A New Inexact Gradient Descent Method with Applications to Nonsmooth Convex Optimization

Pham Duy Khanh* Boris S. Mordukhovich† Dat Ba Tran‡

February 19, 2024

Dedicated to the memory of Andreas Griewank, an outstanding mathematician and a wonderful human being

Abstract. The paper proposes and develops a novel inexact gradient method (IGD) for minimizing $\mathcal{C}^1$-smooth functions with Lipschitzian gradients, i.e., for problems of $\mathcal{C}^{1,1}$ optimization. We show that the sequence of gradients generated by IGD converges to zero. The convergence of iterates to stationary points is guaranteed under the Kurdyka-Łojasiewicz (KL) property of the objective function with convergence rates depending on the KL exponent. The newly developed IGD is applied to designing two novel gradient-based methods of nonsmooth convex optimization such as the inexact proximal point methods (GIPPM) and the inexact augmented Lagrangian method (GIALM) for convex programs with linear equality constraints. These two methods inherit global convergence properties from IGD and are confirmed by numerical experiments to have practical advantages over some well-known algorithms of nonsmooth convex optimization.

Key words: inexact gradient methods, inexact proximal point methods, inexact augmented Lagrangian methods, $\mathcal{C}^{1,1}$ optimization, nonsmooth convex optimization

Mathematics Subject Classification (2020) 90C52, 90C56, 90C25, 90C26

1 Introduction

The paper addresses numerical optimization, the area to which the contribution of Andreas Griewank is difficult to overstate. In particular, his book with Andrea Walther [25] is the classical collection of constructive methods of algorithmic differentiation in nonlinear optimization being a strong inspiration for a great many of applied mathematicians and practitioners.

This paper is devoted to developing new algorithms to solve various classes of optimization problems. Let us start with the following basic problem of $\mathcal{C}^{1,1}$ *optimization*:

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \mathbb{R}^n, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable ($\mathcal{C}^1$-smooth) objective function whose gradient ∇f is globally Lipschitzian on $\mathbb{R}^n$ with some constant $L > 0$. A very natural and classical approach to solve optimization problems of type (1.1) is by using the *gradient descent method* described as follows: given a starting point $x^1 \in \mathbb{R}^n$, construct the iterative procedure

$$x^{k+1} := x^k - \frac{1}{L}\nabla f(x^k) \quad \text{for all} \quad k \in \mathbb{N}. \tag{1.2}$$

*Department of Mathematics, Ho Chi Minh City University of Education, Ho Chi Minh City, Vietnam. E-mails: pdkhanh182@gmail.com, khanhpd@hcmue.edu.vn. Research of this author is funded by the Ministry of Education and Training Research Funding under the project "*First-order approximation methods and applications*"

†Department of Mathematics, Wayne State University, Detroit, Michigan, USA. E-mail: aa1086@wayne.edu. Research of this author was partly supported by the US National Science Foundation under grants DMS-1808978 and DMS-2204519, by the Australian Research Council under grant DP-190100555, and by Project 111 of China under grant D21024.

‡Department of Mathematics, Wayne State University, Detroit, Michigan, USA. E-mail: tranbadat@wayne.edu. Research of this author was partly supported by the US National Science Foundation under grants DMS-1808978 and DMS-2204519.

Largely due to its simplicity, the gradient descent method is broadly used to solve various optimization problems; see, e.g., [7, 13, 17, 40, 41]. However, errors in gradient calculations may appear deterministically in various contexts. In derivative-free optimization problems where only the function values are accessible, the exact gradients are not available but only their approximations via finite differences and simplex gradients are [2, 5, 16]. In addition, many optimization methods for minimizing nonsmooth functions [18, 37, 43, 45, 46, 48, 49] can be treated via the gradient descent method applied to their regularized functions. Errors in gradient calculations for regularized functions, which appear in practical situations, require developing gradient descent methods with *inexact gradient information*.

For these reasons, gradient methods taking errors into account have been developed over the years. Among the major ones, we list the following. Gilmore and Kelley [24] proposed the implicit filtering algorithm for minimizing a noisy smooth function under box constraints, which uses finite difference approximations of the exact gradient. Bertsekas and Tsitsiklis [11] justified the convergence of gradients in gradient methods with errors smaller than the multiplication of the stepsize by an expression depending on the exact gradient. The convergence for incremental and stochastic gradient methods was also established in [11]. Nesterov proposed in [38] and further developed in [39] and his paper with Devolder and Glineur [18] gradient methods with inexact oracle for convex functions. Generalizations of the inexact oracle are discussed in [12, 19] for nonconvex functions. Additionally, accelerated gradient methods considering various types of errors are analyzed in [51]. Optimization methods with stochastic errors are also studied in [22, 22]. Quite recently, the authors of the present paper [31] proposed a general framework of inexact reduced gradient (IRG) methods for minimizing nonconvex $\mathcal{C}^1$-smooth function with or without the Lipschitz continuity of the gradient under the condition that the distance between the exact and the approximate gradient is less than a determined error at each iteration. To the best of our knowledge, [31] is the first paper addressing the global convergence and convergence rates of inexact gradient descent methods with deterministic error for general classes of nonconvex functions.

The present paper designs and justifies the novel *inexact gradient descent* (IGD) method to solve problem (1.1) by using some IRG ideas from [31]. The iterative procedure of IGD is (1.2) with the exact gradient replaced by one of its approximations. The main differences between IGD and its predecessor IRG are that the construction of IGD in Algorithm 1 use directly an approximate gradient as a descent direction instead of introducing another reduced gradient direction as in [31, Algorithm 1]. The convergence analysis of the new IGD method does not involve the notion of null iterations which is essential in the results of [31, Propositions 4.6, 4.7] and the construction of non-null iteration as in [31, (5.23)]. The IGD method also achieves the following fundamental convergence properties:

- Convergence to zero of the gradient sequence.
- Convergence to some stationary point of the iterative sequence under the Kurdyka-Łojasiewicz (KL) property of the objective function.
- Convergence rates depending on the KL exponent of the objective function for the sequences of iterates, gradients, and function values.

It should be mentioned that the results in [18, 38, 39] address only convex optimization problems. Although the implicit filtering method in [24] deals with nonconvex problems, it does not achieve any of the fundamental convergence properties above due to the presence of noise. The gradient methods with errors in [11] establish only the convergence to zero of the gradient sequence. Moreover, the rates of convergence for the gradient sequence and function value sequence are not provided in [31].

The IGD method developed in this paper for $\mathcal{C}^{1,1}$ optimization problems of type (1.1) is applied then to problems of *nonsmooth optimization* to design and justify the following gradient-based methods:

- The *gradient-based inexact proximal point method* (GIPPM) for minimizing general nonsmooth convex functions.
- The *gradient-based inexact augmented Lagrangian method* (GIALM) for minimizing nonsmooth convex functions under linear equality constraints.

The two methods listed above inherit the convergence properties from IGD with establishing in this way the stationarity of accumulation points and the global convergence under the KL property of the objective function with convergence rates depending on the KL exponent. Regarding application aspects, the conducted numerical experiments show that the GIALM method with the new handling of errors exhibit a better performance than the classical inexact augmented Lagrangian method by Rockafellar [46] in practical problems of image processing as well as in random generated examples.

The rest of the paper is organized as follows. Section 2 presents some preliminaries and discussions. The construction and convergence analysis of the main IGD method are given in Section 3. The next Section 4, which is split into two subsections, proposes and justifies the new GIPPM and GIALM methods of nonsmooth convex optimization with their convergence analysis. The numerical experiments and comparisons between the newly developed inexact gradient-based methods with the existing algorithms are given in the final Section 5.

2 Preliminaries and Discussions

First we recall some notions and notation frequently used in the paper. All our considerations are given in the space $\mathbb{R}^n$ with the Euclidean norm $\|\cdot\|$. We use the matrix norm defined by

$$\|A\| := \max\big\{\|Ax\| \;\big|\; \|x\| = 1\big\} \;\text{ for any }\; m \times n \text{ matrix } A.$$

Along with the Euclidean norm $\|\cdot\|$, in Section 5 we also use the ℓ_1 norm $\|\cdot\|_1$ denoted by

$$\|x\|_1 = \sum_{k=1}^{n} |x_k| \;\text{ for all }\; x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

As always, $\mathbb{N} := \{1, 2, \ldots\}$ signifies the collections of natural numbers. The symbol $x^k \xrightarrow{J} \bar{x}$ means that $x^k \to \bar{x}$ as $k \to \infty$ with $k \in J \subset \mathbb{N}$. Recall that $\bar{x}$ is a *stationary point* of a $\mathcal{C}^1$-smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$ if $\nabla f(\bar{x}) = 0$. A $\mathcal{C}^1$-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ is said to have a Lipschitz continuous gradient with the uniform constant $L > 0$ on $\mathbb{R}^n$ if

$$\|\nabla f(x) - \nabla f(y)\| \le L\,\|x - y\| \;\text{ for all }\; x, y \in \mathbb{R}^n.$$

It follows from [28, Lemma A.11] that the Lipschitz continuity of ∇f with constant $L > 0$ implies that

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\,\|y - x\|^2 \;\text{ for all }\; x, y \in \mathbb{R}^n. \tag{2.1}$$

The converse implication may not hold in general as discussed in [31, Section 2].

The following *Kurdyka-Łojasiewicz property* is taken from Absil et al. [1, Theorem 3.4].

Definition 2.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. We say that f satisfies the *KL property* at $\bar{x} \in \mathbb{R}^n$ if there exist a number $\eta > 0$, a neighborhood U of $\bar{x}$, and a nondecreasing function $\psi : (0, \eta) \to (0, \infty)$ such that the function $1/\psi$ is integrable over $(0, \eta)$ and we have

$$\|\nabla f(x)\| \ge \psi\big(f(x) - f(\bar{x})\big) \;\text{ for all }\; x \in U \;\text{ with }\; f(\bar{x}) < f(x) < f(\bar{x}) + \eta. \tag{2.2}$$

Remark 2.2. It has been realized that the KL property is satisfied in broad settings. In particular, it holds at every *nonstationary point* of f; see [3, Lemma 2.1 and Remark 3.2(b)]. Furthermore, it is proved at the seminal paper by Łojasiewicz [35] that any analytic function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the KL property at every point $\bar{x}$ with $\psi(t) = Mt^q$ for some $q \in [0, 1)$. As demonstrated in [31, Section 2], the KL property formulated in Attouch et al. [3] is stronger than the one in Definition 2.1. Typical smooth functions that satisfy the KL property from [3], and hence the one from Definition 2.1, are smooth *semialgebraic* functions and also those from the more general class of functions known as *definable in o-minimal structures*; see [3, 4, 32]. The following preservation properties for general semialgebraic functions can be found, e.g., in [3]:

- Finite sums and products of semialgebraic functions are semialgebraic.
- Functions of the marginal type type

$$\mathbb{R}^n \ni x \mapsto f(x) = \inf_{y \in \mathbb{R}^m} g(x,y),$$

where g is a semialgebraic function of two variables, are semialgebraic.

Next we present based on [1] some descent-type conditions ensuring the global convergence of iterates for smooth functions that satisfy the KL property.

Proposition 2.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function, and let the following conditions hold along a sequence of iterates $\{x^k\} \subset \mathbb{R}^n$ for the function f:*

(H1) (primary descent condition). *There is $\sigma > 0$ such that we have*

$$f(x^k) - f(x^{k+1}) \geq \sigma \left\|\nabla f(x^k)\right\| \cdot \left\|x^{k+1} - x^k\right\|$$

for sufficiently large $k \in \mathbb{N}$.

