c\sigma^2}, \gamma > 0, c > 0$, and $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ generated by Algorithm 1, there exists an $R > 0$ such that $\|(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\| \leq R$ and $L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) > -\infty$ for any $k \in \mathbb{N}$.*

Proof. By Lemma 5.2, we have for any $k \in \mathbb{N}$:

$$\begin{aligned} L_\beta(\boldsymbol{w}^0, \boldsymbol{q}^0, \boldsymbol{\lambda}^0) &\geq L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \\ &= \frac{c}{2}\|\boldsymbol{w}^k\|_{\boldsymbol{K}}^2 + g(\boldsymbol{q}^k) + \frac{\beta}{2}\|\boldsymbol{Z}\boldsymbol{w}^k - \boldsymbol{q}^k + \frac{1}{\beta}\boldsymbol{\lambda}^k\|^2 - \frac{1}{2\beta}\|\boldsymbol{\lambda}^k\|^2 \\ &\geq \frac{c}{2}\|\boldsymbol{w}^k\|_{\boldsymbol{K}}^2 + g(\boldsymbol{q}^k) + \frac{\beta}{2}\|\boldsymbol{Z}\boldsymbol{w}^k - \boldsymbol{q}^k + \frac{1}{\beta}\boldsymbol{\lambda}^k\|^2 - \frac{1}{2\beta\sigma^2}\|\boldsymbol{K}\boldsymbol{w}^k\|^2 \\ &\geq \frac{c}{4}\|\boldsymbol{w}^k\|_{\boldsymbol{K}}^2 + g(\boldsymbol{q}^k) + \frac{\beta}{2}\|\boldsymbol{Z}\boldsymbol{w}^k - \boldsymbol{q}^k + \frac{1}{\beta}\boldsymbol{\lambda}^k\|^2 \geq 0, \end{aligned}$$

where the second inequality is due to $\sigma \cdot \|\boldsymbol{\lambda}^k\| \leq \|\boldsymbol{Z}^\top\boldsymbol{\lambda}^k\| = \|\boldsymbol{K}\boldsymbol{w}^k\|$, the final inequality is by facts that $\boldsymbol{K}$ is idempotent and $\left(\frac{c}{2} - \frac{1}{2\beta\sigma^2}\right) > \frac{c}{4}$. Therefore, the augmented Lagrangian is lower bounded by 0 for any $k \in \mathbb{N}$ and the properness of $L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)$ for any $k \in \mathbb{N}$ directly follows.

Then, similar to the proof of Proposition 3.4, we can prove $\frac{c}{4}\boldsymbol{K} + \frac{\beta}{2}\boldsymbol{Z}^\top\boldsymbol{Z} \succ 0$. Thus the level-set of L_β with respect to $\boldsymbol{w}$ is bounded and we have $\{\boldsymbol{w}^k\}_k$ is bounded. Again, by Equations (A.1b) and (A.1c) and surjective $\boldsymbol{Z}$, we have $\|\boldsymbol{w}^k\| \geq \|\boldsymbol{K}\boldsymbol{w}^k\| = \|\boldsymbol{Z}^\top\boldsymbol{\lambda}^k\| \geq \sigma\|\boldsymbol{\lambda}^k\|$, and $\{\boldsymbol{\lambda}^k\}_k$ is bounded. By Equation (A.1c), $\|\boldsymbol{q}^k\| = \|\boldsymbol{Z}\boldsymbol{w}^k + \frac{1}{\beta}\boldsymbol{\lambda}^k - \frac{1}{\beta}\boldsymbol{\lambda}^{k+1}\| \leq \|\boldsymbol{Z}\|\|\boldsymbol{w}^k\| + \frac{1}{\beta}\|\boldsymbol{\lambda}^k\| + \frac{1}{\beta}\|\boldsymbol{\lambda}^{k+1}\|$, and $\{\boldsymbol{q}^k\}_k$ is bounded. □

Lemma 5.3 (limiting safeguard). *There exists a $\tau_2 > 0$ such that*

$$\begin{aligned} &\operatorname{dist}\left(0, \partial L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1})\right) \\ &\leq \tau_2\|(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) - (\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\|. \end{aligned}$$

Proof. To prove the safeguard property, we need to characterize the limiting subdifferential ∂L_β. Due to the irregularity of $g(\boldsymbol{q})$, a rigorous argument needs to apply a series of subdifferential calculus rules in a carefully chosen order. To this end, by grouping smooth terms in L_β together, by [Rockafellar and Wets, 2009, Exercise 8.8(c)], calmness of L_β from local Lipschitz, [Rockafellar and Wets, 2009, Proposition 8.32], and then with [Rockafellar and Wets, 2009, Proposition 10.5], we have

$$\partial L_\beta(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda}) = \begin{bmatrix} c\boldsymbol{K}\boldsymbol{w} + \boldsymbol{Z}^\top\boldsymbol{\lambda} + \beta\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{w} - \boldsymbol{q}) \\ -\boldsymbol{\lambda} + \beta(\boldsymbol{q} - \boldsymbol{Z}\boldsymbol{w}) + \partial g(\boldsymbol{q}) \\ \boldsymbol{Z}\boldsymbol{w} - \boldsymbol{q} \end{bmatrix}.$$

By Cauchy-Schwarz, it holds

$$\begin{aligned} \operatorname{dist}\left(0, \partial L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1})\right) \leq \|c\boldsymbol{K}\boldsymbol{w}^{k+1} + \boldsymbol{Z}^\top\boldsymbol{\lambda}^{k+1} + \beta\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1})\| + \\ \operatorname{dist}\left(0, -\boldsymbol{\lambda}^{k+1} + \beta(\boldsymbol{q}^{k+1} - \boldsymbol{Z}\boldsymbol{w}^{k+1}) + \partial g(\boldsymbol{q}^{k+1})\right) + \|\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1}\|. \end{aligned}$$

