VARIABLE SMOOTHING ALGORITHM FOR INNER-LOOP-FREE DC COMPOSITE OPTIMIZATIONS

Kumataro Yazawa, Keita Kume, Isao Yamada
Dept. of Information and Communications engineering, Institute of Science Tokyo
Email: {yazawa,kume,isao}@sp.ict.e.titech.ac.jp

***Abstract*—We propose a variable smoothing algorithm for minimizing a nonsmooth and nonconvex cost function. The cost function is the sum of a smooth function and a composition of a difference-of-convex (DC) function with a smooth mapping. At each step of our algorithm, we generate a smooth surrogate function by using the Moreau envelope of each weakly convex function in the DC function, and then perform the gradient descent update of the surrogate function. Unlike many existing algorithms for DC problems, the proposed algorithm does not require any inner loop. We also present a convergence analysis in terms of a DC critical point for the proposed algorithm as well as its application to robust phase retrieval.**

I. Introduction

In this paper, we consider the following nonsmooth and nonconvex optimization problem.

Problem I.1 (DC composite type problem)**.**

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\ F(\boldsymbol{x}) := h(\boldsymbol{x}) + \underbrace{(f-g)}_{\varphi}\circ\mathfrak{S}(\boldsymbol{x}), \tag{1}$$

where

(a) $h:\mathbb{R}^d\to\mathbb{R}$ is differentiable and its gradient $\nabla h:\mathbb{R}^d\to\mathbb{R}^d$ is Lipschitz continuous, i.e., there exists $L_{\nabla h}>0$ such that $\|\nabla h(\boldsymbol{x})-\nabla h(\boldsymbol{y})\|\le L_{\nabla h}\|\boldsymbol{x}-\boldsymbol{y}\|$ $(\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d)$;

(b) $f:\mathbb{R}^n\to\mathbb{R}$ and $g:\mathbb{R}^n\to\mathbb{R}$ are
   (i) Lipschitz continuous (possibly not differentiable),
   (ii) weakly-convex, i.e., there exist $\eta_f,\eta_g>0$ such that $f+\frac{\eta_f}{2}\|\cdot\|^2$ and $g+\frac{\eta_g}{2}\|\cdot\|^2$ are convex (We define $\eta:=\max\{\eta_f,\eta_g\}$ for convenience),
   (iii) prox-friendly, i.e, *their proximity operators* (see Definition II.4) are available as computable operators

   (see Table I for various examples of $f-g$ in applications);

(c) $\mathfrak{S}:\mathbb{R}^d\to\mathbb{R}^n$ is differentiable, and its Fréchet derivative $\mathrm{D}\mathfrak{S}:\mathbb{R}^d\to\mathbb{R}^{n\times d}$ is Lipschitz continuous[1];

(d) F is bounded below, i.e., $\inf_{\boldsymbol{x}\in\mathbb{R}^d}F(\boldsymbol{x})>-\infty$.

This work was supported partially by JSPS Grants-in-Aid (19H04134, 24K23885).

[1]Problem I.1 covers seemingly much more general case of the minimization of $h+\hat{f}\circ\mathfrak{S}_1-\hat{g}\circ\mathfrak{S}_2$, where $\hat{f}:\mathbb{R}^{n_1}\to\mathbb{R}$ and $\hat{g}:\mathbb{R}^{n_2}\to\mathbb{R}$ are Lipschitz continuous, weakly convex and prox-friendly, and $\mathfrak{S}_1:\mathbb{R}^d\to\mathbb{R}^{n_1}$ and $\mathfrak{S}_2:\mathbb{R}^d\to\mathbb{R}^{n_2}$ are continuously differentiable such that their Fréchet derivative are Lipschitz continuous. This fact can be understood through a simple translation of this minimization into Problem I.1 by introducing $f:\mathbb{R}^{n_1+n_2}\to\mathbb{R}:[\boldsymbol{z}_1^T,\boldsymbol{z}_2^T]^T\mapsto\hat{f}(\boldsymbol{z}_1)$, $g:\mathbb{R}^{n_1+n_2}\to\mathbb{R}:[\boldsymbol{z}_1^T,\boldsymbol{z}_2^T]^T\mapsto\hat{g}(\boldsymbol{z}_2)$, and $\mathfrak{S}:\mathbb{R}^d\to\mathbb{R}^{n_1+n_2}:\boldsymbol{x}\mapsto[\mathfrak{S}_1(\boldsymbol{x})^T\mathfrak{S}_2(\boldsymbol{x})^T]^T$.

TABLE I: φ in application areas and their DC decomposition

name	$\varphi(\boldsymbol{z})=(f-g)(\boldsymbol{z})$	$f(\boldsymbol{z})$	$g(\boldsymbol{z})$
ℓ_1 norm	$\sum_{i=1}^n \lvert[\boldsymbol{z}]_i\rvert$	$\sum_{i=1}^n \lvert[\boldsymbol{z}]_i\rvert$	0
MCP [8]	$\sum_{i=1}^n r([\boldsymbol{z}]_i)$ *1	$\sum_{i=1}^n r([\boldsymbol{z}]_i)$	0
Capped ℓ_1 [10]	$\sum_{i=1}^n \min\{\lvert[\boldsymbol{z}]_i\rvert,\beta\}$ (with $\beta\in\mathbb{R}_{++}$)	$\sum_{i=1}^n \lvert[\boldsymbol{z}]_i\rvert$	$\sum_{i=1}^n \max\{\lvert[\boldsymbol{z}]_i\rvert-\beta,0\}$
Trimmed ℓ_1 [11]	$\sum_{i=K+1}^n \lvert[\boldsymbol{z}]_{\downarrow i}\rvert$ *2 (with $0\le K\le n-1$)	$\sum_{i=1}^n \lvert[\boldsymbol{z}]_i\rvert$	$\sum_{i=1}^K \lvert[\boldsymbol{z}]_{\downarrow i}\rvert$

*1 $r(t):=\begin{cases}\lambda|t|-\frac{t^2}{2\beta} & |t|\le\beta\lambda,\\ \frac{\beta\lambda^2}{2} & \text{otherwise}\end{cases}$ with $\lambda,\beta\in\mathbb{R}_{++}$.