(H2) (complementary descent condition). *The following implication holds:*

$$\left[f(x^{k+1}) = f(x^k)\right] \Longrightarrow \left[x^{k+1} = x^k\right]$$

for sufficiently large $k \in \mathbb{N}$.

If $\bar{x}$ is an accumulation point of $\{x^k\}$ and f satisfies the KL property at $\bar{x}$, then $x^k \to \bar{x}$ as $k \to \infty$.

When the sequence under consideration is generated by a linesearch method and satisfies some stronger conditions than (H1) and (H2) in Proposition 2.3, its convergence rates are established in [31] under the KL property with $\psi(t) = Mt^q$ as given below.

Proposition 2.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function. Assume that the sequences $\{x^k\}$ and $\{d^k\}$ satisfy the iterative update $x^{k+1} = x^k + \tau d^k$ with $x^{k+1} \neq x^k$ for all $k \in \mathbb{N}$, and that*

$$f(x^k) - f(x^{k+1}) \geq \alpha \left\|d^k\right\|^2 \quad \text{and} \quad \left\|\nabla f(x^k)\right\| \leq \beta \left\|d^k\right\| \tag{2.3}$$

for sufficiently large $k \in \mathbb{N}$ with some constants $\tau, \alpha, \beta > 0$. Suppose in addition that $\bar{x}$ is an accumulation point of $\{x^k\}$ and that f satisfies the KL property at $\bar{x}$ with $\psi(t) = Mt^q$ for some $M > 0$ and $q \in (0,1)$. Then the following convergence rates are guaranteed:

(i) *For $q \in (0, 1/2]$, the sequence $\{x^k\}$ converges linearly to $\bar{x}$ as $k \to \infty$.*

(ii) *For $q \in (1/2, 1)$, we have $\left\|x^k - \bar{x}\right\| = \mathcal{O}(k^{-\frac{1-q}{2q-1}})$ as $k \to \infty$*

3 Inexact Gradient Descent Method

In this section, which is split into two subsections, we design and provide a convergence analysis of our main *inexact gradient method* (IGD) to solve $\mathcal{C}^{1,1}$ optimization problems of type (1.1).

3.1 Algorithm Formulation and Motivating Examples

Here is the proposed IGD algorithm to solve (1.1) by using inexact gradients.

Algorithm 1 (IGD)**.**

Step 0 (initialization)**.** Select an initial point $x^1 \in \mathbb{R}^n$, initial error ε_1, error reduction factor $\theta \in (0,1)$, and error scaling factor $\mu > 1$. Set $k := 1$.

Step 1 (inexact gradient). Find g^k and the smallest natural number i_k such that

$$\left\|g^k - \nabla f(x^k)\right\| \le \theta^{i_k}\varepsilon_k \text{ and } \left\|g^k\right\| > \mu\theta^{i_k}\varepsilon_k. \tag{3.1}$$

Step 2 (iteration and error update). Set $x^{k+1} := x^k - \frac{1}{L}g^k$ and $\varepsilon_{k+1} := \theta^{i_k}\varepsilon_k$. Increase k by 1 and go back to Step 1.

Let us discuss some important features of Algorithm 1 and present an illustrating figure.

Remark 3.1. We have the following observations:

(i) *The existence of g^k and i_k in Step* 1: The procedure of finding g^k and i_k that satisfy Step 1 can be given as follows. Set $i_k = 0$ and find some g^k such that

$$\left\|g^k - \nabla f(x^k)\right\| \le \theta^{i_k}\varepsilon_k. \tag{3.2}$$

While $\left\|g^k\right\| \le \mu\theta^{i_k}\varepsilon_k$, increase i_k by 1 and recalculate g^k under (3.2). When $\nabla f(x^k) \ne 0$, the existence of g^k and i_k in Step 1 is guaranteed. Indeed, otherwise we get a sequence $\{g_i^k\}$ with

$$\left\|g_i^k - \nabla f(x^k)\right\| \le \theta^i\varepsilon_k \text{ and } \left\|g_i^k\right\| \le \mu\theta^i\varepsilon_k \text{ for all } i \in \mathbb{N}.$$

Since $\theta \in (0,1)$, this implies that $\nabla f(x^k) = 0$, a contradiction. In fact, we can bound the number i_k by considering the two cases as follows.

Case 1: $\varepsilon_k < \frac{1}{\mu+1}\left\|\nabla f(x^k)\right\|$. In this case, we show that $i_k = 0$.

To proceed, find some g^k such that $\left\|g^k - \nabla f(x^k)\right\| \le \varepsilon_k$. Combining this with $\varepsilon_k < \frac{1}{\mu+1}\left\|\nabla f(x^k)\right\|$ and with the triangle inequality yields

$$\begin{aligned}\left\|g^k\right\| &\ge \left\|\nabla f(x^k)\right\| - \left\|g^k - \nabla f(x^k)\right\| \\ &\ge \left\|\nabla f(x^k)\right\| - \varepsilon_k \\ &> \left\|\nabla f(x^k)\right\| - \frac{1}{\mu+1}\left\|\nabla f(x^k)\right\| \\ &= \frac{\mu}{\mu+1}\left\|\nabla f(x^k)\right\| > \mu\varepsilon_k.\end{aligned}$$

This means that g^k satisfies the desired condition with $i_k = 0$.

Case 2: $\varepsilon_k \ge \frac{1}{\mu+1}\left\|\nabla f(x^k)\right\|$. In this case, we show that $i_k \le \log_\theta\left(\frac{1}{\varepsilon_k(\mu+1)}\left\|\nabla f(x^k)\right\|\right) + 1$.

To proceed, assume while arguing by contradiction that $i_k > \log_\theta\left(\frac{1}{\varepsilon_k(\mu+1)}\left\|\nabla f(x^k)\right\|\right)+1$. Define $j := i_k - 1$ and get that $j > 0$. We show that for any $g^k \in \mathbb{R}^n$ such that $\left\|g^k - \nabla f(x^k)\right\| \le \theta^j\varepsilon_k$, the estimate $\left\|g^k\right\| > \theta^j\varepsilon_k$ holds, which violates the construction of i_k. It follows from $j > \log_\theta\left(\frac{1}{\varepsilon_k(\mu+1)}\left\|\nabla f(x^k)\right\|\right)$ and $\theta \in (0,1)$ that

$$\theta^j\varepsilon_k < \frac{1}{\mu+1}\left\|\nabla f(x^k)\right\|.$$

Now take any $g^k \in \mathbb{R}^n$ such that $\left\|g^k - \nabla f(x^k)\right\| \le \theta^j\varepsilon_k$ and deduce that

$$\left\|g^k - \nabla f(x^k)\right\| \le \theta^j\varepsilon_k < \frac{1}{\mu+1}\left\|\nabla f(x^k)\right\|.$$

Applying the triangle inequality gives us

$$\begin{aligned}\left\|g^k\right\| &\ge \left\|\nabla f(x^k)\right\| - \left\|g^k - \nabla f(x^k)\right\| \\ &> \left\|\nabla f(x^k)\right\| - \frac{1}{\mu+1}\left\|\nabla f(x^k)\right\| \\ &= \frac{\mu}{\mu+1}\left\|\nabla f(x^k)\right\| > \mu\theta^j\varepsilon_k.\end{aligned}$$

Therefore, $i_k \le \log_\theta\left(\frac{1}{\varepsilon_k(\mu+1)}\left\|\nabla f(x^k)\right\|\right) + 1$.

(ii) *Geometric illustration of* (3.1): It follows from Step 1 and Step 2 of Algorithm 1 that

$$\left\|g^k - \nabla f(x^k)\right\| \leq \varepsilon_{k+1} \quad \text{and} \quad \left\|g^k\right\| > \mu\varepsilon_{k+1} \quad \text{for all} \quad k \in \mathbb{N}. \tag{3.3}$$

Since $\mu > 1$, the two conditions in (3.3) make the angle between $\nabla f(x^k)$ and g^k smaller than 90^0, which ensures that $-g^k$ is a descent direction of f at x^k. As illustrated in Figure 1, the angle between these vectors can be chosen arbitrarily small by increasing μ.

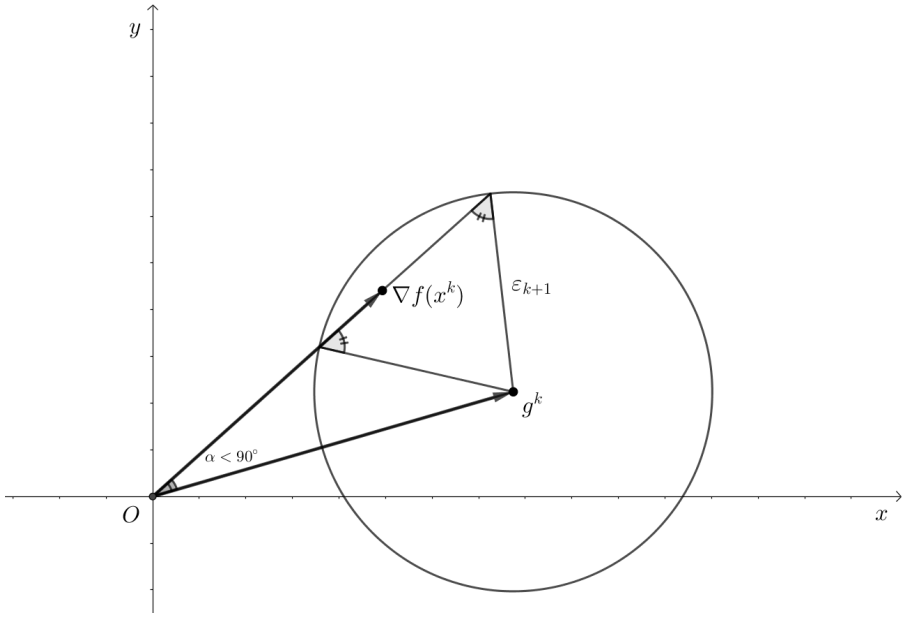

Figure 1: Geometric illustration of (3.1)

Note that the problem of finding g^k satisfying the first condition in (3.1) can be considered more generally as follows. Given any point $x \in \mathbb{R}^n$ and any error value $\varepsilon > 0$, find an approximation $\mathcal{G}$ of the gradient $\nabla f(x)$ such that

$$\|\mathcal{G} - \nabla f(x)\| \leq \varepsilon. \tag{3.4}$$

Let us present some examples where the approximate condition (3.4) appears in practical situations.

Example 1 (Gradient approximation methods)**.** In many problems of *derivative-free optimization* [5, 16], we only have access to function values, which can be used to derive approximate gradients satisfying (3.4) by employing gradient approximation methods.

1. *Forward finite difference* (*FFD*): For a fixed number $\delta > 0$, the FFD formula gives us the approximation of the gradient $\nabla f(x)$ defined by

$$\mathcal{G} := \frac{1}{\delta}\sum_{i=1}^{n}\big(f(x + \delta e_i) - f(x)\big)e_i, \tag{3.5}$$

where e_i is the i^{th} vector in the standard basis of $\mathbb{R}^n$. The following error bound is standard and can be found, e.g., in [42, Section 7.2] and [9, Theorem 2.1]:

$$\|\mathcal{G} - \nabla f(x)\| \leq \frac{L\sqrt{n}\delta}{2}. \tag{3.6}$$

Therefore, the choice of $\delta \in \left(0, \frac{2\varepsilon}{L\sqrt{n}}\right)$ guarantees the error bound (3.4).