By Equation (A.1b), we have

$$\|c\boldsymbol{K}\boldsymbol{w}^{k+1} + \boldsymbol{Z}^\top\boldsymbol{\lambda}^{k+1} + \beta\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1})\| = \|\boldsymbol{Z}^\top(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k)\| \leq \|\boldsymbol{Z}\|\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|.$$

Meanwhile, by Equation (A.1a) and triangle inequality, we compute

$$\operatorname{dist}\left(0, -\boldsymbol{\lambda}^{k+1} + \beta(\boldsymbol{q}^{k+1} - \boldsymbol{Z}\boldsymbol{w}^{k+1}) + \partial g(\boldsymbol{q}^{k+1})\right) \leq \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\| + \gamma\|\boldsymbol{q}^{k+1} - \boldsymbol{q}^k\| + \beta\|\boldsymbol{Z}\|\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\|.$$

Finally, using Equation (A.1c), summing up, we get

$$\operatorname{dist}\left(0, \partial L_\beta(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1})\right) \leq \beta\|\boldsymbol{Z}\|\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\| + \gamma\|\boldsymbol{q}^{k+1} - \boldsymbol{q}^k\| + \left(\frac{\beta+1}{\beta} + \|\boldsymbol{Z}\|\right)\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|,$$

which combing with Cauchy-Schwarz and the choosing

$$\tau_2 \coloneqq \sqrt{3}\max\left\{\beta\|\boldsymbol{Z}\|, \gamma, \left(\frac{\beta+1}{\beta}+\|\boldsymbol{Z}\|\right)\right\},$$

completes the proof. □

Lemma 5.4 (limiting stationarity convergence)**.** *Let $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ be the sequence generated by Algorithm 1. As $k \to +\infty$, we have $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \to (\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ with $\boldsymbol{w}^*$ is limiting-stationary, that is,*

$$0 \in c\boldsymbol{K}\boldsymbol{w}^* + \boldsymbol{Z}^\top \partial g(\boldsymbol{Z}\boldsymbol{w}^*).$$

Proof. The proof is by verifying the conditions in [Attouch et al., 2013, Theorem 2.9]. For H1, it directly holds from Lemma 5.2. H2 is from Lemma 5.3. H3 is due to Lemma 5.1 and continuity of L_β. We still need to show L_β is a KŁ function. To see this, we simply notice that $h_1(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda}) \coloneqq \frac{c}{2}\|\boldsymbol{w}\|_{\boldsymbol{K}}^2 + \frac{\beta}{2}\|\boldsymbol{Z}\boldsymbol{w} - \boldsymbol{q} + \frac{1}{\beta}\boldsymbol{\lambda}\|^2 - \frac{1}{2\beta}\|\boldsymbol{\lambda}\|^2$ is a polynomial function and $h_2(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda}) \coloneqq g(\boldsymbol{q})$ has a piecewise linear graph. Hence the sum $L_\beta(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda}) = h_1(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda}) + h_2(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda})$ is semi-algebraic and thus KŁ Kurdyka [1998]. The proof completes by using [Attouch et al., 2013, Theorem 2.9]. □

Lemma 3.3 (proximal operator)**.** *For $0 < \lambda < 2\rho^2$, it holds*

$$\mathrm{Prox}_{\lambda\phi_\rho}(v) = \begin{cases} \{v\} & \text{for} \quad v < -\frac{\lambda}{2\rho}, \\ \left\{v, v+\frac{\lambda}{\rho}\right\} & \text{for} \quad v = -\frac{\lambda}{2\rho}, \\ \left\{v+\frac{\lambda}{\rho}\right\} & \text{for} \quad -\frac{\lambda}{2\rho} < v \le \rho - \frac{\lambda}{\rho}, \\ \{\rho\} & \text{for} \quad \rho - \frac{\lambda}{\rho} < v \le \rho, \\ \{v\} & \text{for} \quad v > \rho. \end{cases}$$

For $0 < 2\rho^2 \le \lambda$, we have

$$\mathrm{Prox}_{\lambda\phi_\rho}(v) = \begin{cases} \{v\} & \text{for} \quad v < \rho - \sqrt{2\lambda}, \\ \{v, \rho\} & \text{for} \quad v = \rho - \sqrt{2\lambda}, \\ \{\rho\} & \text{for} \quad \rho - \sqrt{2\lambda} < v < \rho, \\ \{v\} & \text{for} \quad v \ge \rho. \end{cases}$$

Proof. As the computation of the proximal operator of the margin-loss Φ_ρ is quite tedious and complicated, we will deduce its explicit expression by calling on existing result Wang et al. [2021]. By definition, we have

$$\mathrm{Prox}_{\lambda\phi_\rho}(v) = \operatorname*{Arg\,min}_x \left\{\frac{1}{2\lambda}(x-v)^2 + \phi_\rho(x)\right\}.$$

Let $y = 1 - \frac{x}{\rho}$ and using identity $\max(a, \min(b, c)) = \min(b, \max(a, c))$, we have

$$\mathrm{Prox}_{\lambda\phi_\rho}(v) = \rho - \rho \cdot \operatorname*{Arg\,min}_y \left\{\frac{\rho^2}{2\lambda}\left(y + \frac{v}{\rho} - 1\right)^2 + \max(0, \min(1, y))\right\} = \rho - \rho \cdot \mathrm{Prox}_{\frac{\lambda}{\rho^2}\ell_r}\left(1 - \frac{v}{\rho}\right),$$

where ℓ_r is defined in [Wang et al., 2021, Equation (2)]. Then, the claim follows from [Wang et al., 2021, Proposition 1], [Wang et al., 2021, Proposition 2]. □

