Journal of Optimization Theory and Applications
https://doi.org/10.1007/s10957-023-02319-9

Inexact Reduced Gradient Methods in Nonconvex Optimization

Pham Duy Khanh[1] · Boris S. Mordukhovich[2] · Dat Ba Tran[2]

Received: 12 June 2023 / Accepted: 24 September 2023

Abstract

This paper proposes and develops new linesearch methods with inexact gradient information for finding stationary points of nonconvex continuously differentiable functions on finite-dimensional spaces. Some abstract convergence results for a broad class of linesearch methods are established. A general scheme for inexact reduced gradient (IRG) methods is proposed, where the errors in the gradient approximation automatically adapt with the magnitudes of the exact gradients. The sequences of iterations are shown to obtain stationary accumulation points when different stepsize selections are employed. Convergence results with constructive convergence rates for the developed IRG methods are established under the Kurdyka–Łojasiewicz property. The obtained results for the IRG methods are confirmed by encouraging numerical experiments, which demonstrate advantages of automatically controlled errors in IRG methods over other frequently used error selections.

Dedicated to the memory of Boris Polyak,
a great mathematician and incredible person

Communicated by Arkadi Nemirovski.

Research of this author is funded by Ho Chi Minh City University of Education Foundation for Science and Technology under Grant Number CS.2023.19.02TD
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.
Research of this author was partly supported by the US National Science Foundation under Grant DMS-1808978.

✉ Pham Duy Khanh
khanhpd@hcmue.edu.vn ; pdkhanh182@gmail.com

✉ Boris S. Mordukhovich
aa1086@wayne.edu

Extended author information available on the last page of the article

Published online: 19 October 2023

Keywords Nonconvex optimization · Inexact reduced gradient methods · Linesearch methods · Kurdyka–Łojasiewicz property · Convergence rates

Mathematics Subject Classification 90C30 · 90C52 · 49M05

1 Introduction

Consider the unconstrained optimization problem formulated as follows:

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

with a continuously differentiable ($\mathcal{C}^1$-smooth) objective function $f : \mathbb{R}^n \to \mathbb{R}$. One of the most natural and classical approaches to solve (1.1) is by using *linesearch methods*; see, e.g., [8, 10, 24, 38, 40, 45]. Given a starting point $x^1 \in \mathbb{R}^n$, such methods construct the iterative procedure

$$x^{k+1} := x^k + t_k d^k \;\; \text{for all} \;\; k \in \mathbb{N}, \tag{1.2}$$

where $t_k \geq 0$ is a stepsize at the kth iteration, and where the direction d^k satisfies the condition

$$\left\langle d^k, \nabla f(x^k) \right\rangle < 0.$$

The classical choice for the direction is $d^k = -\nabla f(x^k)$ when the resulting algorithm is known as the *gradient descent method*; see the aforementioned books and the references therein. If f is twice continuously differentiable ($\mathcal{C}^2$-smooth) and the Hessian matrix $\nabla^2 f(x^k)$ is positive-definite, then d^k is chosen by solving the linear equation

$$-\nabla f(x^k) = \nabla^2 f(x^k) d^k,$$

and it is known as a *Newton direction* [10, 21, 24]. Additionally, more general choices of descent directions widely used are the *gradient related* directions [10, Page 41], *directions satisfying an angle condition* [1, Page 541], etc. Together with the descent directions, stepsizes are usually chosen to ensure the decreasing property of the entire sequence $\left\{ f(x^k) \right\}$ or sometimes only its tail. Well-known stepsize selections are *constant stepsize*, *diminishing stepsize* (not summable), stepsizes following *Armijo rule*, and *Wolfe conditions*; see, e.g., [1, 8, 10, 24, 38, 40, 45].

The stationarity of accumulation points generated by linesearch methods with *gradient related* directions and stepsizes following the *Armijo rule* is established in [10, Proposition 1.2.1]. When the Lipschitz continuity of the gradient is additionally assumed, the same type of convergence is achieved if either the stepsize is constant and directions are gradient related, or the stepsize is diminishing and directions satisfy more involved conditions [10, Proposition 1.2.3]. The global convergence of some linesearch methods to an isolated stationary point relies on the *Ostrowski condition*; see [42] and [21, Theorem 8.3.9]. If there is no guarantee on the isolation of stationary points, algebraic geometry tools introduced by Łojasiewicz [36] and

Kurdyka [33] are used for linesearch methods with directions satisfying the *angle conditions* and the stepsize following the *Wolfe conditions*; see [1, Theorem 4.1]. While rates of convergence of general linesearch methods are not considered, some specific methods achieve certain convergence rates for particular classes of functions. For instance, the gradient descent method achieves a local linear rate of convergence if the objective function is twice differentiable with Lipschitz continuous Hessian as in [38, Theorem 1.2.4], and the (generalized) damped Newton method attains a superlinear convergence rate of under the positive-definiteness of the (generalized) Hessian and some additional assumptions; [27, Theorem 4.5]. Furthermore, a linear rate of convergence for the gradient descent method is achieved under either the *Polyak–Łojasiewicz condition* as in [43] and [26], or under the *weak convexity* of the objective function as in [48].

Due to its simplicity, the gradient descent method is broadly used to solve various optimization problems; see, e.g., [13, 17, 18, 39, 49]. However, errors in gradient calculations may appear in many situations, which can be found in practical problems arising, e.g., in the design of space trajectories [2] and computer-aided design of microwave devices [23]. Moreover, many nonsmooth optimization problems can be transformed into its smoothed versions by using Moreau envelopes [47] and forward-backward envelopes [50]. Nevertheless, gradients of smoothed functions cannot be usually computed precisely, and therefore, various gradient methods with *inexact gradient information* have been suggested. We mention the following major developments in this vein:

- Devolder et al. [20] introduce the notion of *inexact oracle* and analyze behavior of several first-order methods of *smooth convex* optimization employed such an oracle. Nesterov [37] develops new methods of this type for *nonsmooth convex* optimization in the framework of inexact oracle.
- Gilmore and Kelley [23] propose an *implicit filtering algorithm* to deal with certain box constrained optimization problems, where the objective function is a sum of a $\mathcal{C}^1$-smooth function with Lipschitz continuous gradient and a noise function.
- Bertsekas shows in [10, pp. 44–45] that if the objective function is $\mathcal{C}^1$-smooth with a Lipschitz continuous gradient and if the error of inexact gradient is either small relative to the norm of the exact gradient, or proportional to the stepsize, then convergence behavior of gradient methods is similar to the case where there are no errors.

Recently, [22] proposed a frequency-domain analysis of inexact gradient descent methods; [51] analyzed accelerated gradient methods with absolute and relative noise in the gradient; [35] presented a zero-order mini-batch stochastic gradient descent methods. All the convergence results for inexact gradient methods mentioned above assume that the objective function is *either* $\mathcal{C}^1$*-smooth* with a Lipschitz continuous gradient, *or convex*. To the best of our knowledge, general methods of solving nonconvex $\mathcal{C}^1$-smooth optimization problems with inexact information on non-Lipschitzian gradients are *not available* in the literature. One of the reasons for this is that verifying the descent property of the sequence of function values without the Lipschitz continuity of ∇f via the Armijo linesearch requires exact information on gradients. To deal with inexactness, we need a descent direction that allows us to *replace the Armijo* *linesearch procedure* by another one not demanding exact gradients. In addition, a practical inexact gradient method that uses constant stepsize for a general nonconvex $\mathcal{C}^1$-smooth function with the Lipschitz gradient is also not established. Although an inexact gradient method with constant stepsize is proposed in [10, pp. 44–45] by using an error smaller than the norm of the exact gradient, the problem of how to control this error while the exact gradient is unknown is still questionable.