*2 $[\boldsymbol{z}]_{\downarrow i}$ denotes the entry of $\boldsymbol{z}$ whose absolute value is the i-th largest.

The function $f-g$ in (1) is called "Difference-of-convex (DC) function"[2], and thus, we call Problem I.1 "DC composite type problem".

Problem I.1 appears mainly in sparsity-aware signal processing applications, such as image restoration [1], trend filtering [2], compressed sensing [3], and sparse logistic regression [4]. In addition, Problem I.1 also arises in robust estimation including robust phase retrieval [5], [6]. The DC function $f-g$ in (1) is employed, e.g, to induce sparsity of the target signal translated by $\mathfrak{S}$ in sparse signal processing, or to enhance the robustness of the data fidelity to measurement outliers in robust estimation. Such DC functions $f-g$ include (i) convex functions (e.g., ℓ_1 norm [7, Exm.24.22]), (ii) weakly convex functions (e.g., the minimax concave penalty (MCP) [8] and the smoothly clipped absolute deviation (SCAD) [9]), and (iii) DC functions that are not weakly convex (e.g., capped ℓ_1 norm [10], the trimmed ℓ_1 norm [11] and its variant [12]), some of which are summarized in Table I.

In particular, DC functions lacking weak convexity (referred to as "inherently DC" functions in this paper), such as the capped ℓ_1 norm and the trimmed ℓ_1 norm, have been attracting great attention. Indeed, a model using the capped ℓ_1 norm has been reported to effectively reduce the impact of outliers in application of *twin support vector machine* [13]. Moreover, in *robust principal component analysis*, which aims to decompose a matrix into a low rank matrix and a sparse

[2]Although f and g are weakly convex, $f-g$ is actually DC function because it can be expressed as the difference of convex functions $\tilde{f}:=f+\frac{\eta}{2}\|\cdot\|^2$ and $\tilde{g}:=g+\frac{\eta}{2}\|\cdot\|^2$. For this reason, one might think that we could just assume f,g in Problem I.1 to be convex functions instead of weakly convex functions in the first place. However, components of this naive DC decomposition $\tilde{f}$ and $\tilde{g}$ do not satisfy the assumption (b)(i) in Problem I.1, namely Lipschitz continuity. Thus, we assume a weaker condition, i.e., weak convexity, than convexity on f and g.

matrix, a model using the capped ℓ_1 norm in [14] outperforms a model using ℓ_1 norm because the capped ℓ_1 norm avoids over-penalizing a matrix that contains entries of large absolute values, unlike the ℓ_1 norm. On the other hand, the trimmed ℓ_1 norm has been utilized in terms of an exact penalty theory for cardinality-constrained optimization problems [15], i.e.,

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\ h(\boldsymbol{x}) \quad \text{s.t.} \quad \|\mathfrak{S}(\boldsymbol{x})\|_0 \leq K, \tag{2}$$

where $\|\cdot\|_0$ counts the number of non-zero entries of a given vector. More precisely, if the trimmed ℓ_1 norm is used as $f - g$, a global (resp. local) minimizer of Problem I.1 also serves as a global (resp. local) minimizer of the cardinality-constrained problem (2) under certain conditions [15]. Such an exact penalty formulation via Problem I.1 seems to be more tractable than the cardinality-constrained problem (2) from a viewpoint of designing algorithms.

A commonly used existing approach to minimization of DC functions is DC algorithm (DCA) (see, e.g, [16]). At each iteration of DCA, a subtrahend convex function in a DC function is replaced with an affine minorization by utilizing its subgradient, and then minimizes the resulting surrogate function as a subproblem.For Problem I.1, *DC composite algorithm* (DCCA) [17], which is an extension of DCA, can be employed. If an exact solution to the subproblem in DCCA is available, then DCCA has a convergence guarantee in terms of *DC critical point* (see Definition II.2). In practice, however, DCCA requires infinite iterations of an inner loop so as to find the exact solution of the subproblem[3]. The convergence analysis of DCCA does not cover realistic cases where only inexact solutions of the subproblem are available.

In this paper, we propose a variable smoothing algorithm for Problem I.1 that does not require any inner loop for the subproblem. Our algorithm (Algorithm 1) is designed as a gradient descent update of a time-varying smoothed surrogate function of F in (1). With *the Moreau envelopes* (see Definition II.4) ${}^{\mu}f$ of f and ${}^{\mu}g$ of g, the proposed surrogate function is given as $h+({}^{\mu_k}f-{}^{\mu_k}g)\circ\mathfrak{S}$, where $(\mu_k)_{k=1}^{\infty}\subset\mathbb{R}$ is a monotonically decreasing sequence of convergence to zero. We also present an asymptotic convergence analysis of the proposed algorithm in the sense of a DC critical point (see Theorem III.6 and Remark III.7). To verify effectiveness of the proposed model (i.e, Problem I.1) and the proposed algorithm, we conduct numerical experiments in a scenario of the robust phase retrieval (e.g., [5], [6]) with its new optimization model.