Proposition B.3. *g in Equation (MSVM) is prox-bounded for all $\lambda_g > 0$.*

Proof. Trivial as g is lower-bounded by 0 globally. □

Proposition 3.5. *A point $\boldsymbol{w}$ is local minimal for Problem (3.1) if and only if $\boldsymbol{w}$ is d-stationary.*

Proof. The "only if" part is directly from [Rockafellar and Wets, 2009, Theorem 10.1]. For the "if" part, let $\boldsymbol{w}^*$ be a d-stationary point of F with $F'(\boldsymbol{w}^*; \boldsymbol{w} - \boldsymbol{w}^*) \ge 0, \forall \boldsymbol{w} \in \mathbb{R}^{d+1}$ (see Definition 2.7(b)). As $g(\boldsymbol{Z}\boldsymbol{w})$ is piecewise

affine with respect to $\boldsymbol{w}$, by the min-max representation and with [Facchinei and Pang, 2007, Equation (4.2.7)], there exists $\delta > 0$ such that for all $\|\boldsymbol{w} - \boldsymbol{w}^*\| \le \delta$, we have

$$g(\boldsymbol{Z}\boldsymbol{w}) = g(\boldsymbol{Z}\boldsymbol{w}^*) + g'(\boldsymbol{Z}\boldsymbol{w}^*; \boldsymbol{Z}\boldsymbol{w} - \boldsymbol{Z}\boldsymbol{w}^*).$$

Then, we compute

$$\begin{aligned}
F(\boldsymbol{w}) &= \frac{c}{2}\|\boldsymbol{w}^* + (\boldsymbol{w} - \boldsymbol{w}^*)\|_{\boldsymbol{K}}^2 + \sum_{i=1}^n \phi_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}) \\
&\ge \frac{c}{2}\|\boldsymbol{w}^*\|_{\boldsymbol{K}}^2 + c(\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{K}\boldsymbol{w}^* + \sum_{i=1}^n \phi_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}^*) + \sum_{i=1}^n \phi_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}) - \phi_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}^*) \\
&= \frac{c}{2}\|\boldsymbol{w}^*\|_{\boldsymbol{K}}^2 + \sum_{i=1}^n \phi_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}^*) + c(\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{K}\boldsymbol{w}^* + \sum_{i=1}^n \phi_\rho'(\boldsymbol{z}_i^\top \boldsymbol{w}^*; \boldsymbol{z}_i^\top \boldsymbol{w} - \boldsymbol{z}_i^\top \boldsymbol{w}^*) \\
&= \frac{c}{2}\|\boldsymbol{w}^*\|_{\boldsymbol{K}}^2 + \sum_{i=1}^n \phi_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}^*) + F'(\boldsymbol{w}^*; \boldsymbol{w} - \boldsymbol{w}^*) \ge F(\boldsymbol{w}^*),
\end{aligned}$$

which completes the proof. □

C Proof of Theorem 3.6

We recall the definition of KŁ property in [Attouch et al., 2013, Definition 2.4] as follows:

Definition C.1 (KŁ property)**.** *The function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is said to have the Kurdyka–Łojasiewicz (KŁ) property at $\bar{\boldsymbol{x}} \in \operatorname{dom} \partial f$ if there exists $\eta \in (0, +\infty]$, a neighborhood U of $\bar{x}$ and a continuous concave function $\phi : [0, \eta) \to \mathbb{R}_+$ such that*

- *$\phi(0) = 0$;*
- *ϕ is C^1 on $(0, \eta)$;*
- *for all $s \in (0, \eta)$, $\phi'(s) > 0$;*
- *for all $\boldsymbol{x}$ in $U \cap [f(\bar{\boldsymbol{x}}) < f < f(\bar{\boldsymbol{x}}) + \eta]$, the KŁ inequality holds*

$$\phi'(f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}})) \cdot \operatorname{dist}(0, \partial f(\boldsymbol{x})) \ge 1.$$

Now, we are ready to prove the Theorem 3.6.

Theorem 3.6. *Let $\gamma > 0, c > 0$, and*

$$\beta > 1 + \max\left\{ \frac{2\left(1 + \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\right\|^2\right)}{c\sigma^2}, \frac{2}{n\sigma^2}, \frac{4}{cn\sigma^4} \right\}.$$

For the sequence $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ generated by Algorithm 1, the following holds:

- *When $k \to \infty$, we have sequential convergence $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \to (\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ and $\boldsymbol{w}^*$ is a d-stationary point of Problem (3.1).*
- *For any $0 \le T < \infty$, we have*

$$\min_{k \in [T]} \operatorname{dist}\left(0, \partial L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\right) \le \frac{C_1}{\sqrt{T}},$$

  where $C_1 \coloneqq \tau_2 \cdot \sqrt{\frac{L_\beta(\boldsymbol{w}^0, \boldsymbol{q}^0, \boldsymbol{\lambda}^0) - F^*}{\tau_1}}$. See Section 5 for definitions of τ_1, τ_2.
- *Let $\mathcal{S}_d$ be the set of d-stationary points of Problem (3.1). There exist $C_2 < \infty$, $\rho \in [0, 1)$ and finite $\bar{k}$ such that for all $k \ge \bar{k}$:*

$$\operatorname{dist}\left(\boldsymbol{w}^k, \mathcal{S}_d\right) \le C_2 \rho^k.$$

Proof. The proof is divided into several steps.