Having in mind the above discussion, we introduce new *inexact reduced gradient* (IRG) methods to find stationary points for a general class of nonconvex $\mathcal{C}^1$-smooth functions. Although our proposed methods address *smooth* problems, some motivations for them partly come from a certain *nonsmooth* algorithm and *generalized differential* tools of variational analysis. Specifically, to find a Clarke stationary point of a nonsmooth locally Lipschitzian function, the *gradient sampling* (GS) method, introduced by Burke et al. [14] and modified by Kiwiel [30], approximates at each iteration the ε-*generalized gradient* by the convex hull of nearby gradients. In the GS method, the negative projection of the origin onto this convex hull is chosen as the descent direction and the stepsizes are chosen from the backtracking linesearch as in [30, Section 4.1]. Although the GS method works well for nonsmooth problems, using them for smooth functions seems to be challenging due to, in particular, the necessity to solve subproblems of finding projections onto convex hulls. However, replacing the ε-generalized gradient by the Fréchet-type ε-*subdifferential* makes our methods much simpler and suitable for smooth problems. Indeed, the latter construction for a $\mathcal{C}^1$-smooth function at the point in question is just the closed ball centered at its gradient with radius ε. Thus, the projection of the origin onto this ball has an explicit and simple form. The descent direction chosen by this projection also allows us to replace the exact gradient by its approximation and to use a linesearch procedure that does not require exact gradients. Developing this idea, we design our *inexact reduced gradient* methods with non-normalized directions together with some stepsize selections such as backtracking stepsize, constant stepsize, and diminishing stepsize. To the best of our knowledge, the IRG methods that we propose and develop in this paper are *completely new* even in the *exact case*. It should also be emphasized that the proposed IRG methods are *not special versions* of the GS one since the latter needs exact gradients at multiple points in each iteration, while the IRG methods need only *one inexact gradient*. Moreover, the iterative sequence of the GS method is chosen randomly, while IRG iterations are designed *deterministically*. Our main results include the following:

- Designing a *general framework* for IRG methods and revealing their basic properties. The inexact criterion used in IRG methods is universal and appears in various contexts of nonsmooth optimization as well as in its smooth derivative-free counterpart.
- Finding *stationary accumulation points* of iterations in the IRG methods with backtracking stepsizes as well as with either *constant stepsizes*, or *diminishing ones* under an additional descent condition on the objective functions.
- Obtaining the *global convergence* of iterations in the IRG methods with *constructive convergence rates* depending on the exponent of the imposed *Kurdyka–Łojasiewicz* (KL) *property* of the objective functions.

These results are achieved by using our newly developed scheme for general linesearch methods described in the following way. To begin with, some conditions are proposed to ensure the stationary of accumulation points in general linesearch methods. If the KL property is additionally assumed, then the global convergence of the iterative sequence to a stationary point is guaranteed. Moreover, the rates of convergence are established if the stepsize is bounded away from zero.

From a practical viewpoint, our IRG methods automatically adjust the errors required for finding approximate gradients, which will be shown to have numerical advantages over decreasing errors, e.g., $\varepsilon_k = k^{-p}$ as $p \geq 1$, that are frequently used in the existing methods [10, 20, 23]. To elaborate more on this issue, observe that since the magnitude of the exact gradient is small near the stationary points and is larger elsewhere, decreasing errors that do not take the information of the exact or inexact gradients into consideration may encounter the following phenomena:

- *Over approximation*, which happens when the magnitude of the exact gradient is large but the error is too small. In this case, the procedures of finding an approximate gradient may execute longer than needed to obtain a good approximation of the exact gradient.
- *Under approximation*, which happens when the magnitude of the exact gradient is small but the error is too large, which may lead us to an approximate gradient that is not good enough. As a consequence of using such an approximate gradient, the next iterative element can be worse instead of being better than the current one.

In contrast to methods using decreasing errors, our IRG methods, by performing a low-cost checking step in each iteration to determine whether the error for the approximation procedure should decrease or stay the same in the next iteration, use errors that automatically adapt with the magnitudes of the exact gradients to avoid the aforementioned phenomena and exhibit a better performance. Note that the bounded errors are not compared here for inexact gradient methods since they are even worse than the decreasing errors in the sense that employing them may cause the divergence in sequences of iterates, gradients, and function values as illustrated in [44, Section 4].

The rest of the paper is organized as follows. Section 2 discusses basic notions related to the methods. A unified convergence framework for general linesearch methods is developed in Sect. 3. In Sect. 4, we introduce a general form of IRG methods and investigate their principal properties. Our main results about the convergence behavior of the IRG methods with different stepsize selections are given in Sect. 5 by adapting the convergence framework of Section 3. The numerical experiments conducted in Sect. 6 support the theoretical results obtained in Sect. 5 and show that the IRG methods with the new type of automatically controlled errors have a better performance in comparison with the inexact proximal point method in the Least Absolute Deviations Curve-Fitting problem taken from [11]. In Sect. 6, we also compare numerically the performance of our IRG methods with those of the reduced gradient method and the gradient decent method employing the exact gradient calculations for two well-known benchmark functions in global optimization. The last Sect. 7 discusses some directions of our future research.

2 Linesearch Methods and Related Properties

First we recall some notions and notation frequently used in what follows. All our considerations are given in the space $\mathbb{R}^n$ with the Euclidean norm $\|\cdot\|$ and scalar/inner product $\langle\cdot,\cdot\rangle$. We use $\mathbb{N} := \{1, 2, \ldots\}$, $\mathbb{R}_+$, and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ to denote the collections of natural numbers, positive numbers, and the extended real line, respectively. The symbol $x^k \overset{J}{\to} \overline{x}$ means that $x^k \to \overline{x}$ as $k \to \infty$ with $k \in J \subset \mathbb{N}$. For a $\mathcal{C}^1$-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, $\bar{x}$ is a *stationary point* of f if $\nabla f(\bar{x}) = 0$. The function f is said to satisfy the *L-descent condition* with some $L > 0$ if

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

We see that L-descent condition (2.1) means that the graphs of the quadratic functions $f_{L,x}(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$ lie above that of f for all $x \in \mathbb{R}^n$. This condition is equivalent to the convexity of $\dfrac{L}{2}\|x\|^2 - f(x)$ [52, Lemma 4], while being a direct consequence of the L-Lipschitz continuity of ∇f, i.e., the Lipschitz continuity of ∇f with constant L; see, e.g., [10, Proposition A.24] and [24, Lemma A.11]. The converse implication holds when f is convex [6, 52] but fails otherwise. A major class of real-valued functions satisfying the L-descent condition but not having the Lipschitz continuous gradient is given by

$$f(x) := \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c - h(x),$$

where A is an $n \times n$ matrix, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function whose gradient is not Lipschitz continuous, e.g., $h(x) := \|Cx - d\|^4$, where C is an $m \times n$ matrix and $d \in \mathbb{R}^m$. Indeed, we can find some $L > 0$ such that the matrix $LI - A$ is positive-semidefinite, where I is the $n \times n$ identity matrix. It follows from the second-order characterization of convex functions that

$$\frac{L}{2}\|x\|^2 - \left(\frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c\right) = \frac{1}{2}\big\langle (LI - A)x, x\big\rangle - \langle b, x\rangle - c \ \text{ is convex.}$$

Combining this with the convexity of h, we get the convexity of $\dfrac{L}{2}\|x\|^2 - f(x)$, which means that f satisfies the L-descent property (2.1).

Even when ∇f is Lipschitz continuous with constant $L > 0$, f can satisfy the $\tilde{L}$-descent condition with $\tilde{L} < L$. For example, consider the univariate function f together with its gradient ∇f given by

$$f(x) = \begin{cases} \frac{3}{4}x^2 & \text{if } |x| < \frac{2}{3}, \\ -\frac{3}{2}x^2 + 3x - 1 & \text{if } \frac{2}{3} \leq x \leq 1, \\ -\frac{3}{2}x^2 - 3x - 1 & \text{if } -1 \leq x \leq -\frac{2}{3}, \\ |x| - \frac{x^2}{2} & \text{if } |x| > 1 \end{cases}$$

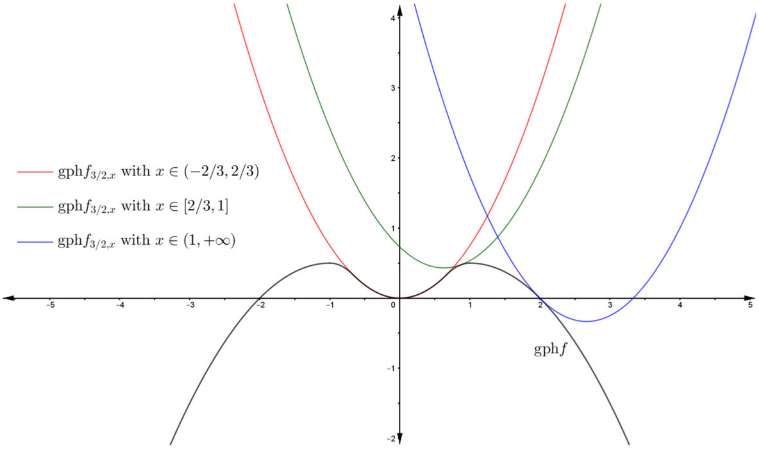

Fig. 1 An illustration for f and $f_{3/2,x}$

$$\text{and } \nabla f(x) = \begin{cases} \frac{3}{2}x & \text{if } |x| < \frac{2}{3}, \\ -3x + 3 & \text{if } \frac{2}{3} \le x \le 1, \\ -3x - 3 & \text{if } -1 \le x \le -\frac{2}{3}, \\ 1 - x & \text{if } x > 1, \\ -1 - x & \text{if } x < 1. \end{cases}$$

The latter representation implies that $L = 3$ is the smallest constant for the Lipschitz continuity of ∇f. Meanwhile, we see in Fig. 1 that f satisfies the $\tilde{L}$-descent property with $\tilde{L} = 3/2$.