Related works. Our algorithm serves as an extension of algorithms [20], [21], [22] proposed for Problem I.1 in a special case where $g \equiv 0$ (more precisely, $\mathfrak{S}$ is linear in [20]). The existing algorithms in [20], [21], [22] can be applied only to optimization problems involving weakly convex functions, while the proposed algorithm can cover even inherently DC functions.

Notation. $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_{++}$ denote respectively the sets of all positive integers, all real numbers and all positive real number. $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ are respectively the Euclidean norm and the standard inner product. For $\boldsymbol{v}\in\mathbb{R}^n$, $[\boldsymbol{v}]_i\in\mathbb{R}$ stands for the i-th entry. We use Id to denote the identity mapping. For Euclidean spaces $\mathcal{X},\mathcal{Y}$ and a continuously differentiable mapping $J:\mathcal{X}\to\mathcal{Y}$, its Fréchet derivative at $\boldsymbol{x}\in\mathcal{X}$ is the linear operator $\mathrm{D}J(\boldsymbol{x}):\mathcal{X}\to\mathcal{Y}$ such that $\lim_{\mathcal{X}\setminus\{\boldsymbol{0}\}\ni\boldsymbol{h}\to\boldsymbol{0}}\frac{J(\boldsymbol{x}+\boldsymbol{h})-J(\boldsymbol{x})-DJ(\boldsymbol{x})[\boldsymbol{h}]}{\|\boldsymbol{h}\|}=0$. In particular with $\mathcal{Y}=\mathbb{R}$, $\nabla J:\mathcal{X}\to\mathcal{X}$ is called the gradient of J if $\nabla J(\boldsymbol{x})\in\mathcal{X}$ at $\boldsymbol{x}\in\mathcal{X}$ satisfies $\mathrm{D}J(\boldsymbol{x})[\boldsymbol{v}]=\langle\nabla J(\boldsymbol{x}),\boldsymbol{v}\rangle$ ($\boldsymbol{v}\in\mathcal{X}$).

[3]For a special case of Problem I.1 where $\mathfrak{S} = \mathrm{Id}$, a variant of DCA introduced in [18] does not require any inner loop for the subproblem because its exact solution can be obtained by using the proximity operator of f. Moreover, another DCA-type algorithm proposed by [19] has a convergence guarantee even though an inexact solution of the subproblem is used, where such an inexact solution can be obtained in finite steps.

II. Preliminary

As an extension of the subdifferential of convex functions, we use the following subdifferential of nonconvex functions. (see, e.g., a recent survey [23] for readers who are unfamiliar with nonsmooth analysis).

Definition II.1 (Regular subdifferential [24, Def. 8.3])**.** For a function $\phi:\mathbb{R}^d\to\mathbb{R}$, *the regular subdifferential* of ϕ at $\bar{\boldsymbol{x}}\in\mathbb{R}^d$, denoted as $\partial\phi(\bar{\boldsymbol{x}})\subset\mathbb{R}^d$, is the set of all vectors $\boldsymbol{v}\in\mathbb{R}^n$ such that

$$\lim_{\delta\searrow 0}\ \inf_{0<\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|<\delta}\frac{\phi(\boldsymbol{x})-\phi(\bar{\boldsymbol{x}})-\langle\boldsymbol{v},\boldsymbol{x}-\bar{\boldsymbol{x}}\rangle}{\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|}\geq 0.$$

If ϕ is convex, this regular subdifferential is equivalent to the convex subdifferential [24, Proposition 8.12]. Furthermore, if ϕ is Fréchet differentiable at $\bar{\boldsymbol{x}}$, $\partial\phi(\bar{\boldsymbol{x}})=\{\nabla\phi(\bar{\boldsymbol{x}})\}$ holds [24, Exercise 8.8(a)].

Unfortunately, finding a global minimizer of Problem I.1 is not realistic due to the severe nonconvexity of F. Instead, in this paper, we focus on finding a DC critical point defined, with the regular subdifferentials, as follows.

Definition II.2 (DC critical point for Problem I.1 [17])**.** A point $\boldsymbol{x}^\star\in\mathbb{R}^d$ is said to be a *DC critical point* for Problem I.1 if

$$\partial(h+f\circ\mathfrak{S})(\boldsymbol{x}^\star)-\partial(g\circ\mathfrak{S})(\boldsymbol{x}^\star)\ni\boldsymbol{0}. \tag{3}$$

Lemma II.3 (Relationship between local minimizer and DC critical point)**.** Let $\boldsymbol{x}^\star\in\mathbb{R}^d$ be a local minimizer of F in Problem I.1. Then, $\boldsymbol{x}^\star$ is a DC critical point for Problem I.1.

From Lemma II.3, being a DC critical point is a necessary condition for being a local minimizer. Moreover, finding such a DC critical point has been used as an acceptable goal in many DC optimization literature [16], [17], [18], [19], [25].

The Moreau envelope plays an important role in this paper for designing the proposed algorithm.