Step 1 (limiting stationary convergence). We first prove the sequence $\{(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\}_k$ generated by Algorithm 1 will converge to a limiting stationary point $(\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ of the augmented Lagrangian L_β, which implies $\boldsymbol{w}^*$ is a limiting stationary point of Problem (3.1). The claim direct follows from Lemma 5.4.

The global rate is a direct corollary of Lemma 5.2 and Lemma 5.3. Specifically,

$$\begin{aligned}
\min_{k\in[T]} \operatorname{dist}\left(0, \partial L_\beta(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k)\right) &\le \tau_2 \cdot \min_{k\in[T]} \left\|(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) - (\boldsymbol{w}^{k-1}, \boldsymbol{q}^{k-1}, \boldsymbol{\lambda}^{k-1})\right\| && \text{(Lemma 5.3)} \\
&\le \tau_2 \cdot \sqrt{\frac{1}{T}\sum_{k=1}^{T} \left\|(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) - (\boldsymbol{w}^{k-1}, \boldsymbol{q}^{k-1}, \boldsymbol{\lambda}^{k-1})\right\|^2} \\
&\le \tau_2 \cdot \sqrt{\frac{L_\beta(\boldsymbol{w}^0, \boldsymbol{q}^0, \boldsymbol{\lambda}^0) - L_\beta(\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)}{\tau_1 T}} = \frac{C_2}{\sqrt{T}}. && \text{(Lemma 5.2)}
\end{aligned}$$

Step 2 (refine stationarity characterization). Then, we will refine the characterization of $(\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ by showing $\boldsymbol{w}^*$ is indeed a d-stationary point. We first rewrite the optimality condition for $\boldsymbol{q}$-subproblem as

$$\boldsymbol{q}^{k+1} \in \operatorname{Prox}_{\frac{1}{\beta+\gamma}g}\left(\frac{\beta}{\beta+\gamma}\left(\boldsymbol{Z}\boldsymbol{w}^k + \frac{1}{\beta}\boldsymbol{\lambda}^k\right) + \frac{\gamma}{\beta+\gamma}\boldsymbol{q}^k\right).$$

Note that, by Equation (A.1c), we observe the following identify:

$$\frac{\beta}{\beta+\gamma}\left(\boldsymbol{Z}\boldsymbol{w}^k + \frac{1}{\beta}\boldsymbol{\lambda}^k\right) + \frac{\gamma}{\beta+\gamma}\boldsymbol{q}^k = \boldsymbol{q}^{k+1} + \frac{\gamma}{\beta+\gamma}\left(\boldsymbol{q}^k - \boldsymbol{q}^{k+1}\right) + \frac{\beta}{\beta+\gamma}\boldsymbol{Z}(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}) + \frac{1}{\beta+\gamma}\boldsymbol{\lambda}^{k+1}.$$

Let $k \to +\infty$. By the convergence of $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \to (\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ in Step 1, global prox-boundedness of g from Proposition B.3, and outer semi-continuous of proximal mapping $\operatorname{Prox}_{\frac{1}{\beta+\gamma}g}$ (see [Rockafellar and Wets, 2009, Example 5.23]), we have

$$\begin{aligned}
\boldsymbol{q}^* = \lim_{k\to\infty} \boldsymbol{q}^k &\in \limsup_{k\to\infty} \operatorname{Prox}_{\frac{1}{\beta+\gamma}g}\left(\boldsymbol{q}^{k+1} + \frac{\gamma}{\beta+\gamma}\left(\boldsymbol{q}^k - \boldsymbol{q}^{k+1}\right) + \frac{\beta}{\beta+\gamma}\boldsymbol{Z}(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}) + \frac{1}{\beta+\gamma}\boldsymbol{\lambda}^{k+1}\right) \\
&\subseteq \operatorname{Prox}_{\frac{1}{\beta+\gamma}g}\left(\boldsymbol{q}^* + \frac{1}{\beta+\gamma}\boldsymbol{\lambda}^*\right).
\end{aligned}$$

By Fermat's rule [Rockafellar and Wets, 2009, Theorem 10.1] on proximal mapping $\operatorname{Prox}_{\frac{1}{\beta+\gamma}g}$, it holds $\boldsymbol{\lambda}^* \in \widehat{\partial} g(\boldsymbol{q}^*)$. Using Equations (A.1b) and (A.1c) and limiting argument, we also have $\boldsymbol{Z}\boldsymbol{w}^* = \boldsymbol{q}^*, 0 = \boldsymbol{K}\boldsymbol{w}^* + \boldsymbol{Z}^\top\boldsymbol{\lambda}^*$, which indicates $0 \in \boldsymbol{K}\boldsymbol{w}^* + \boldsymbol{Z}^\top\widehat{\partial} g(\boldsymbol{Z}\boldsymbol{w}^*)$. As $\boldsymbol{Z}$ is surjective, by chain rule [Rockafellar and Wets, 2009, Exercise 10.7] and sum rule [Rockafellar and Wets, 2009, Exercise 8.8(c)], we have $0 \in \widehat{\partial} F(\boldsymbol{w}^*)$. As F is locally Lipschitz, by [Rockafellar and Wets, 2009, Theorem 9.13], $\boldsymbol{w}^*$ is a d-stationary point.