2. *Centered finite difference* (*CFD*): For a given number $\delta > 0$, the CFD formula gives us the approximation $\nabla f(x)$ defined by

$$\mathcal{G} := \frac{1}{\delta}\sum_{i=1}^{n}\Big(f\big(x+\frac{\delta}{2}e_i\big)-f\big(x-\frac{\delta}{2}e_i\big)\Big)e_i. \tag{3.7}$$

Assume that f is twice continuously differentiable ($\mathcal{C}^2$-smooth) having the Lipschitz continuous Hessian with constant $M > 0$. It follows from [9, Theorem 2.2] that

$$\|\mathcal{G}-\nabla f(x)\| \le \frac{\sqrt{n}M\delta^2}{24}. \tag{3.8}$$

Therefore, the choice of $\delta \in \Big(0, \sqrt{\frac{24\varepsilon}{M\sqrt{n}}}\Big)$ guarantees the error bound required in (3.4).

When f is smoothed from another (possibly *nonsmooth*) function, the exact information about the function value and its gradient is usually not available. However, some specific smoothing techniques allow us to find an approximation of the gradient that satisfies condition (3.4).

Example 2 (Moreau envelopes)**.** Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, lower semicontinuous (l.s.c.), convex function, and let $f := e_\lambda g$ be its Moreau envelope with proximal parameter $\lambda > 0$; see Section 4 for more details. By (4.2), the problem of calculating $\nabla f(x)$ approximately is equivalent to the approximate calculation of $\mathrm{Prox}_{\lambda g}(x)$ approximately. The latter problem is equivalent to

$$\underset{y\in\mathbb{R}^n}{\text{minimize}} \quad \varphi_\lambda(y) := g(y) + \frac{1}{2\lambda}\|y-x\|^2. \tag{3.9}$$

Let $\bar{y}$ be such that $\varphi_\lambda(\bar{y}) - \inf_{y\in\mathbb{R}^n}\varphi_\lambda(y) \le \dfrac{\lambda\varepsilon^2}{2}$, and set $\mathcal{G} := \lambda^{-1}(x-\bar{y})$. Since φ_λ is strongly convex with positive constant $1/\lambda$, we deduce from (4.2) that

$$\begin{aligned}\|\mathcal{G}-\nabla f(x)\| = \lambda^{-1}\|\bar{y}-\mathrm{Prox}_{\lambda g}(x)\| &\le \lambda^{-1}\sqrt{2\lambda\Big(\varphi_\lambda(\bar{y}) - \inf_{y\in\mathbb{R}^n}\varphi_\lambda(y)\Big)}\\ &\le \lambda^{-1}\cdot\sqrt{2\lambda\frac{\lambda}{2}\varepsilon^2} = \varepsilon.\end{aligned}$$

Therefore, $\mathcal{G}$ is an approximation of $\nabla f(x)$ with the error bound ε.

3.2 Convergence Analysis

In this subsection, we establish the convergence properties of Algorithm 1 including the convergence of the sequence of gradients (to zero), the sequence of iterates, and the sequence of function values together with their convergence rates under the KL property of the objective function. Let us begin by connecting the conditions in (3.3) with the following two conditions on the gradient estimates:

$$\big\|g^k-\nabla f(x^k)\big\| \le \nu_1\big\|\nabla f(x^k)\big\| \quad \text{for all} \quad k\in\mathbb{N}, \tag{3.10}$$

$$\big\|g^k-\nabla f(x^k)\big\| \le \nu_2\big\|g^k\big\| \quad \text{for all} \quad k\in\mathbb{N}, \tag{3.11}$$

with some $\nu_1, \nu_2 \in [0,1)$. Condition (3.10) was first introduced and studied by Polyak [44, Section 4.2.3] for the convergence of inexact gradient methods for strongly convex functions. In [10], the convergence of a general linesearch algorithm with noise is established when (3.10) is satisfied with high probability. However, since $\nabla f(x^k)$ is unknown, it is not easy to ensure (3.10). In fact, the authors in [9] write:

"*Clearly, unless we know* $\big\|\nabla f(x^k)\big\|$*, condition* (3.10) *may be hard or impossible to verify or guarantee.*"

Regarding condition (3.11), it was studied in [15] for trust-region methods with inexact gradients for minimizing $\mathcal{C}^1$-smooth and bounded from below functions with Lipschitzian gradients. For linesearch

methods, condition (3.11) is a standing assumption for investigating complexity of the inexact gradient method applied to $\mathcal{C}^2$-smooth *strongly convex* functions in [14], and also for convergence properties of the inexact gradient method applied to *convex smooth* functions in [33].

It turns out that the sequence $\{g^k\}$ generated by Algorithm 1 gives us a *universal approach* to ensure both conditions (3.10) and (3.11) when μ is chosen properly. Indeed, it follows from (3.3) that (3.11) is satisfied with $\nu_2 = 1/\mu$ if $\mu > 1$. Applying the triangle inequality in (3.11) brings us to (3.10) with $\nu_1 = \frac{1}{\mu-1} \in (0,1)$ whenever $\mu > 2$. In addition to this observation, Step 1 of Algorithm 1 also tells us that the error between g^k and $\nabla f(x^k)$ is chosen to be the largest one among $\{\theta^i \varepsilon_k \mid i \in \mathbb{N}\}$, which saves the executed time for finding g^k in practice. In conclusion, Algorithm 1 uses the gradient errors automatically controlled to be as large as possible while maintaining conditions (3.10) and (3.11).

Although conditions (3.10) and (3.11) are used in many contexts for linesearch methods, the global convergence of the sequence of gradients (to zero), the sequence of iterates, and the sequence of function values together with their convergence rates *were not established* in the studies listed above in the case of *nonconvex functions*. These properties are now derived in the next theorem.

Theorem 3.2. *Let $\{x^k\}$ and $\{g^k\}$ be the sequences satisfying the iterative procedure $x^{k+1} = x^k - \frac{1}{L}g^k$ for all $k \in \mathbb{N}$ under the condition (3.11) with $\nu_2 < \frac{1}{2}$. Assume that $\nabla f(x^k) \neq 0$ for all $k \in \mathbb{N}$. Then either $\|x^k\| \to \infty$, or we have the assertions:*

(i) *Both sequences $\{\nabla f(x^k)\}$ and $\{g^k\}$ converge to $0 \in \mathbb{R}^n$ as $k \to \infty$.*

(ii) *If f satisfies the KL property at some accumulation point $\bar{x}$ of $\{x^k\}$, then $x^k \to \bar{x}$ as $k \to \infty$.*

(iii) *If f satisfies the KL property at some accumulation point $\bar{x}$ of $\{x^k\}$ with $\psi(t) = Mt^q$ for $M > 0$ and $q \in (0,1)$, then the following convergence rates are guaranteed:*

- *For $q \in (0, 1/2]$, the sequences $\{x^k\}$, $\{f(x^k)\}$, and $\{\nabla f(x^k)\}$ converge linearly to $\bar{x}$, $f(\bar{x})$, and $0 \in \mathbb{R}^n$ respectively.*
- *For $q \in (1/2, 1)$, we have the convergence rate estimates*

$$\left\|x^k - \bar{x}\right\| = \mathcal{O}\left(k^{-\frac{1-q}{2q-1}}\right), \tag{3.12}$$

$$f(x^k) - f(\bar{x}) = \mathcal{O}\left(k^{-\frac{2-2q}{2q-1}}\right), \tag{3.13}$$

$$\left\|\nabla f(x^k)\right\| = \mathcal{O}\left(k^{-\frac{1-q}{2q-1}}\right). \tag{3.14}$$

Proof. Since ∇f is Lipschitz continuous with constant L, we deduce from the descent condition (2.1), the relationship $x^{k+1} = x^k - \frac{1}{L}g^k$, and the condition (3.11) that

$$\begin{aligned}
f(x^{k+1}) &\le f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\
&= f(x^k) + \left\langle \nabla f(x^k) - g^k, -\frac{1}{L}g^k \right\rangle - \frac{1}{L}\left\|g^k\right\|^2 + \frac{L}{2}\left\|\frac{1}{L}g^k\right\|^2 \\
&\le f(x^k) + \frac{1}{L}\left\|\nabla f(x^k) - g^k\right\| \cdot \left\|g^k\right\| - \frac{1}{2L}\left\|g^k\right\|^2 \\
&\le f(x^k) + \frac{\nu_2}{L}\left\|g^k\right\|^2 - \frac{1}{2L}\left\|g^k\right\|^2 \\
&= f(x^k) - \frac{1-2\nu_2}{2L}\left\|g^k\right\|^2 \quad \text{for all} \quad k \in \mathbb{N}.
\end{aligned} \tag{3.15}$$

It follows from $\nu_2 < \frac{1}{2}$ that the sequence $\{f(x^k)\}$ is decreasing. If $\|x^k\| \to \infty$ as $k \to \infty$, there is nothing to prove, and thus we suppose that $\{x^k\}$ has an accumulation point $\bar{x}$. Then $f(\bar{x})$ is an accumulation point of $\{f(x^k)\}$. Combining this with the decreasing property of $\{f(x^k)\}$, we deduce that $\lim_{k\in\mathbb{N}} f(x^k) = f(\bar{x})$, which implies that $f(x^k) - f(x^{k+1}) \to 0$. Then (3.15) tells us that $g^k \to 0$, which being combined with (3.11) gives us $\nabla f(x^k) \to 0$ as $k \to \infty$ and hence verifies (i).

To justify the assertions in (ii) and (iii), let $\bar{x}$ be some accumulation point of $\{x^k\}$, and let f satisfy the KL property at $\bar{x}$. It follows from (3.11) that

$$\left\|\nabla f(x^k)\right\| \le \left\|g^k\right\| + \left\|\nabla f(x^k) - g^k\right\| \le (1+\nu_2)\left\|g^k\right\| \quad \text{for all} \quad k \in \mathbb{N}. \tag{3.16}$$

Combining the latter with (3.15) and the recurrent relation $x^{k+1} = x^k - \frac{1}{L}g^k$ gives us the estimates

$$f(x^k) - f(x^{k+1}) \ge \frac{1-2\nu_2}{2L}\left\|g^k\right\|^2 = \frac{1-2\nu_2}{2}\left\|g^k\right\|\left\|\frac{1}{L}g^k\right\| \tag{3.17}$$

$$\ge \frac{1-2\nu_2}{2(1+\nu_2)}\left\|\nabla f(x^k)\right\| \cdot \left\|x^{k+1} - x^k\right\| \quad \text{for all} \quad k \in \mathbb{N}. \tag{3.18}$$

Thus assumption (H1) in Proposition 2.3 is satisfied with $\sigma = \frac{1-2\nu_2}{2(1+\nu_2)} > 0$. The second assumption (H2) in Proposition 2.3 is also satisfied since $f(x^k) = f(x^{k+1})$ yields $g^k = 0$ by (3.17) and hence $x^{k+1} = x^k$ by the relationship $x^{k+1} = x^k - \frac{1}{L}g^k$. Thus Proposition 2.3 brings us to $x^k \to \bar{x}$ as $k \to \infty$, which verifies (ii).