Definition II.4 (Moreau envelope, proximity operator [20])**.** Let $\psi:\mathbb{R}^n\to\mathbb{R}$ be an η_ψ-weakly convex function with $\eta_\psi>0$. Its Moreau envelope and proximity operator at $\bar{\boldsymbol{z}}\in\mathbb{R}^n$ with $\mu\in(0,\eta_\psi^{-1})$ are respectively defined as

$${}^{\mu}\psi(\bar{\boldsymbol{z}}):=\min_{\boldsymbol{z}\in\mathbb{R}^n}\left\{\psi(\boldsymbol{z})+\frac{1}{2\mu}\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|^2\right\},$$

$$\mathrm{prox}_{\mu\psi}(\bar{\boldsymbol{z}}):=\underset{\boldsymbol{z}\in\mathbb{R}^n}{\mathrm{argmin}}\left\{\psi(\boldsymbol{z})+\frac{1}{2\mu}\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|^2\right\},$$

where $\mathrm{prox}_{\mu\psi}$ is single-valued due to the strong convexity of $\psi+(2\mu)^{-1}\|\cdot-\bar{\boldsymbol{z}}\|^2$.

The Moreau envelope ${}^{\mu}\psi$ serves as a smoothed surrogate function of ψ because of the next properties.

Algorithm 1 Variable smoothing algorithm for DC composite type problem (Problem I.1)

Input: $\boldsymbol{x}_1 \in \mathbb{R}^d, (\mu_k)_{k=1}^{\infty} \subset (0,(2\eta)^{-1}]$ enjoying (6).
1: **for** $k = 1,2,3,\dots$ **do**
2: Set $F_k := F^{\langle \mu_k \rangle} = h + ({}^{\mu_k}f - {}^{\mu_k}g) \circ \mathfrak{S}$
3: Obtain γ_k by Algorithm 2
4: $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \gamma_k \nabla F_k(\boldsymbol{x}_k)$
5: **end for**

Algorithm 2 Backtracking algorithm to find γ_k

Input: $\gamma_{\text{initial}} > 0,\ \rho \in (0,1),\ c \in (0,1)$
1: $\gamma_k \leftarrow \gamma_{\text{initial}}$
2: **while** $F_k(\boldsymbol{x}_k - \gamma_k \nabla F_k(\boldsymbol{x}_k)) > F_k(\boldsymbol{x}_k) - c\gamma_k \|\nabla F_k(\boldsymbol{x}_k)\|^2$ **do**
3: $\gamma_k \leftarrow \rho\gamma_k$
4: **end while**
Output: γ_k

Fact II.5 (Properties of Moreau envelope)**.** Let $\psi : \mathbb{R}^n \to \mathbb{R}$ be an η_ψ-weakly convex function with $\eta_\psi > 0$. For $\mu \in (0, \eta_\psi^{-1})$, the following hold.

(a) [24, Theorem 1.25] $(\boldsymbol{z} \in \mathbb{R}^n)\ \lim_{\mu \searrow 0} {}^{\mu}\psi(\boldsymbol{z}) = \psi(\boldsymbol{z})$.
(b) [26, Collorary 3.4] ${}^{\mu}\psi$ is continuously differentiable with $(\boldsymbol{z} \in \mathbb{R}^n)\ \nabla {}^{\mu}\psi(\boldsymbol{z}) = \mu^{-1}\big(\boldsymbol{z} - \operatorname{prox}_{\mu\psi}(\boldsymbol{z})\big)$.
(c) [26, Collorary 3.4] $\nabla {}^{\mu}\psi$ is Lipschitz continuous with $L_{\nabla {}^{\mu}\psi} := \max\{\mu^{-1}, \frac{\eta_\psi}{1-\eta_\psi\mu}\}$.

Note that for f and g in Problem I.1, we can compute $\nabla {}^{\mu}f$ and $\nabla {}^{\mu}g$ in closed forms because these functions are assumed to be prox-friendly (see the assumption (b)(iii) of Problem I.1).

III. Variable Smoothing Algorithm for DC Composite Problem

A. Design of Smooth Surrogate Function

In our algorithm, we use the following function as a smooth surrogate function of F in place of the direct utilization of the nonsmooth function F.

Definition III.1 (Surrogate function)**.** Consider Problem I.1. For $\mu \in (0, \eta^{-1})$, we define a surrogate function of the cost function F in Problem I.1 as

$$F^{\langle \mu \rangle} := h + ({}^{\mu}f - {}^{\mu}g) \circ \mathfrak{S}. \tag{4}$$

By Fact II.5(b), $F^{\langle \mu \rangle}$ is also continuously differentiable.

The next theorem suggests how to find a DC critical point in (3) using the surrogate function $F^{\langle \mu \rangle}$.

Theorem III.2. Consider Problem I.1. Suppose that a positive sequence $(\mu_k)_{k=1}^{\infty} \subset (0, \eta^{-1})$ converges to 0. For the function sequence $F_k := F^{\langle \mu_k \rangle}$ $(k \in \mathbb{N})$ with (4) and any convergent sequence $(\boldsymbol{x}_k)_{k=1}^{\infty} \subset \mathbb{R}^d \to \exists \bar{\boldsymbol{x}} \in \mathbb{R}^d$, we have

$$\operatorname{dist}(\boldsymbol{0}, \partial(h + f\circ\mathfrak{S})(\bar{\boldsymbol{x}}) - \partial(g\circ\mathfrak{S})(\bar{\boldsymbol{x}})) \le \liminf_{k\to\infty} \|\nabla F_k(\boldsymbol{x}_k)\|, \tag{5}$$

where $\operatorname{dist}(\boldsymbol{v}, S) := \inf_{\boldsymbol{w}\in S}\|\boldsymbol{v} - \boldsymbol{w}\|$ for a point $\boldsymbol{v} \in \mathbb{R}^d$ and a set $S \subset \mathbb{R}^d$.