Step 3 (linear convergence). By inspecting the analytic expression of proximal operator in Lemma 3.3, it is easy to check that, for any $\lambda > 0, \rho > 0$, $q \in \mathbb{R}$ and $q \in \operatorname{Prox}_{\lambda\phi_\rho}(v)$, where $v \in \mathbb{R}$ is arbitrary, we have

$$|q| \ge \min\left\{\frac{\lambda}{2\rho}, \rho\right\}.$$

Then, for any $k \in \mathbb{N}$, there exists a constant τ_3 with $\min\left\{\frac{1}{2\rho(\beta+\gamma)}, \rho\right\} > \tau_3 > 0$ such that

$$\min_{i\in[n]} |q_i^k| > \tau_3 > 0, \qquad \text{and} \qquad \min_{i\in[n]} |q_i^*| = \min_{i\in[n]} |\boldsymbol{z}_i^\top \boldsymbol{w}^*| > \tau_3 > 0.$$

By the convergence of $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) \to (\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*)$ in Lemma 5.4, there exists $\bar{k}$ such that for all $k \ge \bar{k}$, we have

$$\max_{i\in[n]} |q_i^k - q_i^*| \le \tau_3, \qquad \text{implying} \qquad \min_{i\in[n]} q_i^k \cdot q_i^* > \frac{\tau_3^2}{2} > 0, \qquad \text{and} \qquad \operatorname{sgn}(q_i^k) = \operatorname{sgn}(q_i^*), \quad \forall i \in [n].$$

Let the partition $[n] = \mathcal{I}^k_< \sqcup \mathcal{I}^k_= \sqcup \mathcal{I}^k_>$, where $k \in \mathbb{N} \cup \{*\}$ and

$$\mathcal{I}^k_< := \{i \in [n] : q^k_i < 0\}, \quad \mathcal{I}^k_= := \{i \in [n] : q^k_i = 0\}, \quad \mathcal{I}^k_> := \{i \in [n] : q^k_i > 0\}.$$

Then, we define the following ghost convex problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \widehat{F}(\boldsymbol{w}) := \frac{c}{2}\|\boldsymbol{w}\|^2_{\boldsymbol{K}} + \sum_{i \in \mathcal{I}^*_>} \widehat{\phi}_\rho(\boldsymbol{z}_i^\top \boldsymbol{w}), \qquad \text{where} \qquad \widehat{\phi}_\rho(u) = \max\left(0, 1 - \frac{u}{\rho}\right). \tag{CPLQ}$$

Let $\hat{g}(\boldsymbol{q}) = \sum_{i \in \mathcal{I}^*_>} \widehat{\phi}_\rho(\boldsymbol{z}_i^\top \boldsymbol{w})$. Using the same operator splitting technique $\boldsymbol{Z}\boldsymbol{w} = \boldsymbol{q} \in \mathbb{R}^n$, the augmented Lagrangian can be written as

$$\widehat{L}_\beta(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{\lambda}) := \widehat{F}(\boldsymbol{w}) + \langle \boldsymbol{\lambda}, \boldsymbol{Z}\boldsymbol{w} - \boldsymbol{q}\rangle + \frac{\beta}{2}\|\boldsymbol{Z}\boldsymbol{w} - \boldsymbol{q}\|^2.$$

Setting $\hat{\boldsymbol{w}}^0 = \boldsymbol{w}^{\bar{k}}, \hat{\boldsymbol{q}}^0 = \boldsymbol{q}^{\bar{k}}, \hat{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^{\bar{k}}, (\beta, \gamma) = (\beta, \gamma)$ in Algorithm 1, we have the following ghost convex semi-proximal ADMM algorithm:

Algorithm 2 Ghost Semi-proximal ADMM for Problem (CPLQ)

1: Set $\hat{\boldsymbol{w}}^0 = \boldsymbol{w}^{\bar{k}}, \hat{\boldsymbol{q}}^0 = \boldsymbol{q}^{\bar{k}}, \hat{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^{\bar{k}}, (\beta, \gamma) = (\beta, \gamma)$ in Algorithm 1.
2: **for** $t \in \{0, 1, 2, \dots\}$ **do**
3:

$$\begin{aligned}
\hat{\boldsymbol{q}}^{t+1} \in & \arg\min_{\hat{\boldsymbol{q}}} \hat{g}(\hat{\boldsymbol{q}}) + \frac{\beta}{2}\|\hat{\boldsymbol{q}} - \boldsymbol{Z}\hat{\boldsymbol{w}}^t - \frac{1}{\beta}\hat{\boldsymbol{\lambda}}^t\|^2 + \frac{\gamma}{2}\|\hat{\boldsymbol{q}} - \hat{\boldsymbol{q}}^t\|^2 \\
\hat{\boldsymbol{w}}^{t+1} = & \arg\min_{\hat{\boldsymbol{w}}} \hat{f}(\hat{\boldsymbol{w}}) + \frac{\beta}{2}\|\boldsymbol{Z}\hat{\boldsymbol{w}} - \hat{\boldsymbol{q}}^{t+1} + \frac{1}{\beta}\hat{\boldsymbol{\lambda}}^t\|^2 \\
\hat{\boldsymbol{\lambda}}^{t+1} = & \hat{\boldsymbol{\lambda}}^t + \beta(\boldsymbol{Z}\hat{\boldsymbol{w}}^{t+1} - \hat{\boldsymbol{q}}^{t+1}).
\end{aligned}$$

4: **end for**

We denote the sequence generated by Algorithm 2 as $\{(\hat{\boldsymbol{w}}^t, \hat{\boldsymbol{q}}^t, \hat{\boldsymbol{\lambda}}^t)\}_t$. As $\hat{g}$ is a convex piecewise linear-quadratic function, by [Rockafellar and Wets, 2009, Theorem 12.30], the KKT mapping of $\widehat{F}$ is piecewise polyhedral. Then, by the celebrated results of Robinson [1981], we have the inverse KKT mapping is calm (see also [Dontchev and Rockafellar, 2009, Proposition 3H.1]). Besides, by the strong convexity of $\boldsymbol{w}$-subproblem in the proof of Lemma 5.2, we have $c\boldsymbol{K} + \beta\boldsymbol{Z}^\top\boldsymbol{Z} \succ \mathbb{0}$. Using [Han et al., 2018, Theorem 2], there exists $\rho \in (0, 1)$ and constant $\tau_4 \geq 0$ such that

$$\operatorname{dist}\left((\hat{\boldsymbol{w}}^k, \hat{\boldsymbol{q}}^k, \hat{\boldsymbol{\lambda}}^k), \mathcal{S}_{\mathrm{KKT}}\right) \leq \tau_4 \cdot \rho^k,$$

where $\mathcal{S}_{\mathrm{KKT}}$ is the set KKT points for Equation (CPLQ).