Next we recall by following [10, Section 1.2] some basic stepsize selections for the iterative procedure (1.2). The stepsize sequence $\{t_k\}$ satisfies the *Armijo rule* if there exist a scalar β and a reduction factor $\gamma \in (0, 1)$ such that for all $k \in \mathbb{N}$ we have the representation

$$t_k = \max\left\{t \;\middle|\; f(x^k + td^k) \le f(x^k) + \beta t\left\langle \nabla f(x^k), d^k \right\rangle,\; t = 1, \gamma, \gamma^2, \ldots\right\}. \quad (2.2)$$

This stepsize selection ensures the nonincreasing property of the entire sequence $\left\{f(x^k)\right\}$. However, Armijo stepsizes may be small and thus require a large number of stepsize reducing steps in order to make just small changes of the iterative sequence.

To significantly simplify the iterative sequence design, it is possible to consider a *constant stepsize*, i.e., $t_k := \alpha$ for all $k \in \mathbb{N}$. For this rule, the nonincreasing property of $\left\{f(x^k)\right\}$ is not ensured in general but holds under the L-descent condition (2.1) whenever α is chosen to be sufficiently small with respect to $1/L$, see, e.g., [10, 40]. However, even when the L-descent condition is satisfied for f, while an approximate value of L is unknown, using constant stepsizes becomes inefficient. In such a case, it is possible to use the *diminishing stepsize* selection, i.e.,

$$t_k \downarrow 0 \quad \text{as} \quad k \to \infty \quad \text{and} \quad \sum_{k=1}^{\infty} t_k = \infty. \tag{2.3}$$

Drawbacks of the latter selection are the eventual *slow convergence* due to its small stepsizes and the absence of the descent property for the iterative sequence $\{f(x^k)\}$.

Now, we formulate a general type of directions that plays a crucial role in our subsequent analysis of various linesearch methods.

Definition 2.1 Let $\{x^k\}$ be a sequence in $\mathbb{R}^n$. The direction sequence $\{d^k\}$ is called *gradient associated* with $\{x^k\}$ if we have the implication

$$\text{whenever} \quad d^k \xrightarrow{J} 0 \quad \text{for some infinite set} \quad J \subset \mathbb{N}, \quad \text{it holds that} \quad \nabla f(x^k) \xrightarrow{J} 0. \tag{2.4}$$

It can be easily checked that if either

$$\lim_{k\to\infty} \left\| d^k - \nabla f(x^k) \right\| = 0, \tag{2.5}$$

or there exists some constant $c > 0$ such that

$$\left\| \nabla f(x^k) \right\| \le c \left\| d^k \right\| \quad \text{for all sufficiently large} \quad k \in \mathbb{N}, \tag{2.6}$$

then $\{d^k\}$ is *gradient associated* with $\{x^k\}$. Many methods such as the gradient descent, the generalized damped Newton method [27, 28], and the methods appeared in [10, Proposition 1.2.3] satisfy (2.6), while (2.5) can be considered as a standard condition for inexact gradient directions. It should be also mentioned that the notion of gradient associated directions is different from the notion of gradient-related directions proposed by Bertsekas in [10].

Finally in this section, we discuss two versions of the fundamental KL property playing a crucial role in the results on global convergence and convergence rates established in what follows. The first version, which is mainly used in the paper, is due to Absil et al. [1, Theorem 3.4].

Definition 2.2 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

$$\|\nabla f(x)\| \ge \psi\big(f(x) - f(\bar{x})\big) \tag{2.7}$$

for all $x \in U$ with $f(\bar{x}) < f(x) < f(\bar{x}) + \eta$.

Remark 2.3 By using rather standard arguments, we observe that for a smooth function f, the KL property from Definition 2.2 is weaker than the KL property of f at $\overline{x}$ introduced by Attouch et al. in [5]. It has been realized that KL property in the sense of Attouch et al., and hence, the one from Definition 2.2, is satisfied in broad settings. In particular, it holds at every *nonstationary point* of f; see [5, Lemma 2.1 and Remark 3.2(b)]. Furthermore, it is proved at the original paper by Łojasiewicz [36] that any analytic function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the KL property at every point $\overline{x}$ with $\varphi(t) = Mt^{1-q}$ for some $q \in [0, 1)$. Typical smooth functions satisfying this property are *semialgebraic* functions and also those from the more general class of functions *definable in o-minimal structures*; see [12, 33]. For other examples of functions satisfying the KL property, we refer the reader to [5, 34] and the bibliographies therein.

Next we present the convergence result for linesearch methods under the fulfillment of the *KL property*, which is taken from [1, Theorem 3.4].

Proposition 2.4 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function, and let the sequence of iterations $\{x^k\} \subset \mathbb{R}^n$ satisfy the following conditions:*

(H1) (primary descent condition). *There exists $\sigma > 0$ such that for sufficiently large $k \in \mathbb{N}$ we have*

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

(H2) (complementary descent condition). *For sufficiently large $k \in \mathbb{N}$, we have*

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

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 $k \to \infty$.

Although the convergence of linesearch methods under the KL property is widely exploited, the convergence rates of such methods under conditions (2.8) below have not been established. The following result presents convergence rates of general linesearch methods under these conditions. It should be noted that the proofs in [4, 41] for the convergence rates of proximal-type methods under the KL property cannot be generalized directly to Theorem 2.5. To the best of our knowledge, the results in [26, 43] are the closest to this theorem. However, the known results consider only the convergence of the exact gradient method for smooth functions with Lipschitz gradients under the Polyak–Łojasiewicz (PL) property [43]. Since the exact gradient method is a special case of the linesearch method, while the PL property is a special case of the KL property, we conclude that Theorem 2.5 has a broader range of applications than the results in [26, 43]. The proof of this result can be conducted similarly to the corresponding one from [3, Theorem 1] and thus is omitted.

Theorem 2.5 *Let the sequences* $\{x^k\} \subset \mathbb{R}^n$, $\{t_k\} \subset \mathbb{R}_+$ *and the numbers* $\beta > 0$, $c > 0$ *be such that* $x^{k+1} \neq x^k$ *for all* $k \in \mathbb{N}$, *and that we have*

$$f(x^k) - f(x^{k+1}) \geq \frac{\beta}{t_k} \left\| x^{k+1} - x^k \right\|^2 \ \textit{and} \ \left\| \nabla f(x^k) \right\| \leq \frac{c}{t_k} \left\| x^{k+1} - x^k \right\| \tag{2.8}$$

for sufficiently large $k \in \mathbb{N}$. *Suppose that the sequence* $\{t_k\}$ *is bounded away from* 0, *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* $q \in (0, 1)$. *The following convergence rates are guaranteed:*

(i) *If* $q \in (0, 1/2]$, *then the sequence* $\{x^k\}$ *converges linearly to* $\bar{x}$.
(ii) *If* $q \in (1/2, 1)$, *then there exists a positive constant* ϱ *such that*

$$\left\| x^k - \bar{x} \right\| \leq \varrho k^{-\frac{1-q}{2q-1}} \ \textit{for sufficiently large} \ k \in \mathbb{N}.$$