Theorem III.2 implies that $\bar{\boldsymbol{x}}$ is a DC critical point in the sense of (3) if the right hand side of (5) is zero. Hence, our goal of finding a DC critical point of Problem I.1 is reduced to designing an algorithm to generate a point sequence $(\boldsymbol{x}_k)_{k=1}^{\infty}$ such that $\liminf_{k\to\infty}\|\nabla F_k(\boldsymbol{x}_k)\| = 0$.

B. Proposed Algorithm and Its Convergence Analysis

We propose Algorithm 1 based on the gradient descent method of the smoothed surrogate function $F_k := F^{\langle \mu_k \rangle}$.

We design $(\mu_k)_{k=1}^{\infty} \subset (0,(2\eta)^{-1}]$ to satisfy the following condition (introduced in [22]) so as to establish a convergence analysis of Algorithm 1:

$$\begin{cases} \text{(i) } \lim_{k\to\infty}\mu_k = 0, \quad \text{(ii) } \sum_{k=1}^{\infty}\mu_k = \infty, \\ \text{(iii) } (\exists M \ge 1, \forall k \in \mathbb{N})\ 1 \le \mu_k/\mu_{k+1} \le M. \end{cases} \tag{6}$$

For example, $\mu_k := (2\eta)^{-1}k^{-\frac{1}{\alpha}}$ with $\alpha \ge 1$ enjoys the condition (6) ($\alpha = 3$ is reported to be an appropriate value for a reasonable convergence rate of a special case of Algorithm 1 with $g \equiv 0$ [20], [22]).

To obtain a stepsize γ_k in line 3 of Algorithm 1, we employ the so-called *backtracking algorithm* in Algorithm 2 which has been utilized as a standard stepsize selection for smooth optimization (see, e.g., [27]). The finite termination of Algorithm 2 is guaranteed under the following assumption.

Assumption III.3 (Decent assumption)**.** Consider the surrogate function $F^{\langle \mu \rangle}$ in (4) with $\mu \in (0,(2\eta)^{-1}]$. Assume that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$F^{\langle \mu \rangle}(\boldsymbol{y}) \le F^{\langle \mu \rangle}(\boldsymbol{x}) + \langle \nabla F^{\langle \mu \rangle}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\kappa_\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \tag{7}$$

where $\kappa_\mu = \varpi_1 + \frac{\varpi_2}{\mu}$ with some $\varpi_1, \varpi_2 \in \mathbb{R}_{++}$.

Remark III.4 (Sufficient conditions of Assumption III.3)**.** In analogy with [21, Lemma 3.2], if $\mathfrak{S}$ is Lipschitz continuous, then $\nabla F^{\langle \mu \rangle}$ turns out to be Lipschitz continuous with a Lipschitz constant

$$L_{\nabla F^{\langle \mu \rangle}} = L_{\nabla h} + L_{\mathrm{D}\mathfrak{S}}(L_f + L_g) + \frac{2L_{\mathfrak{S}}^2}{\mu},$$

where each L_Θ denotes a Lipschitz constant of a mapping Θ. Because $\kappa_\mu = L_{\nabla F^{\langle \mu \rangle}}$ enjoys (7) [28, Lemma 5.7], Assumption III.3 is achieved in this case. On the other hand, we also will present, in Section IV, an example where Assumption III.3 is satisfied without the Lipchitz continuity of $\mathfrak{S}$ (see Proposition IV.2).

Here, in order to show $\liminf_{k\to\infty}\|\nabla F_k(\boldsymbol{x}_k)\| = 0$ with $(\boldsymbol{x}_k)_{k=1}^{\infty}$ generated by Algorithm 1, we present the following lemma to see a behavior of the gradient sequence $(\nabla F_k(\boldsymbol{x}_k))_{k=1}^{\infty}$.

Lemma III.5. Consider Problem I.1. Choose arbitrarily a sequence $(\mu_k)_{k=1}^{\infty} \subset (0,(2\eta)^{-1}]$ satisfying (6), an initial point $\boldsymbol{x}_1 \in \mathbb{R}^d$, and inputs of Algorithm 2 $(\gamma_{\text{initial}}, \rho, c) \in \mathbb{R}_{++} \times (0,1) \times (0,1)$. Under Assumption III.3, the following inequality holds for the function sequence F_k and the point sequence $(\boldsymbol{x}_k)_{k=1}^{\infty}$ produced by Algorithm 1:

$$(\underline{k}, \bar{k} \in \mathbb{N} \text{ s.t. } \underline{k} \le \bar{k})\ \min_{\underline{k}\le k\le \bar{k}}\|\nabla F_k(\boldsymbol{x}_k)\| \le \sqrt{\frac{C}{\sum_{k=\underline{k}}^{\bar{k}}\mu_k}},$$

where $C > 0$ is a constant.

By noting that $\sum_{k=1}^{\infty} \mu_k = \infty$ from the condition (ii) in (6), we obtain the next convergence theorem.

Theorem III.6. Under the setting of Lemma III.5, we have

$$\liminf_{k\to\infty} \|\nabla F_k(\boldsymbol{x}_k)\| = 0.$$

Remark III.7 (Interpretation of Theorem III.6)**.** Theorem III.6 means that we can choose a subsequence $(\boldsymbol{x}_{m(l)})_{l=1}^{\infty}$ such that $\lim_{l\to\infty} \|\nabla F_{m(l)}(\boldsymbol{x}_{m(l)})\| = 0$, where $m : \mathbb{N} \to \mathbb{N}$ is monotonically increasing. Every cluster point of $(\boldsymbol{x}_{m(l)})_{l=1}^{\infty}$ is guaranteed to be a DC critical point of Problem 1.1 by applying Theorem III.2.