To justify (iii), we first employ Proposition 2.4 with $d^k = -g^k$ for all $k \in \mathbb{N}$ and $\tau = \frac{1}{L}$. By (3.17) and (3.16), both conditions in (2.3) are satisfied with $\alpha = \frac{1-2\nu_2}{2L}$ and $\beta = 1+\nu_2$. It follows from (3.16) and $\nabla f(x^k) \ne 0$ that $g^k \ne 0$ for all $k \in \mathbb{N}$. Combining this with $x^{k+1} = x^k - \frac{1}{L}g^k$, we get that $x^{k+1} \ne x^k$ for all $k \in \mathbb{N}$. Therefore, all the assumptions in Proposition 2.4 are satisfied, which verifies the convergence rate of $\{x^k\}$ to $\bar{x}$ stated in (iii). Since $\bar{x}$ is an accumulation point of $\{x^k\}$, we also deduce from (i) that $\bar{x}$ is a stationary point of f, i.e., $\nabla f(\bar{x}) = 0$. Therefore, it follows from the descent condition (2.1) and the decreasing property of $\{f(x^k)\}$ that

$$0 \le f(x^k) - f(\bar{x}) \le \left\langle \nabla f(\bar{x}), x^k - \bar{x}\right\rangle + \frac{L}{2}\left\|x^k - \bar{x}\right\|^2 = \frac{L}{2}\left\|x^k - \bar{x}\right\|^2,$$

which justifies the convergence rates of $\{f(x^k)\}$ to $f(\bar{x})$ as asserted in (iii).

It remains to verify the convergence rates for $\{\nabla f(x^k)\}$. Since ∇f is Lipschitz continuous with constant $L > 0$, this follows from the convergence rates for $\{x^k\}$ due to

$$\left\|\nabla f(x^k)\right\| = \left\|\nabla f(x^k) - \nabla f(\bar{x})\right\| \le L\left\|x^k - \bar{x}\right\|,$$

which therefore completes the proof of the theorem. □

Remark 3.3. By the triangle inequality, the estimate in (3.10) with $\nu_1 < \frac{1}{3}$ yields the one in (3.11) with $\nu_2 < \frac{1}{2}$. Therefore, Theorem 3.2 also verifies the global convergence of the sequences in the general inexact gradient methods under (3.10).

Now we are ready to establish the convergence properties of the proposed IGD Algorithm 1.

Theorem 3.4. *Consider Algorithm 1 with $\mu > 2$ and assume that $\nabla f(x^k) \ne 0$ for all $k \in \mathbb{N}$. Then we have that either $\left\|x^k\right\| \to \infty$, or the assertions in* (i)–(iii) *in Theorem 3.2 hold with $\varepsilon_k \downarrow 0$ as $k \to \infty$.*

Proof. It follows from (3.3) that condition (3.11) is satisfied with $\nu_2 = \frac{1}{\mu} < 2$. Then applying Theorem 3.2 yields its assertions (i)–(iii). The convergence to 0 of $\{g^k\}$ in (i) and the inequality $\left\|g^k\right\| > \mu\varepsilon_{k+1}$ for all $k \in \mathbb{N}$ from (3.3) tells us that $\varepsilon_k \downarrow 0$ and thus completes the proof. □

4 Gradient-Based Inexact Methods in Nonsmooth Convex Optimization

In this section, we employ the inexact gradient method for $\mathcal{C}^{1,1}$ optimization developed in Section 3 to design and justify two new gradient-based *inexact methods* to solve problems of *nonsmooth convex optimization*. The reductions of such nonsmooth problems to $\mathcal{C}^{1,1}$ optimization is conducted by using *smoothing procedures* via Moreau envelopes and proximal mappings.

The methods we develop in this way are the following:

(i) The *gradient-based inexact proximal point method* (GIPPM) to minimize l.s.c. convex functions.

(ii) The *gradient-based inexact augmented Lagrangian method* (GIALM) to solve nonsmooth convex programs with equality linear constraints.

Each of these two methods is considered in the corresponding subsection below. First we recall the appropriate constructions of convex analysis used in what follows; see [6, 36, 47] for more details. Given a proper (i.e., with $\operatorname{dom} g := \{x \in \mathbb{R}^n \mid g(x) < \infty\} \neq \emptyset$), l.s.c., convex function $g\colon \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$, the *subdifferential* of g at $x \in \operatorname{dom} g$ is defined by

$$\partial g(x) := \big\{v \in \mathbb{R}^n \;\big|\; \langle v, y - x\rangle \le g(y) - g(x) \;\text{ for all }\; y \in \mathbb{R}^n\big\}.$$

For any proximal parameter $\lambda > 0$, the *Moreau envelope* $e_\lambda g\colon \mathbb{R}^n \to \mathbb{R}$ and the *proximal mapping* $\operatorname{Prox}_{\lambda g}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ are defined, respectively, by

$$\begin{aligned} e_\lambda g(x) &:= \inf_{y\in\mathbb{R}^n}\left\{g(y) + \frac{1}{2\lambda}\|y - x\|^2\right\},\\ \operatorname{Prox}_{\lambda g}(x) &:= \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{g(y) + \frac{1}{2\lambda}\|y - x\|^2\right\}. \end{aligned} \tag{4.1}$$

The following result taken from [6, Proposition 12.28, 12.29] and [47, Theorem 2.26, Proposition 12.19] presents remarkable properties of Moreau envelopes and proximal mappings for convex functions that allow us to pass from convex nonsmooth optimization problems with extended-valued objectives (i.e., incorporating constraints) to problems of $\mathcal{C}^{1,1}$ optimization.

Lemma 4.1. *Let $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, l.s.c., convex function, and let $\lambda > 0$. Then the Moreau envelope $e_\lambda g$ is convex, $\mathcal{C}^1$-smooth, and has the Lipschitz continuous gradient on $\mathbb{R}^n$ with constant $1/\lambda$ that satisfies the relationship*

$$\nabla e_\lambda g(x) = \lambda^{-1}\Big(x - \operatorname{Prox}_{\lambda g}(x)\Big) \;\text{ for all }\; x \in \mathbb{R}^n. \tag{4.2}$$

Furthermore, the proximal mapping $\operatorname{Prox}_{\lambda g}$ is Lipschitz continuous on $\mathbb{R}^n$ with constant 1, and every stationary point of $e_\lambda g$ is a minimizer of g.

4.1 Gradient-Based Inexact Proximal Point Method

We consider here the class of optimization problems given in the form

$$\text{minimize} \quad g(x) \;\text{ subject to }\; x \in \mathbb{R}^n, \tag{4.3}$$

where $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, l.s.c., and convex function. Although problems (1.1) and (4.3) are written in the similar format, there is a huge difference between the $\mathcal{C}^{1,1}$ objective in (1.1) and the extended-real-valued objective in (4.3). Besides different differential properties of the objective functions in (1.1) and (4.3), observe that (1.1) is a problem of unconstrained optimization, while (4.3) incorporates constraints coming from $x \in \operatorname{dom} g$ when the function g is extended-real-valued.

The following gradient-based inexact proximal point method/algorithm GIPPM is now proposed to solve the general class of nonsmooth convex optimization problems (4.3).

Algorithm 2 (GIPPM)**.**

Step 0 (initialization)**.** Choose a proximal parameter $\lambda > 0$, initial point $x^1 \in \mathbb{R}^n$, initial error $\varepsilon_1 > 0$, error reduction factor $\theta \in (0, 1)$, and scaling factor $\mu > 2$. Set $k := 1$.

Step 1 (inexact proximal mapping)**.** Find p^k and the smallest natural number i_k such that

$$\Big\|p^k - \operatorname{Prox}_{\lambda g}(x^k)\Big\| \le \lambda\theta^{i_k}\varepsilon_k \quad\text{and}\quad \Big\|x^k - p^k\Big\| > \lambda\mu\theta^{i_k}\varepsilon_k. \tag{4.4}$$

Step 2 (error and iteration update)**.** Set $x^{k+1} := p^k$, $\varepsilon_{k+1} := \theta^{i_k}\varepsilon_k$. Increase k by 1 and go to Step 1.

Remark 4.2. Similarly to the discussions in Remark 3.1, the procedure of finding p^k and i_k that satisfy (4.4) can be given as follows. Set $i_k = 0$. Find some p^k such that

$$\left\|p^k - \mathrm{Prox}_{\lambda g}(x^k)\right\| \le \lambda\theta^{i_k}\varepsilon_k.$$

While $\left\|x^k - p^k\right\| \le \lambda\mu\theta^{i_k}\varepsilon_k$, increase i_k by 1 and recalculate p^k under the condition above. Then the existence of p^k and i_k satisfying (4.4) is guaranteed when $0 \notin \partial g(x^k)$. Indeed, otherwise for each $k \in \mathbb{N}$ we get a sequence $\{p_i^k\}$ as $i \in \mathbb{N}$ with

$$\left\|p_i^k - \mathrm{Prox}_{\lambda g}(x^k)\right\| \le \lambda\theta^i\varepsilon_k \text{ and } \left\|x^k - p_i^k\right\| \le \lambda\mu\theta^i\varepsilon_k \text{ for all } i \in \mathbb{N}.$$

Since $\theta \in (0,1)$, the latter inequalities mean that $x^k = \mathrm{Prox}_{\lambda g}(x^k)$, which is equivalent to

$$x^k = \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{g(x^k) + \frac{1}{2\lambda}\left\|x - x^k\right\|^2\right\}.$$

By using the subdifferential Fermat rule, we get $0 \in \partial g(x^k)$, a contradiction.

Now we are ready to establish convergence properties of Algorithm 2.

Theorem 4.3. *Let $\{x^k\}$ be the iterative sequence generated by Algorithm 2 with $\mu > 2$. Assume that $0 \notin \partial g(x^k)$ for all $k \in \mathbb{N}$. Then either $\|x^k\| \to \infty$ as $k \to \infty$, or we have the assertions:*

- **(i)** *Every accumulation point of $\{x^k\}$ is a minimizer of g, and both sequences $\{\|x^k - p^k\|\}$ and $\{\varepsilon_k\}$ converge to 0 as $k \to \infty$.*
- **(ii)** *If $e_\lambda g$ satisfies the KL property at an accumulation point $\bar{x}$ of $\{x^k\}$, then $x^k \to \bar{x}$ as $k \to \infty$.*
- **(iii)** *If $e_\lambda g$ satisfies the KL property at an accumulation point $\bar{x}$ of $\{x^k\}$ with $\psi(t) = Mt^q$ for some $M > 0$ and $q \in (0,1)$, then the following convergence rates are guaranteed as $k \to \infty$:*
  - *For $q \in (0, 1/2]$, the sequence $\{x^k\}$ converges linearly to $\bar{x}$ as $k \to \infty$.*
  - *For $q \in (1/2, 1)$, we have $\|x^k - \bar{x}\| = \mathcal{O}\big(k^{-\frac{1-q}{2q-1}}\big)$ as $k \to \infty$.*

Proof. Define $f := e_\lambda g$. By Lemma 4.1, we have that ∇f is Lipschitzian with constant $L = 1/\lambda$ and

$$\nabla f(x) = \lambda^{-1}\big(x - \mathrm{Prox}_{\lambda g}(x)\big) \text{ for all } x \in \mathbb{R}^n.$$

Defining $g^k := \lambda^{-1}(x^k - p^k)$, the inequalities in (4.4) can be rewritten as

$$\begin{aligned}\left\|g^k - \nabla f(x^k)\right\| &= \left\|g^k - \nabla e_\lambda g(x^k)\right\| \\ &= \left\|\lambda^{-1}(x^k - p^k) - \lambda^{-1}(x^k - \mathrm{Prox}_{\lambda g}(x^k))\right\| \\ &= \lambda^{-1}\left\|\mathrm{Prox}_{\lambda g}(x^k) - p^k\right\| \le \theta^{i_k}\varepsilon_k \text{ and}\end{aligned}$$

$$\left\|g^k\right\| = \lambda^{-1}\left\|x^k - p^k\right\| > \mu\theta^{i_k}\varepsilon_k,$$

respectively. Furthermore, it follows from Step 2 of Algorithm 2 that the iterative procedure in this algorithm can be represented by

$$x^{k+1} = p^k = x^k - \lambda g^k = x^k - \frac{1}{L}g^k,$$

Therefore, $\{x^k\}$ is the iterative sequence generated by Algorithm 1 with $f = e_\lambda g$. Then Theorem 3.4(i) tells us that every accumulation point of $\{x^k\}$ is a stationary point of $e_\lambda g$, which is actually a minimizer of g by Lemma 4.1. Finally, assertions (ii) and (iii) follow from Theorem 3.4(ii), (iii), respectively. □

Next we discuss some features of Algorithm 2 and its relationships with other developments.