Then, we will show for any $k \in \left(\mathbb{N} \backslash [\bar{k} - 1]\right) \cup \{*\}$, $(\boldsymbol{w}^k, \boldsymbol{q}^k, \boldsymbol{\lambda}^k) = (\hat{\boldsymbol{w}}^{k-\bar{k}}, \hat{\boldsymbol{q}}^{k-\bar{k}}, \hat{\boldsymbol{\lambda}}^{k-\bar{k}})$, which directly indicates the linear convergence. The proof is by induction on k. For $k = \bar{k}$, it is by definition the claim holds. Then, we assume for $\{\bar{k}, \dots, k\}$ the claim holds. Then we need to show $(\boldsymbol{w}^{k+1}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^{k+1}) = (\hat{\boldsymbol{w}}^{k-\bar{k}+1}, \hat{\boldsymbol{q}}^{k-\bar{k}+1}, \hat{\boldsymbol{\lambda}}^{k-\bar{k}+1})$. For $\boldsymbol{q}^{k+1} = \hat{\boldsymbol{q}}^{k-\bar{k}+1}$, we consider the optimality condition Equation (A.1a) of the $\boldsymbol{q}$-subproblem in Algorithm 1. By $\min_{i \in [n]} |q^{k+1}_i| > \tau_3$, [Rockafellar and Wets, 2009, Proposition 10.5], [Rockafellar and Wets, 2009, Theorem 10.49], we have

$$\widehat{\partial} g(\boldsymbol{q}^{k+1}) = \bigoplus_{i=1}^n \partial \phi_\rho(q^{k+1}_i) = \bigoplus_{i=1}^n \delta_{\mathcal{I}^{k+1}_>}(i) \cdot \partial \widehat{\phi}_\rho(q^{k+1}_i) = \partial \hat{g}(\boldsymbol{q}^{k+1}),$$

where $\delta_C(i) := 1$ if $i \in C$ otherwise $\delta_C(i) := 0$, and we use $\mathcal{I}^k_> = \mathcal{I}^*_>$ for all $k \geq \bar{k}$. Then, it is clear that $\boldsymbol{q}^{k+1}$ satisfies the optimality condition of the $\boldsymbol{q}$-subproblem of Algorithm 2, by whose strong convexity, we have $\boldsymbol{q}^{k+1} = \hat{\boldsymbol{q}}^{k+1}$. Similarly, we can show by examining the optimality condition that $\boldsymbol{w}^{k+1} = \hat{\boldsymbol{w}}^{k+1}$ and $\boldsymbol{\lambda}^{k+1} = \hat{\boldsymbol{\lambda}}^{k+1}$. Taking limit we have $(\boldsymbol{w}^*, \boldsymbol{q}^*, \boldsymbol{\lambda}^*) = (\hat{\boldsymbol{w}}^*, \hat{\boldsymbol{q}}^*, \hat{\boldsymbol{\lambda}}^*)$. The final building block is to show if $(\hat{\boldsymbol{w}}^*, \hat{\boldsymbol{q}}^*, \hat{\boldsymbol{\lambda}}^*) \in \mathcal{S}_{\mathrm{KKT}}$ then

$\hat{\boldsymbol{w}}^* \in \mathcal{S}_{\mathrm{d}}$, which indicates $\operatorname{dist}\left(\hat{\boldsymbol{w}}^k, \mathcal{S}_{\mathrm{d}}\right) \leq \operatorname{dist}\left(\hat{\boldsymbol{w}}^k, \Pi_1(\mathcal{S}_{\mathrm{KKT}})\right) \leq \operatorname{dist}\left((\hat{\boldsymbol{w}}^k, \hat{\boldsymbol{q}}^k, \hat{\boldsymbol{\lambda}}^k), \mathcal{S}_{\mathrm{KKT}}\right)$. To this end, for any KKT point $(\hat{\boldsymbol{w}}^*, \hat{\boldsymbol{q}}^*, \hat{\boldsymbol{\lambda}}^*)$, we write down the KKT system of Problem (CPLQ) as

$$\left\{\begin{array}{lcl} 0 & = & c\boldsymbol{K}\hat{\boldsymbol{w}}^* + \boldsymbol{Z}^\top \hat{\boldsymbol{\lambda}}^*, \\ 0 & \in & -\hat{\boldsymbol{\lambda}}^* + \partial \hat{g}(\hat{\boldsymbol{q}}^*), \\ 0 & = & \boldsymbol{Z}\hat{\boldsymbol{w}}^* - \hat{\boldsymbol{q}}^*. \end{array}\right.$$

Then we have $0 \in c\boldsymbol{K}\hat{\boldsymbol{w}}^* + \boldsymbol{Z}^\top \partial \hat{g}(\boldsymbol{Z}\hat{\boldsymbol{w}}^*)$, where $\hat{g}$ is convex. Therefore, the claim follows from $\min_{i\in[n]} |q_i^*| > \tau_3$, [Rockafellar and Wets, 2009, Proposition 10.5], [Rockafellar and Wets, 2009, Theorem 10.49], which implies $\widehat{\partial} g(\boldsymbol{Z}\hat{\boldsymbol{w}}^*) = \widehat{\partial} g(\hat{\boldsymbol{q}}^*) = \partial \hat{g}(\hat{\boldsymbol{q}}^*) = \partial \hat{g}(\boldsymbol{Z}\hat{\boldsymbol{w}}^*)$. Thus we have $\hat{\boldsymbol{w}}^* \in \mathcal{S}_{\mathrm{d}}$, which completes the proof. □