IV. Application to Robust Phase Retrieval

A. Optimization model in robust phase retrieval

To demonstrate the effectiveness of the proposed model (Problem I.1) and Algorithm 1, we carried out numerical experiments in a scenario of *the phase retrieval*. The phase retrieval is widely used, e.g., for crystallography [29], optical imaging [30], and astronomy [31]. The phase retrieval is a problem of estimating an original signal $\boldsymbol{x}^\star$ or $-\boldsymbol{x}^\star \in \mathbb{R}^d$ from the magnitude measurement

$$\boldsymbol{b}^\star := (A\boldsymbol{x}^\star) \odot (A\boldsymbol{x}^\star) := \left[\langle \boldsymbol{a}_1, \boldsymbol{x}^\star\rangle^2, \langle \boldsymbol{a}_2, \boldsymbol{x}^\star\rangle^2, ..., \langle \boldsymbol{a}_n, \boldsymbol{x}^\star\rangle^2\right]^T$$

where $A = [\boldsymbol{a}_1^T, \boldsymbol{a}_2^T ..., \boldsymbol{a}_n^T]^T \in \mathbb{R}^{n\times d}$, and $\odot$ means the element-wise product. While this $\boldsymbol{b}^\star$ is clean measurement, a measurement in real-world applications may be corrupted by a noise. In particular, [5] considers the measurements $\boldsymbol{b} \in \mathbb{R}^n$ with outliers as:

$$[\boldsymbol{b}]_i := \begin{cases} \langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle^2 & i \in \mathcal{I}_{\text{in}} \\ \xi_i & i \in \mathcal{I}_{\text{out}} \end{cases}$$

where $\mathcal{I}_{\text{in}}, \mathcal{I}_{\text{out}} \subset \{1, 2, ..., n\}$ denote disjoint index sets of inliers and outliers such that $\mathcal{I}_{\text{in}} \cup \mathcal{I}_{\text{out}} = \{1, 2, \ldots, n\}$, and $\xi_i > 0$ is a random noise.

In order to circumvent performance degradation caused by outliers, a robust phase retrieval [5] has been formulated as

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\ \|(A\boldsymbol{x}) \odot (A\boldsymbol{x}) - \boldsymbol{b}\|_1 \tag{8}$$

with the ℓ_1 norm $\|\cdot\|_1 : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{z} \mapsto \sum_{i=1}^n |[\boldsymbol{z}]_i|$. Although good estimation results have been reported by solving (8) with *Proximal linear algorithm* [5] and *Inexact proximal linear algorithm* [6], it is questionable whether the ℓ_1 norm in (8) can adequately suppress the effects of the outliers. To explain this, we rewrite the cost function in (8) as $\sum_{i\in\mathcal{I}_{\text{in}}} \left|\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle^2 - \langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle^2\right| + \sum_{i\in\mathcal{I}_{\text{out}}} \left|\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle^2 - \xi_i\right|$. If the cardinality $\#\mathcal{I}_{\text{out}}$ and each ξ_i are large, then the second summation also becomes large even if $\boldsymbol{x}$ is close to $\boldsymbol{x}^\star$ or $-\boldsymbol{x}^\star$. Such a situation may lead to performance degradation.

To resolve this issue, we propose the following reformulation of the robust phase retrieval.

Problem IV.1 (Proposed model for robust phase retrieval)**.** For given $A \in \mathbb{R}^{n\times d}$ and $\boldsymbol{b} \in \mathbb{R}^n$,

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\ F(\boldsymbol{x}) := \varphi((A\boldsymbol{x}) \odot (A\boldsymbol{x}) - \boldsymbol{b}), \tag{9}$$

where φ is chosen from Table I in Section I.
(Note: this problem is a special case of Problem I.1, where $h \equiv 0$, $f - g = \varphi$, and $\mathfrak{S} : \boldsymbol{x} \mapsto (A\boldsymbol{x}) \odot (A\boldsymbol{x}) - \boldsymbol{b}$.)

The proposed model (9) with nonconvex φ such as MCP [8], the capped ℓ_1 norm [10], and the trimmed ℓ_1 norm [11] seems to be more robust for the large outliers than the existing model (8). Indeed, the function for each entry in MCP and the capped ℓ_1 does not exceed a certain tunable constant value even if its entry has a large absolute value, while ℓ_1 norm does not have such desirable property. Alternatively, a large entry tends to be excluded from the summation in the trimmed ℓ_1 norm because the K largest absolute values are ignored therein. Therefore, these nonconvex φ are expected to remedy over-penalization in the model (8).

Algorithm 1 can be employed for Problem IV.1 since it is the special case of Problem I.1. Furthermore, the proposed convergence analysis in Theorem III.6 can be applied to Problem IV.1 because Assumption III.3 is achieved as shown in the following proposition. To the best of the authors' knowledge, Algorithm 1 is the first inner-loop-free algorithm applicable to Problem IV.1 with nonconvex φ.

Proposition IV.2. For Problem IV.1 with φ chosen from Table 1, $F^{\langle\mu\rangle}$ in (4) with $\mu \in (0, (2\eta)^{-1}]$ satisfies Assumption III.3.