3 A Unified Convergence Framework for Some Linesearch Methods

In this section, we establish properties for a general class of linesearch methods of type (1.2), which provide major tools for convergence analysis of IRG methods in Sect. 5. One of the most important results desired for linesearch methods is as follows:

$$\text{every accumulation point of } \{x^k\} \text{ is a stationary point of } f. \tag{3.1}$$

By the continuity of the gradient mapping, the desired property (3.1) automatically holds if for each accumulation point $\bar{x}$ of $\{x^k\}$ we can find an infinite set $J \subset \mathbb{N}$ such that $x^k \xrightarrow{J} \bar{x}$ and $\nabla f(x^k) \xrightarrow{J} 0$. If the exact information on the gradient is unknown, while $\{d^k\}$ is gradient associated with $\{x^k\}$, i.e., (2.4) is satisfied, then property (3.1) is satisfied when

$$x^k \xrightarrow{J} \bar{x} \ \text{and} \ d^k \xrightarrow{J} 0 \ \text{for some infinite set} \ J \subset \mathbb{N}. \tag{3.2}$$

We are going to show that (3.2) holds whenever 0 is an accumulation point of $\{d^k\}$ and

$$\sum_{k=1}^{\infty} \left\| x^{k+1} - x^k \right\| \cdot \left\| d^k \right\| < \infty. \tag{3.3}$$

The following new result gives us a unified convergence analysis for many linesearch methods.

Lemma 3.1 *Let* $\{x^k\}$ *and* $\{d^k\}$ *be sequences satisfying* (3.3). *If* $\bar{x}$ *is an accumulation point of* $\{x^k\}$ *and if* 0 *is an accumulation point of* $\{d^k\}$, *then there exists an infinite* *set* $J \subset \mathbb{N}$ *such that*

$$x^k \xrightarrow{J} \bar{x} \;\; and \;\; d^k \xrightarrow{J} 0. \tag{3.4}$$

Proof If $x^k \to \bar{x}$ as $k \to \infty$, the conclusion obviously holds, so suppose that $x^k \nrightarrow \bar{x}$. It suffices to show that for any $\delta > 0$ sufficiently small and any $N \in \mathbb{N}$ there is a number $k_N \geq N$ such that

$$\|x^{k_N} - \bar{x}\| < \delta \;\text{ and }\; \|d^{k_N}\| < \delta.$$

Fix such $\delta > 0$ and $N \in \mathbb{N}$. Since δ is sufficiently small and $x^k \nrightarrow \bar{x}$, suppose that the set

$$A_1 := \left\{k \geq N \;\middle|\; \|x^k - \bar{x}\| \geq \delta\right\} \text{ is infinite.}$$

As $\bar{x}$ is an accumulation point of $\left\{x^k\right\}$ and 0 is an accumulation point of $\left\{d^k\right\}$, we get that

$$\begin{aligned} A_2 :&= \left\{k \geq N \;\middle|\; \|x^k - \bar{x}\| < \delta\right\} \text{ is infinite,}\\ A_3 :&= \left\{k \geq N \;\middle|\; \|x^k - \bar{x}\| < \delta/2\right\} \text{ is infinite,}\\ A_4 :&= \left\{k \geq N \;\middle|\; \|d^k\| < \delta\right\} \text{ is infinite.} \end{aligned}$$

It suffices to verify that $A_2 \cap A_4 \neq \emptyset$. Suppose on the contrary that $A_2 \cap A_4 = \emptyset$, i.e.,

$$\|d^k\| \geq \delta \;\text{ for all }\; k \in A_2.$$

By (3.3), we have the estimates

$$\begin{aligned} \infty > \sum_{k=1}^{\infty} \left\|x^{k+1} - x^k\right\| \cdot \left\|d^k\right\| &\geq \sum_{k \in A_2} \left\|x^{k+1} - x^k\right\| \cdot \left\|d^k\right\| \\ &\geq \delta \sum_{k \in A_2} \left\|x^{k+1} - x^k\right\|, \end{aligned}$$

which ensure the series convergence

$$\sum_{k \in A_2} \left\|x^{k+1} - x^k\right\| < \infty. \tag{3.5}$$

Taking any number $K \in A_3$, we also have $K \in A_2$. Since A_1 is infinite and A_1, A_2 form a partition of the set $\{N, N+1, \ldots\}$ including K, there exists a number $\widehat{K} \in A_1$ with $\widehat{K} > K$ such that $K, K+1, \ldots, \widehat{K} - 1 \in A_2$. Then, we have the estimates

$$\left\|x^{\widehat{K}} - x^K\right\| \geq \left\|x^{\widehat{K}} - \bar{x}\right\| - \left\|x^K - \bar{x}\right\| \geq \delta - \delta/2 = \delta/2.$$

Using the triangle inequality and (3.5) gives us

$$\delta/2 \le \left\| x^{\widehat{K}} - x^K \right\| \le \sum_{i=K}^{\widehat{K}-1} \left\| x^{i+1} - x^i \right\| \le \sum_{i \ge K,\ i \in A_2} \left\| x^{i+1} - x^i \right\| \xrightarrow[K \in A_3]{K \to \infty} 0,$$

which brings us to a contradiction that completes the proof of the lemma. □

Remark 3.2 Let us now present some observations in Lemma 3.1.

- **(i)** Lemma 3.1 develops a unified convergence analysis framework for linesearch methods (1.2) under a general condition on the directions d^k without specifying a class of functions f and stepsizes t_k. Note to this end that Bertsekas develops in [10, Section 1.2] some general schemes for linesearch methods (1.2) with *specific* classes of functions, directions, and stepsizes, while Absil et al. [1, Theorem 4.1] present convergence properties of linesearch methods for smooth functions under the angle and Wolfe conditions for directions. Devolder et al. [20] develop schemes only for convex functions. To the best of our knowledge, the framework presented in Lemma 3.1 is more general than the other schemes mentioned above.
- **(ii)** Since Lemma 3.1 does not require any assumption on f, its usage is not limited to linesearch methods for smooth functions. Convergence analysis of different nonsmooth optimization methods can be found in [15, 30–32, 41] and the references therein.
- **(iii)** If 0 is not an accumulation point of $\{d^k\}$, then (3.3) implies that $\{x^k\}$ is convergent. Indeed, the negation of the statement that 0 is an accumulation point of $\{d^k\}$ yields the existence of $\tau > 0$ and $K \in \mathbb{N}$ such that

$$\|d^k\| \ge \ \tau \text{ for all } \ k \ge K.$$

It follows from (3.3) that

$$\tau \sum_{k=K}^{\infty} \left\| x^{k+1} - x^k \right\| \le \sum_{k=K}^{\infty} \left\| x^{k+1} - x^k \right\| \cdot \left\| d^k \right\| < \infty,$$

which implies that $\{x^k\}$ is a Cauchy sequence, and thus, it converges.

Next we recall the classical results from [21, 42] that describe important properties of the set of accumulation points generated by a sequence satisfying the Ostrowski condition; see (3.6).

Lemma 3.3 *Let* $\left\{x^k\right\}$ *be a sequence satisfying the Ostrowski condition*

$$\lim_{k \to \infty} \|x^{k+1} - x^k\| = 0. \tag{3.6}$$

Then, the following assertions hold:

(i) *If $\{x^k\}$ is bounded, then the set of accumulation points of $\{x^k\}$ is nonempty, compact, and connected in $\mathbb{R}^n$.*
(ii) *If $\{x^k\}$ has an isolated accumulation point, then this sequence converges to it.*

Now, we are ready to establish the main result of this section revealing major convergence properties of a general class of linesearch methods.