Remark 4.4. Observe the following:

(i) By using the properties of semialgebraic functions listed in Remark 2.2 and the Moreau envelope construction in (4.1), we can verify that the assumption on the KL property of $e_\lambda g$ in Theorem 3.4 is satisfied if g is a semialgebraic function.

(ii) Applying Theorem 3.2 and using the arguments similar to Theorem 4.3 allow us to get the convergence properties for Algorithm 2 if (4.4) is replaced by more general conditions

$$\left\|p^k - \text{Prox}_{\lambda g}(x^k)\right\| \le \nu_2 \left\|x^k - p^k\right\| \quad \text{with} \quad \nu_2 < \frac{1}{2},$$

$$\left\|p^k - \text{Prox}_{\lambda g}(x^k)\right\| \le \nu_1 \left\|x^k - \text{Prox}_{\lambda g}(x^k)\right\| \text{ with } \nu_1 < \frac{1}{3}.$$

(iii) Let us highlight important improvements of our GIPPM over the two versions of the classical inexact proximal point method (IPPM) of Rockafellar [45] applied to (3.9) as follows:

$$x^{k+1} = p^k, \ \left\|p^k - \text{Prox}_{\lambda g}(x^k)\right\| \le \delta_k, \ \sum_{k=1}^{\infty} \delta_k < \infty, \tag{A}$$

$$x^{k+1} = p^k, \ \left\|p^k - \text{Prox}_{\lambda g}(x^k)\right\| \le \delta_k \left\|x^k - p^k\right\|, \ \sum_{k=1}^{\infty} \delta_k < \infty. \tag{B}$$

To this end, we mentioned first that in deriving convergence results, our approach in Algorithm 2 is by employing the IGD method applied to the Moreau envelope $e_\lambda g$, while the one by Rockafellar in (A) and (B) is using the properties related to the maximal monotonicity of the subgradient mapping ∂g. This creates various differences between our GIPPM and Rockafellar's IPPM. In comparison with (A) of IPPM, our GIPPM can achieve a linear convergence rate while the convergence rate of (A) is not established in [45]. In practice, the sequence of errors $\{\delta_k\}$ in (A) is usually chosen as $\delta_k := ck^{-p}$ with $p > 1$ and $c > 0$. If δ_k is too small, the procedure of finding approximate proximal points p^k takes a lot of time while if δ_k is too large, p^k becomes unreliable. Note that in each iteration, the GIPPM algorithm verifies whether the approximate proximal point is acceptable and then decreases the error if necessary.

Regarding the version (B) of IPPM, it follows from the choice of δ_k that $\delta_k \to 0$ as $k \to \infty$. This ensures that for any $\nu_2 < \frac{1}{2}$ there is $N \in \mathbb{N}$ with $\delta_k < \nu_2$, which gives us

$$\left\|p^k - \text{Prox}_{\lambda g}(x^k)\right\| \le \nu_2 \left\|x^k - p^k\right\| \quad \text{as} \quad k \ge N.$$

Therefore, the convergence properties of the algorithm with the procedure (B) can be deduced from (ii). In addition to the advantage on the generality in convergence analysis, our GIPPM method is also promising in practical implementation. Indeed, since ν_2 is a constant while δ_k is decreasing to 0, the procedure of finding p^k under the conditions in (4.4) of Algorithm 2 takes significantly less time than the procedure (B) of IPPM while maintaining impressive convergence results achieved in Theorem 4.3.

4.2 Gradient-Based Inexact Augmented Lagrangian Method

In this subsection, we propose the gradient-based inexact augmented Lagrangian method/algorithm GIALM to solve equality-constrained convex programs given in the form:

$$\begin{aligned} &\text{minimize} \quad h(x) \\ &\text{subject to} \quad Ax = b, \end{aligned} \tag{P}$$

where $h\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, l.s.c., convex function, and where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Assume in what follows that the feasible set $\{x \in \mathbb{R}^n \mid Ax = b\}$ of (P) is nonempty. Given a parameter $\lambda > 0$, for each pair $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$, consider the *Lagrange function* and the *augmented Lagrangian* defined as in [27], respectively, by

$$\ell(x, y) := h(x) + \langle y, Ax - b\rangle \quad \text{and} \quad \mathcal{L}_\lambda(x, y) := \ell(x, y) + \frac{\lambda}{2}\|Ax - b\|^2. \tag{4.5}$$

The *dual Lagrange function* $d\colon \mathbb{R}^m \to [-\infty, \infty)$ is given by $d(y) := \inf_{x\in\mathbb{R}^n} \ell(x, y)$, and the corresponding *dual problem* is formulated as

$$\begin{array}{ll} \text{maximize} & d(y)\\ \text{subject to} & y \in \mathbb{R}^m. \end{array} \tag{D}$$

It follows from [20, Lemma 9] that the function $g := -d$ is l.s.c. and convex. Therefore, when g is proper, its proximal mapping is well-defined and can be represented via the dual function d as

$$\operatorname{Prox}_{\lambda g}(y) = \operatorname*{argmin}_{w\in\mathbb{R}^m}\left\{g(w) + \frac{1}{2\lambda}\|w - y\|^2\right\} = \operatorname*{argmax}_{w\in\mathbb{R}^m}\left\{d(w) - \frac{1}{2\lambda}\|w - y\|^2\right\}. \tag{4.6}$$

Now we present the construction of the gradient-based inexact augmented Lagrangian method.

Algorithm 3 (GIALM)**.**

Step 0. Choose a starting point $y^1 \in \mathbb{R}^m$, proximal parameter $\lambda > 0$, initial error $\varepsilon_1 > 0$, error reduction factor $\theta \in (0, 1)$, and scaling factor $\mu > 2$. Set $k := 1$.

Step 1. Find some x^{k+1} and the smallest natural number i_k such that

$$\mathcal{L}_\lambda(x^{k+1}, y^k) - \inf_{x\in\mathbb{R}^n} \mathcal{L}_\lambda(x, y^k) \le \frac{\lambda\theta^{2i_k}\varepsilon_k^2}{2} \quad \text{and} \quad \left\|b - Ax^{k+1}\right\| > \mu\theta^{i_k}\varepsilon_k. \tag{4.7}$$

Step 2. Set $y^{k+1} := y^k + \lambda(Ax^{k+1} - b)$ and $\varepsilon_{k+1} := \theta^{i_k}\varepsilon_k$. Increase k by 1 and go back to Step 1.

As a part of our convergence analysis for GIALM, we derive the following two propositions of their independent interest. The first proposition provides a useful representation of the augmented Lagrangian function different from (4.5).

Proposition 4.5. *For any $z \in \mathbb{R}^n$, $\eta \in \mathbb{R}^m$, and $\lambda > 0$, we have the representation*

$$\mathcal{L}_\lambda(z, \eta) = \sup_{w\in\mathbb{R}^n}\left\{\ell(z, w) - \frac{1}{2\lambda}\|w - \eta\|^2\right\}.$$

Proof. Taking into account the construction of ℓ in (4.5), we see that

$$w \mapsto \ell(z, w) - \frac{1}{2\lambda}\|w - \eta\|^2 = h(z) + \langle w, Az - b\rangle - \frac{1}{2\lambda}\|w - \eta\|^2$$

is a quadratic concave function, and thus it attains the global maximum when $Az - b = \frac{1}{\lambda}(w - \eta)$, i.e., at $w = \lambda(Az - b) + \eta$. Thus we have

$$\begin{aligned}\sup_w\left\{\ell(z, w) - \frac{1}{2\lambda}\|w - \eta\|^2\right\} &= h(z) + \langle\lambda(Az - b) + \eta, Az - b\rangle - \frac{1}{2\lambda}\|\lambda(Az - b)\|^2\\ &= h(z) + \langle\eta, Az - b\rangle + \frac{\lambda}{2}\|Az - b\|^2\\ &= \mathcal{L}_\lambda(z, \eta),\end{aligned}$$

which therefore verifies the claim of the proposition. □

Next we obtain a relationship between the augmented Lagrangian (4.5) and the proximal mapping (4.6). The result of this type was stated in [46, Proposition 6] but only for the iterative sequence of the inexact augmented Lagrangian method for convex programs with inequality constraints. Using a similar technique, a general result for problem (P) without considering a specific iterative sequence is presented in the next proposition.

Proposition 4.6. *Suppose that the function g defined above is proper. Then we have*

$$\frac{1}{2\lambda}\left\|y+\lambda(Ax-b)-\operatorname{Prox}_{\lambda g}(y)\right\|^2 \le \mathcal{L}_\lambda(x,y)-\inf_{z\in\mathbb{R}^n}\mathcal{L}_\lambda(z,y)$$

for all pairs $(x,y)\in\mathbb{R}^n\times\mathbb{R}^m$.

Proof. Fix any $(x,y)\in\mathbb{R}^n\times\mathbb{R}^m$ and denote $p=y+\lambda(Ax-b)$. Take an arbitrary $\eta\in\mathbb{R}^m$ and define the function $\vartheta\colon\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ by

$$\vartheta(z,w):=\ell(z,w)-\frac{1}{2\lambda}\left\|w-\eta\right\|^2=h(z)+\langle w,Az-b\rangle-\frac{1}{2\lambda}\left\|w-\eta\right\|^2.$$

It is clear that the function ϑ is l.s.c. convex in z, and continuous concave in w with $\vartheta(0,w)\to-\infty$ as $\|w\|\to\infty$. Applying Proposition 4.5 and then [50, Theorem 2.7] to $\vartheta(z,w)$ with taking [50, Remark 2.2] into account, we get the equalities

$$\begin{aligned}\inf_z\mathcal{L}_\lambda(z,\eta)&=\inf_z\sup_w\left\{\ell(z,w)-\frac{1}{2\lambda}\left\|w-\eta\right\|^2\right\}\\&=\sup_w\inf_z\left\{\ell(z,w)-\frac{1}{2\lambda}\left\|w-\eta\right\|^2\right\}\\&=\sup_w\left\{d(w)-\frac{1}{2\lambda}\left\|w-\eta\right\|^2\right\}.\end{aligned}\tag{4.8}$$

Combining this with the construction of $\mathcal{L}_\lambda(x,y)$ in (4.5) gives us

$$\begin{aligned}\mathcal{L}_\lambda(x,y)+\frac{1}{\lambda}\langle p-y,\eta-y\rangle&=h(x)+\langle y,Ax-b\rangle+\frac{\lambda}{2}\left\|Ax-b\right\|^2+\langle Ax-b,\eta-y\rangle\\&=h(x)+\langle\eta,Ax-b\rangle+\frac{\lambda}{2}\left\|Ax-b\right\|^2\\&=\mathcal{L}_\lambda(x,\eta)\ge\inf_z\mathcal{L}_\lambda(z,\eta)\\&=\sup_w\left\{d(w)-\frac{1}{2\lambda}\left\|w-\eta\right\|^2\right\}\\&\ge d\big(\operatorname{Prox}_{\lambda g}(y)\big)-\frac{1}{2\lambda}\left\|\operatorname{Prox}_{\lambda g}(y)-\eta\right\|^2.\end{aligned}\tag{4.9}$$