B. Numerical experiments

In order to evaluate estimation performance of the robust phase retrieval via the proposed model (9), we applied Algorithm 1 to the model (9) using ℓ_1 norm, MCP, the capped ℓ_1 norm, and the trimmed ℓ_1 norm as φ (see Table I in Section I). Note that (9) with ℓ_1 norm is the same as the existing model (8), e.g., in [5], [6].

Our experimental setup inspired by [6] is as follows. We drew each entry of $A \in \mathbb{R}^{200\times 50}$ from the normal distribution $\mathcal{N}(0, 1)$. Each entry of the original signal $\boldsymbol{x}^\star \in \mathbb{R}^{50}$ was chosen from 1 or -1 with a probability of 0.5 respectively. The number of outliers $\mathcal{I}_{\text{out}}$ was 10, that was 5% of all entries of $\boldsymbol{b} \in \mathbb{R}^{200}$, and the position of outliers was randomly chosen. The value of each outliers ξ_i was given by

$$\xi_i = \Omega \tan\left(\frac{\pi}{2} u_i\right)\ (\geq 0), \tag{10}$$

where u_i was drawn from the uniform distribution of $[0, 1]$, and $\Omega > 0$ was used to control the magnitude of ξ_i. We used the parameter $(\lambda, \beta) = (1, 2000)$ and $(2, 500)$ for MCP, $\beta = 1000$ for the capped ℓ_1 norm[4], and $K = 5\,(= \#\mathcal{I}_{\text{out}})$ and 10 for the trimmed ℓ_1 norm. In Algorithm 1, we employed $\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{0}, I_{50})$ for a random initial point, $\mu_k = k^{-\frac{1}{3}}$ $(k \in \mathbb{N})$ for parameters of the Moreau envelope[5], and $(\gamma_{\text{initial}}, \rho, c) = (1, 0.8, 0.0001)$ for inputs of Algorithm 2. We stopped Algorithm 1 when $\|\nabla F_k(\boldsymbol{x}_k)\| < 0.001$ held or the iteration k reached to 10000. For each φ in Table I and each $\Omega \in \{100, 1000, 2000, 5000, 10000\}$ in (10), we performed 50 trials of estimation with the model (9). We judged

[4]The functions for each entry in MCP and the capped ℓ_1 norm is constant τ when an input is far from the origin (see Table I in Section I). We set parameters of MCP and the capped ℓ_1 norm so that $\tau = 1000$.

[5]Since $(2\eta)^{-1} \geq 1$ holds for f and g with parameters used in this experiment, we have $(\mu_k)_{k=1}^{\infty} \subset (0, (2\eta)^{-1}]$, and $(\mu_k)_{k=1}^{\infty}$ also satisfies (6).

TABLE II: Success rate [%]

Ω	Existing ℓ_1	Proposed MCP ($\lambda = 1$)	MCP ($\lambda = 2$)	Capped ℓ_1	Trimmed ℓ_1 ($K = 5$)	Trimmed ℓ_1 ($K = 10$)
10	90	86	88	**92**	66	54
1000	60	**80**	78	70	78	72
3000	56	**84**	80	**84**	82	74
5000	54	80	82	**86**	84	78
10000	54	80	82	86	**88**	78

TABLE III: Averaged time [sec]

Ω	Existing ℓ_1	Proposed MCP ($\lambda = 1$)	MCP ($\lambda = 2$)	Capped ℓ_1	Trimmed ℓ_1 ($K = 5$)	Trimmed ℓ_1 ($K = 10$)
10	**2.50** (2.50)	3.20 (3.22)	3.30 (3.32)	3.30 (3.31)	5.74 (**0.69**)	8.26 (1.52)
1000	**2.55** (2.57)	3.20 (3.22)	3.26 (3.26)	3.32 (3.35)	3.74 (**0.55**)	4.95 (0.71)
3000	**2.47** (2.50)	3.23 (3.25)	3.22 (3.23)	2.83 (2.75)	3.62 (0.76)	4.48 (**0.52**)
5000	2.55 (2.61)	3.11 (3.12)	3.16 (3.17)	**2.36** (2.21)	3.12 (0.76)	3.61 (**0.46**)
10000	2.49 (2.55)	3.13 (3.14)	3.08 (3.08)	**1.65** (1.39)	2.50 (0.75)	3.74 (**0.48**)

The values out of and in the parentheses are the averaged time taken for all estimations and only for successful estimations, respectively.

that an estimation succeeds if the relative error, defined as $\min\{\|\boldsymbol{x}^\star - \boldsymbol{x}^\diamond\|, \|\boldsymbol{x}^\star + \boldsymbol{x}^\diamond\|\}/\|\boldsymbol{x}^\star\|$, achieves a smaller value than 10^{-3}, where $\boldsymbol{x}^\diamond$ is the final estimate of Algorithm 1. As in [6], we used an estimation performance criterion called "success rate" that is the percentage of the successful estimation out of 50 estimations.

Table II shows the success rates for each φ and each Ω. From Table II, the model (9) with nonconvex functions, i.e., MCP, the capped ℓ_1 norm, and the trimmed ℓ_1 norm, keep high success rates even with large outliers while the model (9) with the ℓ_1 norm does not. In particular, the inherently DC functions, the capped ℓ_1 norm and the trimmed ℓ_1 norm with $K = 5$, have higher success rates than others when outliers are large. (Note that the result of the trimmed ℓ_1 with $K = 5$ is based on the utilization of the number of outliers.) Table III demonstrates the averaged CPU time for Algorithm 1 to be terminated. Algorithm 1 for the model (9) achieves the fastest average convergence speed (i) with the capped ℓ_1 among all estimations for huge $\Omega \in \{5000, 10000\}$, and (ii) with the trimmed ℓ_1 among successful estimations for all Ω. As above, the proposed model using inherently DC functions φ have better results in both estimation performance and speed than the existing model using ℓ_1 norm.