D More Example

Counterexample. Let $d = 1, n = 6, c = 1, \rho = 1$. We consider the following data points multiset:

$$\mathbb{S} := \Big\{(x_i, y_i) : i \in [n]\Big\} = \Big\{(1,+1),(1,+1),(-1,-1),(-1,-1),(0,+1),(0,-1)\Big\} \subseteq \mathbb{R}^d \times \{+1,-1\}.$$

We claim that $\boldsymbol{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is a global minimizer for Problem (3.1) on $\{(x_i, y_i) : i \in [n]\}$. However, $\boldsymbol{w}$ does not satisfy the condition in [Suzumura et al., 2017, Theorem 4]. For the proof, we reorganized the data points into

$$\boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}, \qquad \text{with } \operatorname{rank}(\boldsymbol{Z}) = d + 1 = 2 < n.$$

Let $\phi_1(t) = \min\{1, \max\{1 - t, 0\}\}$. Then, Problem (1.1) on $\mathbb{S}$ reads

$$\nu := \min_{\theta\in\mathbb{R}, b\in\mathbb{R}} f(\theta, b) := \frac{\theta^2}{2} + 2\phi_1(\theta + b) + 2\phi_1(\theta - b) + \phi_1(b) + \phi_1(-b). \tag{CE}$$

To get an optimizer of Problem (CE), we consider the following 5 cases:

- $b = 0$
  - $\theta = 0$: $\nu_1 := \inf_\theta f(\theta, b) = 6$.
  - $\theta \neq 0$: $\nu_2 := \inf_\theta f(\theta, b) = 2 + \left(\inf_\theta \frac{\theta^2}{2} + 4\phi_1(\theta)\right) = \frac{5}{2}$ when $\theta = 1$.
- $b \neq 0$
  - $|\theta| = |b|$: by symmetry of b, we assume $\theta = b$. Then, Problem (CE) can be rewritten as

    $$\min_{b\neq 0} \frac{b^2}{2} + 2\phi_1(2b) + 2 + \phi_1(b) + \phi_1(-b).$$

    By monotonicity of ϕ_1, we can assume $b > 0$. Let the convex hinge loss be $h(t) = \max\{1 - t, 0\}$. Then, we get the following convex reformulation

    $$\nu_3 := \frac{7}{2} = \inf_{b>0} \frac{b^2}{2} + 2h(2b) + 2 + h(b) + 1.$$

  - $|\theta| \neq |b|$: by monotonicity of g and symmetry of b, we assume $w \geq 0$ and $b > 0$.
    - $|\theta| > |b|$:

      $$\nu_4 := \frac{5}{2} = \inf_{\theta\geq 0, b>0} \frac{\theta^2}{2} + 2h(\theta + b) + 2h(\theta - b) + h(b) + 1.$$

⋆ $|\theta| < |b|$:

$$\nu_5 := 3 = \inf_{\theta \geq 0, b > 0} \frac{\theta^2}{2} + 2h(\theta + b) + 2 + h(b) + 1.$$

In summary, we claim $\theta^* = 1, b^* = 0$ is an optimal solution of Problem (CE) as $f(1,0) = \frac{5}{2} = \min\{\nu_i\}_{[5]} \leq \nu$. In that case, let $\boldsymbol{q} := \boldsymbol{y}\mathbb{1}_d \odot (\boldsymbol{X}\theta^* - b^*\boldsymbol{y}) = [1\ 1\ 1\ 1\ 0\ 0]^\top$. On the other hand, in [Suzumura et al., 2017, Theorem 4], the KKT-type condition requires $q_i \neq 0$ for all $i \in [n]$, which is not satisfied by a global minimizer $\theta^* = 1, b^* = 0$. Thus, [Suzumura et al., 2017, Theorem 4] is not necessary. Besides, we note that above example provide an explicit case that if $\boldsymbol{Z}$ is not surjective, then the dual characterization in Proposition 3.2 may not hold. In the language of variational analysis, that is because $0 \in \widehat{\partial}\big(f(\boldsymbol{Z}\cdot)\big)(\boldsymbol{w}^*)$ whereas $\widehat{\partial}(f)(\boldsymbol{Z}\boldsymbol{w}^*) = \emptyset$.

E DCA-SpADMM Hybrid Algorithm

The hybrid method (Algorithm 3) runs DCA until convergence and then starts SpADMM from that DC-critical point. The detailed hybrid scheme can be summarized as follows:

Algorithm 3 Hybrid Semi-proximal ADMM and DCA for Problem (3.1)

Input: $\boldsymbol{Z} \in \mathbb{R}^{(d+1)\times n}, \boldsymbol{w}^0$, choose $\gamma > 0, c > 0$ and $\beta > 1 + \max\left\{\frac{2\left(1+\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\right\|^2\right)}{c\sigma^2}, \frac{2}{n\sigma^2}, \frac{4}{cn\sigma^4}\right\}$.