Theorem 3.4 *Let $\{x^k\}$ be a sequence generated by a linesearch method (1.2) such that:*

(a) *$\{d^k\}$ is gradient associated with $\{x^k\}$;*
(b) 0 *is an accumulation point of* $\{d^k\}$;
(c) $\sum_{k=1}^{\infty} t_k \|d^k\|^2 < \infty.$

Then, every accumulation point of $\{x^k\}$ is a stationary point of f. Moreover, if $\{t_k\}$ is bounded from above, then the following assertions hold:

(i) *If $\{x^k\}$ is bounded, then the set of accumulation points of $\{x^k\}$ is nonempty, compact, and connected.*
(ii) *If $\{x^k\}$ has an isolated accumulation point, then this sequence converges to it.*

Proof Let $\bar{x}$ be an accumulation point of $\{x^k\}$. Note that (c) is equivalent to (3.3) under the linesearch relationship $x^{k+1} = x^k + t_k d^k$ in (1.2). Applying Lemma 3.1 with taking into account (b) and (c), we can find an infinite set $J \subset \mathbb{N}$ such that $x^k \overset{J}{\to} \bar{x}$ and $d^k \overset{J}{\to} 0$. Then, (a) implies that $\nabla f(x^k) \overset{J}{\to} 0$. Employing the continuity of ∇f, we have

$$\nabla f(\bar{x}) = \lim_{k \overset{J}{\to} \infty} \nabla f(x^k) = 0,$$

which tells us that $\bar{x}$ is a stationary point of f. Suppose now that $\{t_k\}$ is bounded from above by some $\tau > 0$. Using (c), we immediately get

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2 = \sum_{k=1}^{\infty} t_k^2 \|d^k\|^2 \leq \tau \sum_{k=1}^{\infty} t_k \|d^k\|^2 < \infty.$$

This leads us to $\left\|x^{k+1} - x^k\right\| \to 0$ and verifies assertions (i) and (ii) by applying Lemma 3.3. □

Theorem 3.4 also allows us to ensure the stationarity of accumulation points generated by linesearch methods (1.2) applied to functions satisfying the L-descent condition (2.1), where the stepsize is either constant or diminishing, and where the direction is *gradient associated* while satisfying the following *sufficient descent* condition

$$\langle \nabla f(x^k), d^k \rangle \leq -\kappa \|d^k\|^2 \ \text{ for all } \ k \in \mathbb{N} \tag{3.7}$$

with some constant $\kappa > 0$. Note that condition (3.7) is different from the gradient associated condition from Definition 2.1, since from (3.7) we only have that $\nabla f(x^k) \xrightarrow{J} 0$ yields $d^k \xrightarrow{J} 0$ but the reverse implication may not hold. In addition to the gradient descent and generalized Newton methods discussed above, there exist many other linesearch methods using direction (3.7), e.g., the boosted difference of convex functions algorithm as in [3, Proposition 4]), the inexact Levenberg–Marquardt method as in [19, Algorithm 3.1]), and the GS method for nonsmooth functions with non-normalized direction given in [30, Section 4.1].

We have the following effective consequence of Theorem 3.4.

Corollary 3.5 *Let $\{x^k\}$ be a sequence generated by a linesearch method* (1.2)*. Suppose that* $\inf f(x^k) > -\infty$*, and that we have the conditions:*

(a) *f satisfies the L-descent condition* (2.1) *for some $L > 0$;*
(b) *the sequence $\{d^k\}$ is gradient associated with $\{x^k\}$ and satisfies* (3.7) *for some $\kappa > 0$;*
(c) *the sequence $\{t_k\}$ is not summable, i.e.,*

$$\sum_{k=1}^{\infty} t_k = \infty, \tag{3.8}$$

and there are numbers $\delta > 0$ and $N \in \mathbb{N}$ such that

$$t_k \le \frac{2\kappa - \delta}{L} \ \textit{for all} \ k \ge N. \tag{3.9}$$

Then, every accumulation point of $\{x^k\}$ is stationary for f, and the assertions (i), (ii) in Theorem 3.4 *hold. Moreover, if $\{t_k\}$ is bounded away from* 0*, then $\nabla f(x^k) \to 0$ as $k \to \infty$.*

Proof It follows from (3.9) in condition (c) that

$$\kappa - \frac{Lt_k}{2} \ge \frac{\delta}{2} \ \text{for all} \ k \ge N.$$

Since f satisfies L-descent condition (2.1), we deduce from (3.7) and the latter inequality that

$$\begin{aligned} f(x^{k+1}) &\le f(x^k) + t_k \left\langle \nabla f(x^k), d^k \right\rangle + \frac{Lt_k^2}{2} \left\| d^k \right\|^2 \\ &\le f(x^k) - t_k \kappa \left\| d^k \right\|^2 + \frac{Lt_k^2}{2} \left\| d^k \right\|^2 \\ &= f(x^k) - t_k \left\| d^k \right\|^2 \left(\kappa - \frac{Lt_k}{2} \right) \\ &\le f(x^k) - \frac{\delta}{2} t_k \left\| d^k \right\|^2 \ \text{for all} \ k \ge N. \end{aligned} \tag{3.10}$$

Then, summing up the relationships $\frac{\delta}{2}t_k \left\|d^k\right\|^2 \le f(x^k) - f(x^{k+1})$ over $k = N, N+1, \ldots$ and using the assumption $\inf f(x^k) > -\infty$ give us

$$\sum_{k=N}^{\infty} t_k \left\|d^k\right\|^2 < \infty. \tag{3.11}$$

Now, we show that 0 is an accumulation point of $\left\{d^k\right\}$. Suppose on the contrary that there exist a positive number u and a natural number $K \ge N$ such that

$$\left\|d^k\right\| \ge u \;\text{ for all }\; k \ge K.$$

Using this together with (3.11) implies that $\sum_{k=K}^{\infty} t_k < \infty$, which contradicts (3.8). Therefore, 0 is an accumulation point of $\left\{d^k\right\}$. Combining the latter with (b), (3.11), and (3.9) allows us to confirm that all the assumptions of Theorem 3.4 are satisfied. Thus, every accumulation point of $\{x_k\}$ is a stationary point of f, and both assertions in (i) and (ii) hold.

If finally $\{t_k\}$ is bounded away from 0, it follows from (3.11) that $d^k \to 0$ as $k \to \infty$. Since the sequence $\left\{d^k\right\}$ is gradient associated with $\left\{x^k\right\}$ by (b), we get $\nabla f(x^k) \to 0$. □

By employing the iterative procedure $x^{k+1} = x^k + t_k d^k$, the conditions in (2.8) can be rewritten as the following estimates:

$$f(x^k) - f(x^{k+1}) \ge \beta t_k \left\|d^k\right\|^2 \;\text{ and }\; \left\|\nabla f(x^k)\right\| \le c \left\|d^k\right\|.$$

4 General Scheme for Inexact Reduced Gradient Methods

In this section, we design a general framework for our novel IRG methods and establish their basic properties prior to constructing particular methods of this type with various stepsize selections. Now, we are ready to formulate our *general algorithmic framework* (the *Master Algorithm*) for IRG methods without considering yet particular stepsize selections.

Algorithm 1 (**general framework for IRG methods**)

Step 0 (initialization) Select an initial point $x^1 \in \mathbb{R}^n$, initial radii $\varepsilon_1, r_1 > 0$, radius reduction factors $\mu, \theta \in (0, 1)$.

Step 1 (inexact gradient and stopping criterion) Choose g^k such that

$$\left\|g^k - \nabla f(x^k)\right\| \le \varepsilon_k. \tag{4.1}$$

Step 2 (radius update) If $\|g^k\| \le r_k + \varepsilon_k$, then set $r_{k+1} := \mu r_k$, $\varepsilon_{k+1} := \theta\varepsilon_k$, $d^k := 0$, and go to Step 3. Otherwise, set $r_{k+1} := r_k$, $\varepsilon_{k+1} := \varepsilon_k$, and

$$d^k := -\frac{\|g^k\| - \varepsilon_k}{\|g^k\|} g^k. \tag{4.2}$$

Step 3 (stepsize) Choose $t_k > 0$ by a specific rule.

Step 4 (iteration update) Set $x^{k+1} := x^k + t_k d^k$.

Step 5 Increase k by 1 and go back to Step 1.

Let us make some comments on the constructions of Algorithm 1. The first remark concerns the novelty in the choice of errors and directions.