V. Conclusion

We presented an inner-loop-free variable smoothing algorithm for nonsmooth DC composite type problems with its convergence analysis. The proposed algorithm was designed to find a DC critical point by generating the sequence of points at which the gradient of the smooth surrogate function approaches zero. The numerical experiments in a scenario of robust phase retrieval demonstrated the effectiveness of the proposed optimization model (9) using inherently DC functions.

References

[1] J. You, Y. Jiao, X. Lu, and T. Zeng, "A nonconvex model with minimax concave penalty for image restoration," *Journal of Scientific Computing*, vol. 78, no. 2, pp. 1063–1086, 2019.

[2] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "ℓ_1 trend filtering," *SIAM review*, vol. 51, no. 2, pp. 339–360, 2009.

[3] X. Huang, Y. Liu, L. Shi, S. Van Huffel, and J. A. Suykens, "Two-level ℓ_1 minimization for compressed sensing," *Signal Processing*, vol. 108, pp. 459–475, 2015.

[4] Z. Lu and X. Li, "Sparse recovery via partial regularization: models, theory, and algorithms," *Mathematics of Operations Research*, vol. 43, no. 4, pp. 1290–1316, 2018.

[5] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 471–529, 2019.

[6] Z. Zheng, S. Ma, and L. Xue, "A new inexact proximal linear algorithm with adaptive stopping criteria for robust phase retrieval," *IEEE Transactions on Signal Processing*, vol. 72, pp. 1081–1093, 2024.

[7] H. Bauschke and P. L. Combettes, "Convex analysis and monotone operator theory in Hilbert spaces," 2011.

[8] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, pp. 894–942, 2010.

[9] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[10] T. Zhang, "Multi-stage convex relaxation for learning with sparse regularization," *Advances in neural information processing systems*, 2008.

[11] Z. Luo, Y. Wang, and X. Zhang, "New improved penalty methods for sparse reconstruction based on difference of two norms," Technical report, 2013. doi: 10.13140/RG. 2.1. 3256.3369, Tech. Rep., 2015.

[12] T. Sasaki, Y. Bandoh, and M. Kitahara, "Sparse regularization based on reverse ordered weighted l 1-norm and its application to edge-preserving smoothing," in *IEEE ICASSP 2024*.

[13] C. Wang, Q. Ye, P. Luo, N. Ye, and L. Fu, "Robust capped l1-norm twin support vector machine," *Neural Networks*, vol. 114, pp. 47–59, 2019.

[14] Q. Sun, S. Xiang, and J. Ye, "Robust principal component analysis via capped norms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.

[15] S. Yagishita and J. Gotoh, "Exact penalization at d-stationary points of cardinality- or rank-constrained problem," 2024. [Online]. Available: https://arxiv.org/abs/2209.02315

[16] H. A. Le Thi and T. Pham Dinh, "Open issues and recent advances in DC programming and DCA," *Journal of Global Optimization*, vol. 88, no. 3, pp. 533–590, 2024.

[17] H. A. Le Thi, V. N. Huynh, and T. P. Dinh, "Minimizing compositions of differences-of-convex functions with smooth mappings," *Mathematics of Operations Research*, vol. 49, no. 2, pp. 1140–1168, 2024.

[18] J. Gotoh, A. Takeda, and K. Tono, "DC formulations and algorithms for sparse optimization problems," *Mathematical Programming*, vol. 169, pp. 141–176, 2018.

[19] Y. Zhang and I. Yamada, "An inexact proximal linearized DC algorithm with provably terminating inner loop," *Optimization*, pp. 1–33, 2024.

[20] A. Böhm and S. J. Wright, "Variable smoothing for weakly convex composite functions," *Journal of optimization theory and applications*, vol. 188, pp. 628–649, 2021.

[21] K. Kume and I. Yamada, "A variable smoothing for nonconvexly constrained nonsmooth optimization with application to sparse spectral clustering," in *IEEE ICASSP 2024*.

[22] ——, "A variable smoothing for weakly convex composite minimization with nonconvex constraint," *arXiv preprint arXiv:2412.04225*, 2024.

[23] J. Li, A. M.-C. So, and W.-K. Ma, "Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions," *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 18–31, 2020.

[24] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[25] K. Sun and X. A. Sun, "Algorithms for difference-of-convex programs based on difference-of-moreau-envelopes smoothing," *INFORMS Journal on Optimization*, vol. 5, no. 4, pp. 321–339, 2023.

[26] T. Hoheisel, M. Laborde, and A. Oberman, "A regularization interpretation of the proximal point method for weakly convex functions," *J. Dyn. Games*, vol. 7, no. 1, pp. 79–96, 2020.

[27] N. Andrei, *Nonlinear conjugate gradient methods for unconstrained optimization*. Springer, 2020.

[28] A. Beck, *First-order methods in optimization*. SIAM, 2017.

[29] R. P. Millane, "Phase retrieval in crystallography and optics," *Journal of the Optical Society of America A*, vol. 7, no. 3, pp. 394–411, 1990.

[30] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: a contemporary overview," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[31] C. Fienup and J. Dainty, "Phase retrieval and image reconstruction for astronomy," *Image recovery: theory and application*, vol. 231, p. 275, 1987.