Set $r = 0$.
for all $k \in \{0, 1, 2, \dots\}$ **do**
  if $r = 0$ **then**
    Set $\boldsymbol{g}^k = -\frac{1}{\rho}\boldsymbol{Z}^\top \operatorname{sgn}\left(-\frac{1}{\rho}\boldsymbol{Z}^\top \boldsymbol{w}^k\right)$.
    Compute

$$\boldsymbol{w}^{k+1} \in \operatorname*{Arg\,min}_{\boldsymbol{w}} \frac{c}{2}\|\boldsymbol{w}\|_{\boldsymbol{K}}^2 + \sum_{i=1}^n \max\left(1 - \frac{\boldsymbol{z}_i^\top \boldsymbol{w}}{\rho}, 0\right) - \boldsymbol{w}^\top \boldsymbol{g}_k.$$

    if $\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\| \leq 10^{-5}$ **then**
      Set $r = 1, \boldsymbol{q}^{k+1} = \boldsymbol{Z}\boldsymbol{w}^{k+1}, \boldsymbol{\lambda}^{k+1} = 0$.
    end if
  else

$$\begin{aligned}
\boldsymbol{q}^{k+1} &\in \operatorname*{Arg\,min}_{\boldsymbol{q}} L_\beta(\boldsymbol{w}^k, \boldsymbol{q}, \boldsymbol{\lambda}^k) + \frac{\gamma}{2}\|\boldsymbol{q} - \boldsymbol{q}^k\|^2,\\
\boldsymbol{w}^{k+1} &= \operatorname*{arg\,min}_{\boldsymbol{w}} L_\beta(\boldsymbol{w}, \boldsymbol{q}^{k+1}, \boldsymbol{\lambda}^k),\\
\boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \beta(\boldsymbol{Z}\boldsymbol{w}^{k+1} - \boldsymbol{q}^{k+1}).
\end{aligned}$$

  end if
end for

F More Numerical Results

F.1 Objective Value Curve

In this section, we provide extra numerical results in the setting of Section 6.

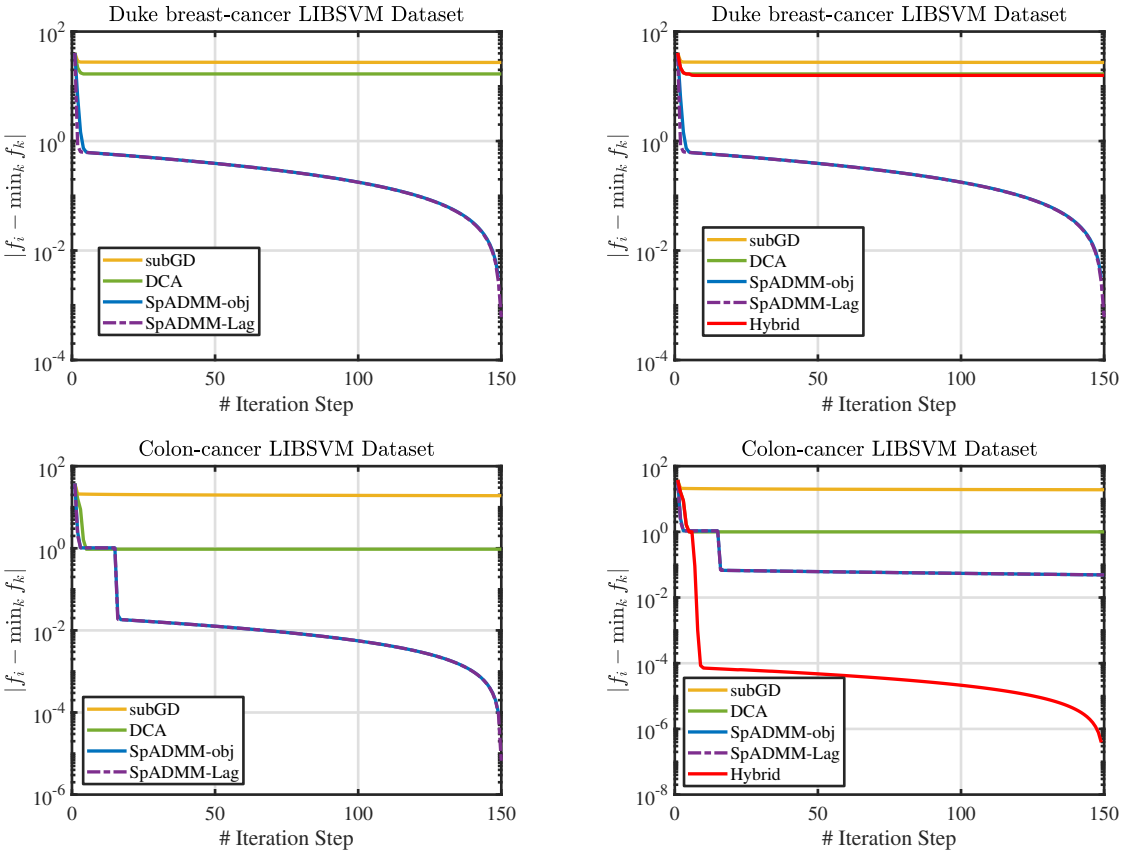

Figure 5: Performance of subGD (subgradient), DCA, SpADMM (Algorithm 1), and Hybrid Algorithm 3 on Real-world Data. The min in log-scaled y-axis is taken over all aforementioned algorithms.

F.2 Wall-Clock Time

To demonstrate the efficiency of various algorithms in the numerical experiments, we record the wall-clock time of running these algorithms for 10^2 iterative steps in Table 2.

Table 2: Wall-Clock Time of subGD (subgradient), DCA, SpADMM (Algorithm 1), and Hybrid Algorithm 3.

10^2 steps	Wall-Clock Time (s)			
	Toy	Leukemia	Duke	Colon
subGD	0.0173	24.0051	23.9875	1.7640
DCA	19.6547	60.6936	72.4713	42.5783
SpADMM	**0.0138**	**1.1031**	**1.1090**	**0.0828**
Hybrid	0.7206	2.9015	3.9245	1.9675