Employing (4.8) again with $\eta:=y$ and taking into account the construction of the proximal mapping in (4.6), we arrive at the equalities

$$\begin{aligned}\inf_z\mathcal{L}_\lambda(z,y)&=\sup_w\left\{d(w)-\frac{1}{2\lambda}\left\|w-y\right\|^2\right\}\\&=d\big(\operatorname{Prox}_{\lambda g}(y)\big)\big)-\frac{1}{2\lambda}\left\|\operatorname{Prox}_{\lambda g}(y)-y\right\|^2.\end{aligned}$$

The latter expression combined with (4.9) tells us that

$$\mathcal{L}_\lambda(x,y)-\inf_z\mathcal{L}_\lambda(z,y)\ge\frac{1}{2\lambda}\Big(\left\|\operatorname{Prox}_{\lambda g}(y)-y\right\|^2-\left\|\operatorname{Prox}_{\lambda g}(y)-\eta\right\|^2-2\langle p-y,\eta-y\rangle\Big).$$

Choosing finally $\eta:=\operatorname{Prox}_{\lambda g}(y)+y-p$ brings us to

$$\begin{aligned}\mathcal{L}_\lambda(x,y)-\inf_z\mathcal{L}_\lambda(z,y)&\ge\frac{1}{2\lambda}\Big(\left\|\operatorname{Prox}_{\lambda g}(y)-y\right\|^2-\left\|p-y\right\|^2-2\langle p-y,\operatorname{Prox}_{\lambda g}(y)-p\rangle\Big)\\&=\frac{1}{2\lambda}\left\|p-\operatorname{Prox}_{\lambda g}(y)\right\|^2,\end{aligned}$$

which therefore completes the proof of the proposition. □

Remark 4.7. The procedure of finding x^{k+1} and i_k in Step 1 of Algorithm 3 can be performed similarly to Algorithm 1 and Algorithm 2 in the following way. Set $i_k = 0$ and find some x^{k+1} with

$$\mathcal{L}_\lambda(x^{k+1}, y^k) - \inf_{x\in\mathbb{R}^n} \mathcal{L}_\lambda(x, y^k) \le \frac{\lambda\theta^{2i_k}\varepsilon_k^2}{2}. \tag{4.10}$$

While $\|b - Ax^{k+1}\| \le \mu\theta^{i_k}\varepsilon_k$, increase i_k by 1 and recalculate x^{k+1} under condition (4.10). This procedure terminates finitely when the function g defined above is proper, and when y^k is not a solution to (D). Indeed, if (4.10) does not stop, for each $k \in \mathbb{N}$ we get $\{x_i^{k+1}\}$, $i \in \mathbb{N}$, such that

$$\mathcal{L}_\lambda(x_i^{k+1}, y^k) - \inf_{x\in\mathbb{R}^n} \mathcal{L}_\lambda(x, y^k) \le \frac{\lambda\theta^{2i}\varepsilon_k^2}{2} \quad \text{and} \quad \|b - Ax_i^{k+1}\| \le \mu\theta^i\varepsilon \ \text{ for all } \ i \in \mathbb{N}. \tag{4.11}$$

Combining the first inequality in (4.11) with Proposition 4.6 gives us

$$\frac{1}{2\lambda}\left\|y^k + \lambda(Ax_i^{k+1} - b) - \mathrm{Prox}_{\lambda g}(y^k)\right\|^2 \le \frac{\lambda\theta^{2i}\varepsilon_k^2}{2}.$$

Since $\theta \in (0,1)$, letting $i \to \infty$ with taking into account the second inequality in (4.11), we get that $y^k = \mathrm{Prox}_{\lambda g}(y^k)$, which implies by the subdifferential Fermat rule that $0 \in \partial g(y^k)$. Therefore, y^k is a solution to the dual problem (D) due to its convexity. This is a contradiction.

Now we are ready to establish the convergence results of our GIALM algorithm.

Theorem 4.8. *Let the sequences $\{x^k\}$, $\{y^k\}$ are generated by Algorithm 3. Assume that $\sup_{y\in\mathbb{R}^m} d(y) > -\infty$ and y^k is not a solution to (D) whenever $k \in \mathbb{N}$. Then either $\|y^k\| \to \infty$ as $k \to \infty$, or we have:*

- **(i)** *Every accumulation point of $\{y^k\}$ is a solution to (D). If in addition the sequence $\{y^k\}$ is bounded, then every accumulation point of $\{x^k\}$ is a solution to (P).*
- **(ii)** *If $e_\lambda g$ satisfies the KL property at an accumulation point $\bar{y}$ of $\{y^k\}$, then $y^k \to \bar{y}$ as $k \to \infty$.*
- **(iii)** *If $e_\lambda g$ satisfies the KL property at an accumulation point $\bar{y}$ of $\{y^k\}$ with $\psi(t) = Mt^q$ for some $M > 0$ and $q \in (0,1)$, then we get the convergence rates:*
  - *If $q \in (0, 1/2]$, then the sequence $\{y^k\}$ converges linearly to $\bar{y}$ as $k \to \infty$.*
  - *If $q \in (1/2, 1)$, then $\|y^k - \bar{y}\| = \mathcal{O}\big(k^{-\frac{1-q}{2q-1}}\big)$ as $k \to \infty$.*

Proof. (i) Since $\sup_{y\in\mathbb{R}^m} d(y) > -\infty$, the function $g = -d$ in (D) is proper. It follows from Proposition 4.6 and the first inequality in (4.7) that

$$\frac{1}{2\lambda}\left\|y^k + \lambda(Ax^{k+1} - b) - \mathrm{Prox}_{\lambda g}(y^k)\right\|^2 \le \frac{\lambda\theta^{2i_k}\varepsilon_k^2}{2}.$$

Defining $p^k := y^k + \lambda(Ax^{k+1} - b)$, the latter can be equivalently written as

$$\left\|p^k - \mathrm{Prox}_{\lambda g}(y^k)\right\| \le \lambda\theta^{i_k}\varepsilon_k \ \text{ for all } \ k \in \mathbb{N}.$$

Moreover, the second inequality in (4.7) also tells us that

$$\left\|y^k - p^k\right\| > \lambda\mu\theta^{i_k}\varepsilon_k \ \text{ for all } \ k \in \mathbb{N}.$$

Therefore, $\{y^k\}$ is the iterative sequence generated by Algorithm 2 to find minimizers of g. Then Theorem 4.3(i) tells us that every accumulation point of $\{y^k\}$ is a minimizer of g, i.e., it is a solution to (D). Furthermore, we have $\varepsilon_k \downarrow 0$ and $\|y^k - p^k\| \to 0$ as $k \to \infty$.

Assume next that $\{y^k\}$ is bounded and show that every accumulation point of $\{x^k\}$ is a solution to (P). Since $\|y^k - p^k\| \to 0$, we deduce from the construction of p^k that

$$Ax^k - b \to 0 \ \text{ as } \ k \to \infty. \tag{4.12}$$

Picking any feasible solution x^* to (P) tells us that $Ax^* = b$. It follows from the first inequality in (4.7) and Step 2 of Algorithm 3 that

$$\mathcal{L}_\lambda(x^{k+1}, y^k) \le \inf_{x\in\mathbb{R}^n} \mathcal{L}_\lambda(x, y^k) + \frac{\lambda\varepsilon_{k+1}^2}{2} \le \mathcal{L}_\lambda(x^*, y^k) + \frac{\lambda\varepsilon_{k+1}^2}{2}.$$

Combining this with the construction of $\mathcal{L}_\lambda$ in (4.5) and the feasibility of x^* gives us

$$\begin{aligned} h(x^{k+1}) + \left\langle y^k, Ax^{k+1} - b\right\rangle + \frac{\lambda}{2}\left\|Ax^{k+1} - b\right\|^2 &\le h(x^*) + \left\langle y^k, Ax^* - b\right\rangle + \frac{\lambda}{2}\|Ax^* - b\|^2 + \frac{\lambda}{2}\varepsilon_{k+1}^2 \\ &= h(x^*) + \frac{\lambda}{2}\varepsilon_{k+1}^2 \ \text{ for all } \ k \in \mathbb{N}. \end{aligned} \tag{4.13}$$

Taking now any accumulation point $\bar{x}$ of $\{x^k\}$, we find some infinite subset $J \subset \mathbb{N}$ such that $x^k \xrightarrow{J} \bar{x}$. Then it follows from (4.12) that $A\bar{x} = b$, which verifies the feasibility of $\bar{x}$. Invoking (4.12) again together with the boundedness of $\{y^k\}$ gives us the convergence

$$\left\langle y^{k-1}, Ax^k - b\right\rangle + \frac{\lambda}{2}\left\|Ax^k - b\right\|^2 \to 0 \ \text{ as } \ k \to \infty. \tag{4.14}$$

Combining the latter with the lower semicontinuity of h, the estimate (4.13), and the condition $\varepsilon_k \downarrow 0$ tells us that

$$\begin{aligned} h(\bar{x}) &\le \liminf_{k\xrightarrow{J}\infty} h(x^k) \\ &= \liminf_{k\xrightarrow{J}\infty} \left\{h(x^k) + \left\langle y^{k-1}, Ax^k - b\right\rangle + \frac{\lambda}{2}\left\|Ax^k - b\right\|^2\right\} \\ &\le \liminf_{k\xrightarrow{J}\infty} \left\{h(x^*) + \frac{\lambda}{2}\varepsilon_k^2\right\} = h(x^*). \end{aligned}$$

Since x^* is any feasible solution to (P), we conclude that $\bar{x}$ is an optimal solution to (P).

Finally, assertions (ii) and (iii) follow from Theorem 4.3(ii), (iii), respectively. □

5 Numerical Experiments

In this section, we compare numerical aspects of the newly designed GIALM and the classical IALM in solving the basic Lasso problem given by

$$\text{minimize} \quad \frac{1}{2}\|Ax - b\|^2 + \gamma\|x\|_1 \ \text{ subject to } \ x \in \mathbb{R}^n, \tag{5.1}$$

where A is an $m \times n$ matrix, b is a vector in $\mathbb{R}^m$, and $\gamma > 0$. By omitting constants and defining $p(x) := \gamma\|x\|_1$ and $c := A^*b$, we observe that Lasso problem (5.1) is equivalent to

$$\text{maximize} \quad -\frac{1}{2}\|Ax\|^2 + \langle c, x\rangle - p(x), \quad x \in \mathbb{R}^n \tag{5.2}$$

This is the dual problem of the following primal convex program with linear equality constraints:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|y\|^2 + p^*(z) \\ &\text{subject to} && -A^*y - z = -c, \end{aligned} \tag{5.3}$$

where the indicator function $p^*(\cdot) = \delta_{\gamma\mathbb{B}^\infty}(\cdot)$ is the convex conjugate of p with $\mathbb{B}^\infty := \{z \in \mathbb{R}^n \mid \max|z_i| \le \gamma\}$. Indeed, the corresponding Lagrange function and the dual function of (5.3) are given, respectively, by

$$\ell(y, z, x) = \frac{1}{2}\|y\|^2 + p^*(z) + \langle x, c - A^*y - z\rangle,$$

$$d(x)=\inf_{y,z}\ell(y,z,x)=\inf_{y}\left\{\frac{1}{2}\|y\|^2-\langle y,Ax\rangle\right\}+\inf_{z}\left\{p^*(z)-\langle x,z\rangle\right\}+\langle c,x\rangle$$
$$=-\frac{1}{2}\|Ax\|^2+\langle c,x\rangle-p^{**}(x).$$

Since p is convex and continuous, we know from basic convex analysis that $p^{**}=p$, and thus d is exactly the objective function in (5.2). Since (5.3) has the form of (P), we can apply GIALM to solve (5.3), which brings us to the following algorithm.