Remark 4.1 We have the following observations on the error criterion used in Algorithm 1:

- **(i)** The inexact criterion (4.1) is universal and appears in many contexts even when only information on function values is available as in derivative-free optimization [16]. In addition, as shown in [20, Section 2.2], condition (4.1) is also satisfied if f is convex, smooth, and equipped with a first-order oracle, which covers various well-known models in nonsmooth optimization, e.g., Nesterov's smoothing techniques, Moreau–Yosida regularization, augmented Lagrangians as in [20, Section 3] and [29, Example 1].
- **(ii)** From (4.1), the radius ε_k can be considered as an automatically controlled error for the calculation of $\nabla f(x^k)$, which does not need to decrease after each iteration. This is different than the choice $\varepsilon_k = ck^{-p}$ for $p \ge 1, c > 0$ frequently used in the well-known methods [10, 20, 23]. Moreover, Steps 1 and 2 also show that $\|\nabla f(x^k)\| \le r_k + 2\varepsilon_k$ when ε_k is reduced. Therefore, we can conclude intuitively that $\|\nabla f(x^k)\|$ is decreasing when ε_k is decreasing.
- **(iii)** In the *exact case* when $g^k = \nabla f(x^k)$ for all $k \in \mathbb{N}$, we label our methods as the *reduced gradient* (RG) ones, which are different from the standard gradient descent method. Indeed, it follows from Step 2 that d^k is either 0 or is given by

$$d^k = -\left(\frac{\|\nabla f(x^k)\| - \varepsilon_k}{\|\nabla f(x^k)\|}\right) \nabla f(x^k). \tag{4.3}$$

  Therefore, the vector d^k in (4.3) has the same direction as $-\nabla f(x^k)$, but its length is $\|\nabla f(x^k)\|$ reduced by ε_k for each $k \in \mathbb{N}$.

Further we discuss and illustrate behavior of Algorithm 1 at the major steps of iterations.

Remark 4.2 Notice first that:

- **(i)** If $d^k \neq 0$, it follows from (4.2) and the definition of projections that $d^k = -\text{Proj}(0, \mathbb{B}(g^k, \varepsilon_k))$.

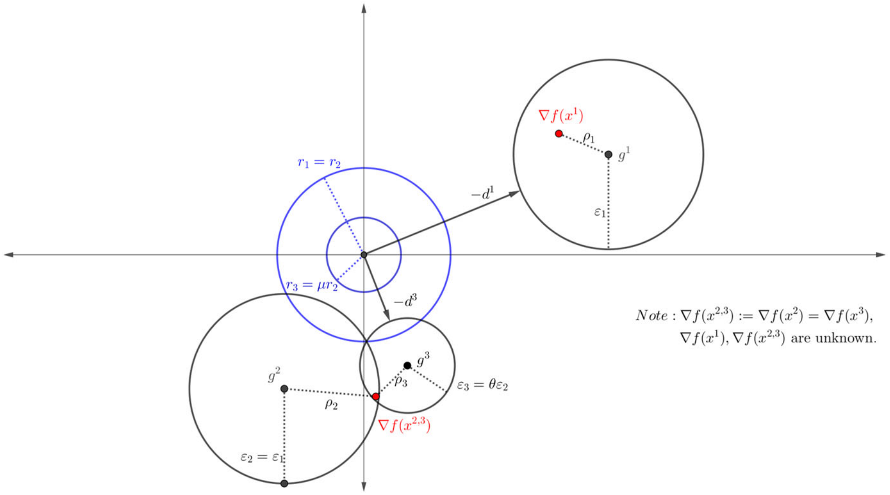
$\nabla f(x^1)$
ρ_1
g^1
$r_1 = r_2$
$-d^1$
ε_1
$r_3 = \mu r_2$
$-d^3$
$Note : \nabla f(x^{2,3}) := \nabla f(x^2) = \nabla f(x^3)$,
$\nabla f(x^1), \nabla f(x^{2,3})$ are unknown.
g^3
g^2
ρ_3
$\varepsilon_3 = \theta\varepsilon_2$
ρ_2
$\nabla f(x^{2,3})$
$\varepsilon_2 = \varepsilon_1$

Fig. 2 An illustration for IRG methods

(ii) An illustration for Algorithm 1 can be seen in Fig. 2.
Let g^1 be an approximate gradient of $\nabla f(x^1)$ at the 1st iteration. Then, Fig. 2 shows that the two balls $\mathbb{B}(g^1, \varepsilon_1)$ and $\mathbb{B}(0, r_1)$ do not intersect. This means by Step 2 that $r_2 = r_1$, $\varepsilon_2 = \varepsilon_1$, and $d^1 = -\text{Proj}(0, \mathbb{B}(g^1, \varepsilon_1))$. Then, we have a new point $x^2 = x^1 + t_1 d^1$ after choosing the stepsize $t_1 > 0$ as in Step 3 and Step 4.
At the 2nd iteration, it can be seen in Fig. 2 that the two balls $\mathbb{B}(g^2, \varepsilon_2)$ and $\mathbb{B}(0, r_2)$ intersect each other. Thus by Step 2 of Algorithm 1, the radii r_2, ε_2 are reduced to $r_3 = \mu r_2$ and $\varepsilon_3 = \theta\varepsilon_2$, while the direction d^2 is zero. The latter means that the iterative point x^2 stays the same, i.e., $x^3 = x^2$ from Step 4.
At the 3rd iteration, although $\nabla f(x^3) = \nabla f(x^2)$, we still need to recalculate an approximate gradient g^3 with a new error ε_3. In this iteration, the two balls $\mathbb{B}(g^3, \varepsilon_3)$ and $\mathbb{B}(0, r_3)$ do not intersect, and hence, the procedure is similar to that at the first iteration.

(iii) For each $k \in \mathbb{N}$, we have from Step 2 and Step 3 the equivalences

$$x^{k+1} = x^k \Longleftrightarrow d^k = 0 \Longleftrightarrow r_{k+1} = \mu r_k \Longleftrightarrow \varepsilon_{k+1} = \theta\varepsilon_k \Longleftrightarrow \left\| g^k \right\| \le r_k + \varepsilon_k. \tag{4.4}$$

The next proposition verifies the *decent property* of Algorithm 1.

Proposition 4.3 *In Algorithm 1, $\{d^k\}$ satisfies the sufficient descent condition with constant* 1, *i.e.,*

$$\left\langle \nabla f(x^k), d^k \right\rangle \le -\left\| d^k \right\|^2 \quad \textit{for all} \;\; k \in \mathbb{N}. \tag{4.5}$$

Proof Note that (4.5) automatically holds if $d^k = 0$. Supposing that $d^k \neq 0$ and using the construction of d^k in Step 2, we have the expression

$$-d^k = \mathrm{Proj}\big(0, \mathbb{B}(g^k, \varepsilon_k)\big).$$

It follows from (4.1) that $\|g^k - \nabla f(x^k)\| \leq \varepsilon_k$, which means that $\nabla f(x^k) \in \mathbb{B}(g^k, \varepsilon_k)$. Invoking the projection description for convex sets yields

$$\left\langle 0 + d^k, \nabla f(x^k) + d^k \right\rangle \leq 0,$$

which is in turn equivalent to

$$\left\langle \nabla f(x^k), d^k \right\rangle \leq - \left\| d^k \right\|^2$$

and thus verifies the claim in (4.5). □

Now, we introduce the notion of null iterations and establish some properties of such iterations related to the IRG methods.

Definition 4.4 The kth iteration of Algorithm 1 is called a *null iteration* if $x^{k+1} = x^k$. The set of all null iterations is denoted by

$$\mathcal{N} := \left\{k \in \mathbb{N} \mid x^{k+1} = x^k\right\}.$$

The next proposition collects important properties of null iterations.

Proposition 4.5 *Let* $\left\{x^k\right\}, \left\{g^k\right\}, \left\{d^k\right\}, \{\varepsilon_k\}$, *and* $\{r_k\}$ *be sequences generated by Algorithm* 1. *The following assertions hold:*

(i) $k \in \mathcal{N}$ *if and only if either one of the equivalent conditions in* (4.4) *holds.*

(ii) $\varepsilon_k \downarrow 0$ *if and only if* $r_k \downarrow 0$, *which is equivalent to the set* $\mathcal{N}$ *being infinite.*

(iii) *If* $\mathcal{N}$ *is finite, then we have*

$$\left\|g^k\right\| > r_N + \varepsilon_N \ \ \textit{and} \ \ \left\|d^k\right\| > r_N$$

for all $k \geq N$, *where* $N := \max \mathcal{N} + 1$.