Algorithm 4 (GIALM for solving (5.3))**.**

Step 0. Choose a starting point $x^1\in\mathbb{R}^n$, proximal parameter $\lambda>0$, initial error $\varepsilon_1>0$, error reduction factor $\theta\in(0,1)$, and scaling factor $\mu>2$. Set $k:=1$.

Step 1. Find some (y^{k+1},z^{k+1}) and the smallest natural number i_k such that

$$\mathcal{L}_\lambda(y^{k+1},z^{k+1},x^k)-\inf_{y,z}\mathcal{L}_\lambda(y,z,x^k)\le\frac{\lambda\theta^{2i_k}\varepsilon_k^2}{2}\quad\text{and}\quad\left\|A^*y^{k+1}+z^{k+1}-c\right\|>\mu\theta^{i_k}\varepsilon_k.\tag{5.4}$$

Step 2. Set $x^{k+1}:=x^k-\lambda(A^*y^{k+1}+z^{k+1}-c)$ and $\varepsilon_{k+1}:=\theta^{i_k}\varepsilon_k$. Increase k by 1 and go to Step 1.

The augmented Lagrangian $\mathcal{L}_\lambda(y,z,x)$ in (5.4) is written now as

$$\mathcal{L}_\lambda(y,z,x)=\frac{1}{2}\|y\|^2+p^*(z)-\langle x,A^*y+z-c\rangle+\frac{\lambda}{2}\|A^*y+z-c\|^2.$$

Given $i_k\in\mathbb{N}$, we employ the technique in [34] to find (y^{k+1},z^{k+1}) that satisfy the first inequality in (5.4). Define the mapping $\Psi_k:\mathbb{R}^m\to\mathbb{R}^n$ and the function $\psi_k:\mathbb{R}^m\to\mathbb{R}$ by

$$\Psi_k(y):=\operatorname*{argmin}_z\mathcal{L}_\lambda(y,z,x^k)\quad\text{and}\quad\psi_k(y):=\inf_z\mathcal{L}_\lambda(y,z,x^k).\tag{5.5}$$

It follows from [34, Section 3.2] that

$$\Psi_k(y)=\operatorname{Prox}_{p^*/\lambda}\left(\frac{x^k}{\lambda}-A^*y+c\right),$$
$$\psi_k(y)=\frac{1}{2}\|y\|^2+\frac{1}{2\lambda}\left\|\operatorname{Prox}_{\lambda p}(x^k-\lambda(A^*y-c))\right\|^2-\frac{\|x^k\|^2}{2\lambda}.$$

In addition, ψ_k is strongly convex with constant 1 and $\mathcal{C}^1$-smooth with the Lipschitzian gradient

$$\nabla\psi_k(y)=y+A\operatorname{Prox}_{\lambda p}(x^k-\lambda(A^*y+c))\quad\text{for all}\quad y\in\mathbb{R}^m.$$

Note that there are explicit formulas for calculating $\operatorname{Prox}_{\lambda p}$ and $\operatorname{Prox}_{p^*/\lambda}$; see, e.g., [7, Section 6.9] and [21]. However, we do not need these formulas and exact computations of proximal mappings in our algorithms and numerical experiments, since we only focus on solving (4.7) *inexactly* to illustrate the efficiency of GIALM in what follows.

To proceed further, deduce from the definitions of Ψ_k and ψ_k in (5.5) the relationships

$$\psi_k(y)=\mathcal{L}_\lambda(y,\Psi_k(y),x^k)\quad\text{and}\quad\inf_{y,z}\mathcal{L}_\lambda(y,z,x^k)=\inf_y\psi_k(y).$$

As a consequence, (y^{k+1},z^{k+1}) satisfying the first inequality in (5.4) can be obtained from

$$\psi_k(y^{k+1})-\inf_y\psi_k(y)\le\frac{\lambda\theta^{2i_k}\varepsilon_k^2}{2}\quad\text{and}\quad z^{k+1}=\Psi_k(y^{k+1}).\tag{5.6}$$

The standard characterization of strong convexity gives us the error bound

$$\psi_k(y^{k+1})-\inf_y\psi_k(y)\le\frac{1}{2}\left\|\nabla\psi_k(y^{k+1})\right\|^2,$$

which implies that the first inequality in (5.6) is satisfied if

$$\left\|\nabla\psi_k(y^{k+1})\right\| \le \omega_k := \sqrt{\lambda}\theta^{i_k}\varepsilon_k. \tag{5.7}$$

Since $\nabla\psi_k$ is Lipschitz continuous, we can apply the gradient descent method to ψ_k with the stepsize $1/L$, where L is the Lipschitz constant of $\nabla\psi_k$, and find an approximate minimizer y^{k+1} under (5.7). We also put $\varepsilon_1 = 1$ and $\theta = 0.8$ in the initial setting of Algorithm 3. The two selections of scaling factor are $\mu = 3$ and $\mu = 1.1$, which correspond to the versions GIALM-3 and GIALM-1.1, respectively.

Now we recall the iterative procedure of classical IALM from [46] to solve (D). Given an initial point $x^1 \in \mathbb{R}^n$, IALM uses the following updates for (y^{k+1}, z^{k+1}) and then x^{k+1} at each iteration:

$$\mathcal{L}_\lambda(y^{k+1}, z^{k+1}, x^k) - \inf_{y,z} \mathcal{L}_\lambda(y, z, x^k) \le \delta_k^2 \ \text{ and } \ x^{k+1} = x^k - \lambda(A^*y^{k+1} + z^{k+1} - c), \tag{5.8}$$

where the sequence $\{\delta_k\}$ is positive and summable. Using arguments similar to the above tells us that for each $k \in \mathbb{N}$ the pair (y^{k+1}, z^{k+1}) satisfying the inequality in (5.8) can be obtained from

$$\left\|\nabla\psi_k(y^{k+1})\right\| \le \omega_k := \sqrt{2}\delta_k \ \text{ and } \ z^{k+1} = \Psi_k(y^{k+1}),$$

where ψ_k and Ψ_k are given in (5.5). In our implementation, we choose $\omega_k := k^{-q}$ for all $k \in \mathbb{N}$ with some $q > 1$ to ensure the summability of $\{\delta_k\}$. Consider also the two specific choices of $q = 2$ and $q = 3$, which correspond to the versions IALM-2 and IALM-1.5, respectively.

Our two numerical experiments presented below are conducted by using a computer with the parameters: 10th Gen Intel(R), Core(TM), i5-10400 (6-Core 12M Cache, 2.9GHz to 4.3GHz), 16GB RAM memory, and the code written in MATLAB R2022b.

5.1 An Example in Image Processing

The setup for the experiment in this subsection follows from that in [8]. More specifically, we first assume the reflexive boundary conditions [26] for the 256×256 cameraman test image. Then we let this image go through a Gaussian blur of size 9×9 and standard deviation 4 followed by a standard Gaussian noise with standard deviation 10^{-3}. Then vector b in (5.1) represents the observed image, and A is the blurred operator. The regularization parameter is chosen as $\gamma = 10^{-4}$. The proximal parameter for the tested methods is chosen as $\lambda = 5$, and the initial point is $x^1 = b$. In this numerical experiment, we first generate an (approximate) optimal value f^* with its associated (approximate) solution by running GIALM-1.1 in 1000 iterations. The original, blurred, and optimal images are presented in Figure 2.

Figure 2: Deblurring of the cameraman

Then we record the results obtained by the tested methods after running 500 iterations. Their function values compared with f^* and the total number of iterations in the subproblems they execute are given in Figure 3 below. It can be seen that GIALM methods have a better performance than the

classical IALM methods. Indeed, the left graph shows that after around 5s, the values generated by the GIALM methods are always lower than those for the IALM methods. This is explained in more detail in the right graph, where we see that the IALM methods execute a much larger total number of gradient descent iterations in comparison with the subproblems of the GIALM methods.

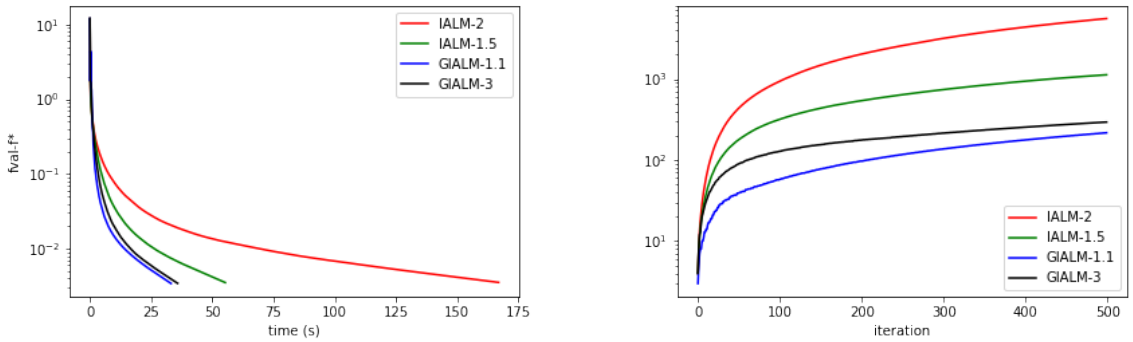

Figure 3: Function values (left) and total number of iterations in subproblems (right)

5.2 Randomly Generated Examples

In the second experiment, we test the GIALM and IALM methods with different sizes of data. To proceed, the $m \times n$ matrix A and vector $b \in \mathbb{R}^m$ are generated randomly with i.i.d. (identically and independently distributed) standard Gaussian entries. We employ the following stopping criterion used in [29, 30, 34] (see the discussions therein):

$$\eta_k := \frac{\left\| x^k - \text{Prox}_{\gamma\|\cdot\|_1}(x^k - A^*(Ax^k - b)) \right\|}{1 + \|x^k\| + \|Ax^k - b\|} \leq 10^{-6}. \tag{5.9}$$

We also stop the algorithms if they reach either the time limit of 4000 seconds, or the maximum number of iterations of 200,000. The initial points are chosen as $x^1 = 0 \in \mathbb{R}^n$ for all the algorithms. The selections of the tested parameter γ are 10^{-3} and $10^{-3}\|A^*b\|_\infty$, where $\|x\|_\infty := \max\{|x_i|, i = 1, \ldots, n\}$ for any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. While running the tests, we choose the proximal parameter as $\lambda = 0.01$ to get the best performance for the tested methods. The detailed information and the results are presented in the table below, where 'TN', 'iter', 'η_k', and 'time' stand, respectively, for test number, the total number of iterations, the residual (5.9) in the last iteration, and CPU running time of the methods. The tests using the regularization parameter $\gamma = 10^{-3}\|A^*b\|_\infty$ are signified by the asterisks (*) after their test numbers.

TN	m	n	IALM-1.5			IALM-2			GIALM-1.1			GIALM-3		
			iter	η_k	time	iter	η_k	time	iter	η_k	time	iter	η_k	time
1*	500	1000	40688	1.0E-06	68.19	40725	1.0E-06	524.31	40495	1.0E-06	20.78	40634	1.0E-06	25.08
2*	1000	1000	1963	1.0E-06	28.16	2110	1.0E-06	80.55	1985	1.0E-06	6.21	2065	1.0E-06	10.23
3*	1000	2000	28444	1.0E-06	879.53	8020	9.1E-05	4000	28613	1.0E-06	291.99	28546	1.0E-06	407
4*	2000	2000	2447	1.0E-06	979.43	2588	1.0E-06	3999.08	2520	1.0E-06	215.27	2560	1.0E-06	430.91
5*	2000	4000	3917	2.0E-04	4000	608	4.8E-03	4000	28929	1.0E-06	2594.49	21416	2.2E-06	4000
6*	4000	4000	952	9.8E-07	3701.33	153	7.8E-03	4000	1076	9.9E-07	1318.44	1114	1.0E-06	2704.56
7	500	1000	200000	2.1E-05	253.4	200000	2.1E-05	1802.58	200000	3.1E-05	96.91	200000	2.1E-05	98.29
8	1000	1000	75875	1.5E-04	4000	35188	5.2E-04	4000	200000	1.0E-05	233.99	200000	1.0E-05	244.08
9	1000	2000	87898	8.1E-05	4000	16867	7.2E-04	4000	200000	3.1E-05	1548.4	200000	2.1E-05	1626.42
10	2000	2000	1668	1.6E-02	4000	928	3.1E-02	4000	183214	6.1E-06	4000	109370	4.7E-05	4000
11	2000	4000	6852	1.7E-03	4000	2298	4.5E-03	4000	76571	3.6E-04	4000	58342	1.1E-04	4000
12	4000	4000	184	1.2E-01	4000	118	1.8E-01	4000	19637	1.4E-03	4000	6219	3.8E-03	4000

Table 1: Results of IALM and GIALM in random tests

It can be seen that in Tests 1-6, GIALM-1.1 exhibits the best performance since it achieves the main stopping criterion and has the lowest running time. In Test 7, GIALM-3 is the best since it achieves the smallest residual with the running time lower than IALM-1.5 and IALM-2. In the other

tests, GIALM methods also have better performances, i.e., they achieve a smaller residual with the running time less than for the IALM methods. The role of the controlling error of GIALM is presented clearly in Test 12, where GIALM-1.1 executes nearly 20,000 iterations while IALM-2 executes only 118 iterations and thus stagnates at a solution with a much larger residual η_k. This is because the error of the subproblems in IALM, not being controlled as in GIALM, becomes too small. Therefore, the subproblems of IALM waste much more time to stop. This observation is confirmed by the Figure 4 below, which illustrates the residual η_k and the error ω_k in the subproblems of the algorithms.

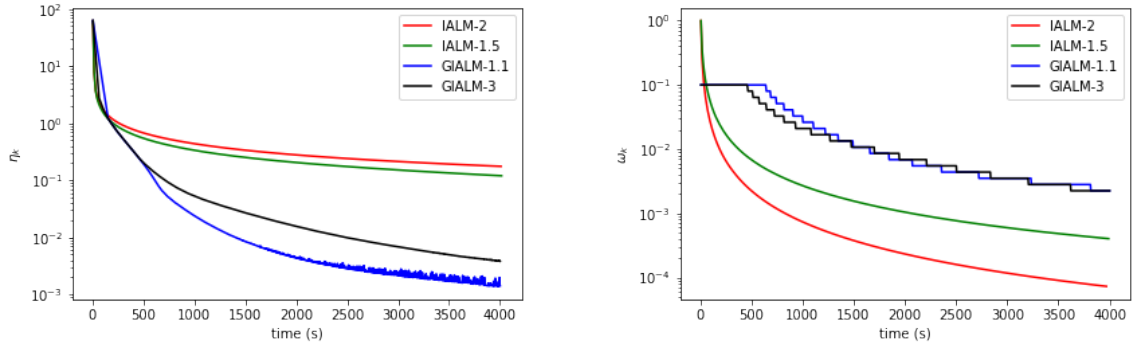

Figure 4: Value of residual η_k (left) and error ω_k (right) in Test 12

In conclusion, GIALM performs better than the classical IALM in these numerical experiments.

Acknowledgements

The authors are very grateful to the editors and the anonymous referees for their helpful remarks and suggestions, which allowed us to significantly improve the original presentation.

References

[1] P.-A. Absil, R. Mahony and B. Andrews, Convergence of the iterates of descent methods for analytic cost functions, *SIAM J. Optim.* **16** (2005), 531–547.

[2] M. Ahookhosh, K Amini and S. Bahrami, Two derivative-free projection approaches for systems of large-scale nonlinear monotone equations, *Numer. Algor.* **64** (2013), 21–42.

[3] H. Attouch, J. Bolte, P. Redont and A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Łojasiewicz property, *Math. Oper. Res.* **35** (2010), 438–457.

[4] H. Attouch, J. Bolte and B. F. Svaiter, Convergence of descent methods for semialgebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, *Math. Program.* **137** (2013), 91–129.

[5] C. Audet and W. Hare, *Derivative-Free and Blackbox Optimization*, Springer, Cham, Switzerland, 2017.

[6] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd edition, Springer, New York, 2017.

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

[8] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* **2** (2009), 183–202.

[9] A. S. Berahas, L. Cao, K. Choromanski and K. Scheinberg, A theoretical and empirical comparison of gradient approximations in derivative-free optimization, *Found. Comput. Math.* **22** (2022), 507–560.

[10] A. S. Berahas, L. Cao and K. Scheinberg, Global convergence rate analysis of a generic linesearch algorithm with noise, *SIAM J. Optim.* **31** (2021), 1489–1518.

[11] D. P. Bertsekas and J. N. Tsitsillis, Gradient convergence in gradient methods with errors, *SIAM J. Optim.* **10** (2000), 627–642.

[12] L. Bogolubsky, P. Dvurechensky, A. Gasnikov, G. Gusev, Yu. Nesterov, A. M. Raigorodskii, A. Tikhonov and M. Zhukovskii, Learning supervised pagerank with gradient-based and gradient-free optimization methods, *Adv. Neural Inf. Process. Syst.* **29** (2016), 1–9.

[13] L. Bottou, F. E. Curtis and J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.* **60** (2018), 223–311.

[14] R. H. Byrd, G. M. Chin and J. Nocedal, Sample size selection in optimization methods for machine learning, *Math. Program.* **134** (2012), 127–155.

[15] R. G. Carter, On the global convergence of trust region algorithms using inexact gradient information, *SIAM J. Numer. Anal.* **28** (1991), 251–265.

[16] A. R. Conn, K. Scheinberg and L. N. Vicente, *Introduction to Derivative-Free Optimization*, SIAM, Philadelphia, PA, 2009.

[17] H. B. Curry, The method of steepest descent for non-linear minimization problems, *Q. Appl. Math.* **2** (1944), 258–261.

[18] O. Devolder, F. Glineur and Yu. Nesterov, First-order methods of smooth convex optimization with inexact oracle, *Math. Program.* **146** (2014), 37–75.

[19] P. Dvurechensky, Gradient method with inexact oracle for composite non-convex optimization, arxiv:1703.09180 (2017).

[20] J. Eckstein and W. Yao, Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives, *Pacific J. Optim.* **11** (2015), 619–644.

[21] G. Chierchia, E. Chouzenoux, P. L. Combettes and J.-C. Pesquet, *The Proximity Operator Repository*, http://proximity-operator.net/index.html (2016).

[22] S. Ghadimi and G. Lan, Stochastic first- and zeroth-order methods for nonconvex stochastic programming, *SIAM J. Optim.* **23** (2013), 2341–2368.

[23] S. Ghadimi, G. Lan and H. Zhang, Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, *Math. Program.* **155** (2014), 438—-57.

[24] P. Gilmore and C. T. Kelley, An implicit filtering algorithm for optimization of functions with many local minima, *SIAM J. Optim.* **5** (1995), 269–285.

[25] A. Griewank and A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2nd edition, SIAM, Philadelphia, PA, 2008.

[26] P. C. Hansen, J. G. Nagy and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*, SIAM, Philadelphia, PH, 2006.

[27] M. R. Hestenes, Multiplier and gradient methods, *J. Optim. Theory Appl.* **4** (1969), 303–320.

[28] A. F. Izmailov and M. V. Solodov, *Newton-Type Methods for Optimization and Variational Problems*, Springer, New York, 2014.

[29] P. D. Khanh, B. S. Mordukhovich, V. T. Phat and D. B. Tran, Generalized damped Newton algorithms in nonsmooth optimization via second-order subdifferentials, *J. Global Optim.* **86** (2023), 93–122.

[30] P. D. Khanh, B. S. Mordukhovich, V. T. Phat and D. B. Tran, Globally convergent coderivative-based generalized Newton method in nonsmooth optimization, *Math. Program.*, doi:10.1007/s10107-023-01980-2 (2023).

[31] P. D. Khanh, B. S. Mordukhovich and D. B. Tran, Inexact reduced gradient methods in nonconvex optimization, *J. Optim. Theory Appl.*, doi: 10.1007/s10957-023-02319-9 (2024).

[32] K. Kurdyka, On gradients of functions definable in o-minimal structures, *Ann. Inst. Fourier* **48** (1998), 769–783.

[33] G. Lan and R. D. C. Monteiro, Iteration-complexity of first-order augmented Lagrangian methods for convex programming, *Math. Program.* **155** (2016), 511–547

[34] X. Li, D. Sun and K.-C. Toh, A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems, *SIAM J. Optim.* **28**, 433–458.

[35] S. Łojasiewicz, *Ensembles Semi-Analytiques*, Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette, France, 1965.

[36] B. S. Mordukhovich and N. M. Nam, *Convex Analysis and Beyond, I: Basic Theory*, Springer, Cham, Switzerland, 2022.

[37] V. Nedelcu, I. Necoara and Q. Tran-Dinh, Computational complexity of inexact gradient augmented Lagrangian methods: Application to constrained MPC, *SIAM J. Control Optim.* **52** (2014), 3109–3134.

[38] Yu. Nesterov, Smooth minimization of non-smooth functions, *Math. Program.* **103** (2005), 127–152.

[39] Yu. Nesterov, Universal gradient methods for convex optimization problems, *Math. Program.* **152** (2015), 381–404.

[40] Yu. Nesterov, *Lectures on Convex Optimization*, 2nd edition, Springer, Cham, Switzerland, 2018.

[41] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, New York, 2015.

[42] J. Nocedal and S. J. Wright, *Numerical Optimization*, New York, 1999.

[43] A. Patrascu, I. Necoara, and Q. Tran-Dinh, Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization, *Optim. Lett.* **111** (2017), 609–626.

[44] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[45] R. T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM J. Control Optim.* **14** (1976), 877–898.

[46] R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, *Math. Oper. Res.* **1** (1996) 97–116.

[47] R. T. Rockafellar and R. J-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[48] A. Themelis, B. Hermans and P. Patrinos, A new envelope function for nonsmooth DC optimization, *Proc. 59th IEEE Conf. Dec. Control*, pp. 4697–4702, 2020.

[49] A. Themelis, L. Stella and P. Patrinos, Forward–backward quasi-Newton methods for nonsmooth optimization problems, *Comput. Optim. Appl.* **67** (2017), 443–487.

[50] H. Tuy, *Convex Analysis and Global Optimization*, 2nd edition, Springer, New York, 2016.

[51] A. Vasin, A. Gasnikov, P. Dvurechenskiy and V. Spokoiny, Accelerated gradient methods with absolute and relative noise in the gradient, *Optim. Methods Softw.* **38** (2023), 1180—-1229.