(iv) *If* $\mathbb{N} \setminus \mathcal{N}$ *is finite, then* $\nabla f(x^K) = 0$ *and* $\left\{x^k\right\}_{k \geq K}$ *is a constant sequence, where we denote* $K := \max\{\mathbb{N} \setminus \mathcal{N}\} + 1$. *Otherwise,* $\bar{x}$ *is an accumulation point of* $\left\{x^k\right\}$ *if and only if it is an accumulation point of* $\left\{x^k\right\}_{k \in \mathbb{N} \setminus \mathcal{N}}$, *and therefore* $x^k \to \bar{x}$ *if and only if* $x^k \xrightarrow{\mathbb{N}\setminus\mathcal{N}} \bar{x}$.

Proof Assertions (i) and (ii) follow directly from Definition 4.4. To verify (iii), observe that for any natural number $k \geq N$, the kth iteration is not a null one and then deduce from (i) that

$$\varepsilon_{k+1} = \varepsilon_k = \varepsilon_N, \quad r_{k+1} = r_k = r_N, \quad \text{and} \quad \left\|g^k\right\| > r_k + \varepsilon_k = r_N + \varepsilon_N.$$

Together with Step 2 in Algorithm 1, this ensures that

$$\left\|d^k\right\| = \left\|g^k\right\| - \varepsilon_k > r_k = r_N.$$

which readily justifies assertion (i).

The proof of (iv) is a bit more involved. Supposing that the set $\mathbb{N} \setminus \mathcal{N}$ is finite, we have that $k \in \mathcal{N}$ for all $k \geq K$. This means by (i) that

$$x^{k+1} = x^k,\ \ \varepsilon_{k+1} = \theta\varepsilon_k,\ \ r_{k+1} = \mu r_k,\ \ \text{and}\ \ \left\|g^k\right\| \leq r_k + \varepsilon_k\ \ \text{whenever}\ \ k \geq K.$$

This tells us that $x^k = x^K$ for all $k \geq K$, and that $\varepsilon_k \downarrow 0,\ r_k \downarrow 0$, and $g^k \to 0$ as $k \to \infty$. Taking the limit as $k \to \infty$ in $\|g^k - \nabla f(x^k)\| \leq \varepsilon_k$ gives us $\nabla f(x^k) \to 0$, which yields $\nabla f(x^K) = 0$.

Supposing otherwise that the set $\mathbb{N} \setminus \mathcal{N}$ is infinite, we obviously get that every accumulation point of $\left\{x^k\right\}_{k\in\mathbb{N}\setminus\mathcal{N}}$ is an accumulation point of $\left\{x^k\right\}$. Conversely, taking any accumulation point $\bar{x}$ of $\left\{x^k\right\}$, it suffices to show that

$$\text{for any}\ \ \delta > 0,\ N \in \mathbb{N}\ \ \text{there exists}\ \ k_N \in \mathbb{N} \setminus \mathcal{N},\ k_N \geq N\ \ \text{with}\ \ \left\|x^{k_N} - \bar{x}\right\| < \delta.$$

To verify this, fixing $\delta > 0$ and $N \in \mathbb{N}$ and remembering that $\bar{x}$ is an accumulation point of $\left\{x^k\right\}$, we find $K \geq N$ such that $\|x^K - \bar{x}\| < \delta$. If $K \in \mathbb{N} \setminus \mathcal{N}$, choose $k_N := K$. Otherwise, using that $\mathbb{N} \setminus \mathcal{N}$ is infinite allows us to find $\widehat{K} \in \mathbb{N} \setminus \mathcal{N}$ for which $K, K+1, \ldots, \widehat{K} - 1 \in \mathcal{N}$. This ensures that

$$x^{\widehat{K}} = x^{\widehat{K}-1} = \ldots = x^{K+1} = x^K,$$

and therefore, with $k_N := \widehat{K}$, we get that $\|x^{k_N} - \bar{x}\| < \delta$. Since δ was chosen arbitrarily, this clearly shows that $\bar{x}$ is an accumulation point of $\left\{x^k\right\}_{k\in\mathbb{N}\setminus\mathcal{N}}$ and thus completes the proof. □

The last proposition here establishes relationships between convergence properties of the sequences $\left\{g^k\right\}$ and $\left\{d^k\right\}$ in Algorithm 1.

Proposition 4.6 *Let $\left\{g^k\right\}, \left\{d^k\right\}$, $\{\varepsilon_k\}$, and $\{r_k\}$ be sequences generated by Algorithm 1. Then, for any $k \in \mathbb{N}$ we have the estimates*

$$\left\|d^k\right\| \leq \left\|g^k\right\| \leq \left\|d^k\right\| + \varepsilon_k + r_k. \tag{4.6}$$

Consequently, the following assertions hold:

(i) $\varepsilon_k \downarrow 0$ *if and only if there is an infinite set $J \subset \mathbb{N}$ such that* $g^k \xrightarrow{J} 0$.

(ii) *For any infinite set $J \subset \mathbb{N}$, we have the equivalence*

$$g^k \xrightarrow{J} 0 \Longleftrightarrow d^k \xrightarrow{J} 0.$$

Proof Fix any $k \in \mathbb{N}$. If $k \in \mathcal{N}$, then we get by Proposition 4.5(i) that $d^k = 0$ and $\|g^k\| \le \varepsilon_k + r_k$. Otherwise, Step 2 in Algorithm 1 yields $\|d^k\| = \|g^k\| - \varepsilon_k \le \|g^k\|$. In both cases, (4.6) holds.

To deduce (i) from (4.6), suppose that $\varepsilon_k \downarrow 0$ as $k \to \infty$. By Proposition 4.5(ii) we have that $r_k \downarrow 0$ and the set $\mathcal{N}$ is infinite. Then, for any $\delta > 0$ and $N \in \mathbb{N}$ there is $k \ge N$ with $k \in \mathcal{N}$ such that

$$\left\|g^k\right\| \le r_k + \varepsilon_k < \delta,$$

where the first inequality follows from (4.4). Thus, we can construct an infinite set $J \subset \mathbb{N}$ such that $g^k \xrightarrow{J} 0$. If conversely the sequence $\{\varepsilon_k\}$ does not converge to 0, then the set $\mathcal{N}$ is finite by Proposition 4.5(ii). Using Proposition 4.5(iii) confirms that $\left\{g^k\right\}$ is bounded away from 0, which tells us that such an index set J does not exist.

To verify now assertion (ii), observe that assuming $g^k \xrightarrow{J} 0$ implies by the first inequality in (4.6) that $d^k \xrightarrow{J} 0$. Conversely, suppose that $d^k \xrightarrow{J} 0$ and deduce from Proposition 4.5(i) that the set $\mathcal{N}$ is infinite. Then, it follows from Proposition 4.5(ii) that $\varepsilon_k \downarrow 0$ and $r_k \downarrow 0$ as $k \to \infty$. Using the second inequality in (4.6), we arrive at $g^k \xrightarrow{J} 0$ and thus complete the proof of the proposition. □

Finally in this section, we deduce from the obtained results the following desired property of the direction sequence $\left\{d^k\right\}$ in Algorithm 1.

Corollary 4.7 *The sequence $\left\{d^k\right\}$ in Algorithm 1 is gradient associated with $\left\{x^k\right\}$.*

Proof It follows from Proposition 4.6 that the convergence $d^k \xrightarrow{J} 0$ yields $g^k \xrightarrow{J} 0$ and $\varepsilon_k \downarrow 0$. Thus, we get $\nabla f(x^k) \xrightarrow{J} 0$ by taking into account $\left\|g^k - \nabla f(x^k)\right\| \le \varepsilon_k$ from (4.1). This shows therefore that the sequence $\left\{d^k\right\}$ is gradient associated with $\left\{x^k\right\}$. □

5 Inexact Reduced Gradient Methods with Stepsize Selections

In this section, we develop novel IRG methods with the following selections of stepsize rules: *backtracking stepsize*, *constant stepsize*, and *diminishing stepsize*. Address first an IRG method with the *backtracking linesearch*. Choose a linesearch scalar $\beta \in (0, 1)$, a reduction factor $\gamma \in (0, 1)$, and an artificial stepsize of null iterations $\tau \in (0, 1)$. Consider the *Master Algorithm* 1 with the stepsize sequence $\{t_k\}$ in Step 3 calculated as follows. If $d^k = 0$, then put $t_k := \tau$. Otherwise, we set

$$t_k := \max\left\{t \;\middle|\; f(x^k + td^k) \le f(x^k) - \beta t\left\|d^k\right\|^2,\ t = 1,\ \gamma,\ \gamma^2, \dots\right\}. \tag{5.1}$$

The next proposition shows that the stepsize sequence $\{t_k\}$ is well defined.

Proposition 5.1 *If $d^k \neq 0$, then there exists a positive number $\bar{t}$ such that*

$$f(x^k + td^k) \le f(x^k) - \beta t \left\| d^k \right\|^2 \textit{ for all } t \in [0, \bar{t}],$$

which ensures the existence of t_k in (5.1).

Proof Suppose that $d^k \neq 0$ and get by Proposition 4.3 that $\langle \nabla f(x^k), d^k \rangle \le - \left\| d^k \right\|^2$. By the differentiability of f at x^k, for each $t > 0$ sufficiently small we have

$$\begin{aligned} f(x^k + td^k) - f(x^k) &= t \left\langle \nabla f(x^k), d^k \right\rangle + o(t) \le -t \left\| d^k \right\|^2 + o(t) \\ &= -\beta t \left\| d^k \right\|^2 + t \left((\beta - 1) \left\| d^k \right\|^2 + \frac{o(t)}{t} \right). \end{aligned}$$

Since $o(t)/t \to 0$ as $t \downarrow 0$ and since $(\beta - 1) \left\| d^k \right\|^2 < 0$, there exists $\bar{t} > 0$ such that

$$f(x^k + td^k) \le f(x^k) - \beta t \left\| d^k \right\|^2 \text{ for all } t \in (0, \bar{t}].$$

Therefore, the selection of t_k in (5.1) is well defined. □

Now, we are ready to establish a major result about the *stationarity of accumulation points* of the iterative sequence generated by Algorithm 1 with the backtracking line search.

Theorem 5.2 *Let $\{x^k\}$ be the sequence of iterations generated by Algorithm 1 with the sequence of stepsizes $\{t_k\}$ being chosen via the backtracking linesearch as in (5.1). Assume in addition to the inexact gradient condition (4.1) that $\left\| g^k - \nabla f(x^k) \right\| \le \rho_k$ with $\rho_k \downarrow 0$ as $k \to \infty$. Suppose furthermore that $\inf f(x^k) > -\infty$. Then, the following assertions hold:*

- **(i)** *$\varepsilon_k \downarrow 0$ and $r_k \downarrow 0$ as $k \to \infty$.*
- **(ii)** *Every accumulation point of $\{x^k\}$ is a stationary point of f.*
- **(iii)** *If the sequence $\{x^k\}$ is bounded, then the set of accumulation points of $\{x^k\}$ is nonempty, compact, and connected.*
- **(iv)** *If $\{x^k\}$ has an isolated accumulation point, then the entire sequence $\{x^k\}$ converges to this point.*

Proof From the choice of $\{t_k\}$ and Step 4 in Algorithm 1, for every $k \in \mathbb{N}$ we have

$$\beta t_k \left\| d^k \right\|^2 \le f(x^k) - f(x^{k+1}). \tag{5.2}$$

Since $\inf f(x^k) > -\infty$, summing up on both sides of (5.2) over $k = 1, 2, \ldots$ and using the relation $x^{k+1} = x^k + t_k d^k$, we get that

$$\sum_{k=1}^{\infty} t_k \left\| d^k \right\|^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \left\| x^{k+1} - x^k \right\| \cdot \left\| d^k \right\| < \infty. \tag{5.3}$$

To verify assertion (i), recall by Proposition 4.5 (ii) that the convergence $\varepsilon_k \downarrow 0$ is equivalent to $r_k \downarrow 0$ and to the set of null iterations $\mathcal{N}$ being infinite. Assume on the contrary that $\mathcal{N}$ is finite. By Proposition 4.5(iii) with $N = \max \mathcal{N} + 1$, we have

$$\left\| g^k \right\| > r_N + \varepsilon_N \text{ and } \left\| d^k \right\| > r_N \text{ for all } k \geq N. \tag{5.4}$$

Then, (5.3) gives us $\sum_{k=1}^{\infty} t_k < \infty$ and thus $t_k \downarrow 0$ as $k \to \infty$. Choosing a larger number N if necessary, we get that $t_k < 1$ for all $k \geq N$. For such k, it follows from the exit condition of the algorithm that

$$-\gamma^{-1}\beta t_k \left\| d^k \right\|^2 < f(x^k + \gamma^{-1}t_k d^k) - f(x^k). \tag{5.5}$$

By the classical mean value theorem, there exists some $\tilde{x}^k \in [x^k, x^k + \gamma^{-1}t_k d^k]$ such that

$$f(x^k + \gamma^{-1}t_k d^k) - f(x^k) = \gamma^{-1}t_k \left\langle d^k, \nabla f(\tilde{x}^k) \right\rangle.$$

The latter equality together with (5.5) tells us that

$$\left\langle -d^k, \nabla f(\tilde{x}^k) \right\rangle \leq \beta \left\| d^k \right\|^2 \text{ for all } k \geq N. \tag{5.6}$$

Using (5.3) and (5.4), we have $\sum_{k=1}^{\infty} \left\| x^{k+1} - x^k \right\| < \infty$, and thus, $\left\{x^k\right\}$ converges to some $\bar{x} \in \mathbb{R}^n$. The continuity of ∇f ensures that $\nabla f(x^k) \to \nabla f(\bar{x})$. Then, employing $\left\| g^k - \nabla f(x^k) \right\| \leq \rho_k \to 0$ yields $g^k \to \nabla f(\bar{x})$ as $k \to \infty$. It follows from Step 2 that

$$-d^k = \frac{\|g^k\| - \varepsilon_k}{\|g^k\|} g^k = \frac{\|g^k\| - \varepsilon_N}{\|g^k\|} g^k \text{ for all } k \geq N.$$

Letting $k \to \infty$ leads us to the equalities

$$-d^k \to \bar{g} := \frac{\|\nabla f(\bar{x})\| - \varepsilon_N}{\|\nabla f(\bar{x})\|} \nabla f(\bar{x}) = \mathrm{Proj}\big(0, \mathbb{B}(\nabla f(\bar{x}), \varepsilon_N)\big). \tag{5.7}$$

Using $t_k \downarrow 0$, we get that $\tilde{x}^k \to \bar{x}$, and thus $\nabla f(\tilde{x}^k) \to \nabla f(\bar{x})$ as $k \to \infty$. Combining the latter with (5.6), (5.7), and the projection characterization verifies the estimates

$$\|\bar{g}\|^2 \leq \langle \bar{g}, \nabla f(\bar{x}) \rangle \leq \beta \|\bar{g}\|^2. \tag{5.8}$$

This tells us that $\bar{g} = 0$, which contradicts the condition $\|\bar{g}\| \geq r_N$ by (5.4). Therefore, we arrive at $\varepsilon_k \downarrow 0$ and $r_k \downarrow 0$ as $k \to \infty$, which completes the proof of assertion (i).

To justify assertions (ii)–(iv), recall from Corollary 4.7 that $\left\{d^k\right\}$ is *gradient associated* with $\left\{x^k\right\}$. Since $\varepsilon_k \downarrow 0$, we deduce from Proposition 4.6 that 0 is an accumulation point of $\{d^k\}$. Combining these facts with (5.3) and $t_k \le 1$ whenever $k \in \mathbb{N}$ ensures that all the assumptions of Theorem 3.4 are satisfied. Therefore, we verify assertions (ii)–(iv) and finish the proof of the theorem. □

Next we consider problem (1.1) with the objective function f satisfying the *L-descent condition* for some $L > 0$. The following result establishes convergence properties of IRG method, which uses either *diminishing* or *constant stepsizes*.

Theorem 5.3 *Let $\{x^k\}$ be the sequence generated by Algorithm 1, where*

(a) *f satisfies the L-descent condition;*
(b) *either $\{t_k\}$ is diminishing, i.e.,*

$$t_k \downarrow 0 \ \ as \ \ k \to \infty \ \ and \ \ \sum_{k=1}^{\infty} t_k = \infty, \tag{5.9}$$

or there exist $\delta, \delta' > 0$ such that $\delta' \le \dfrac{2-\delta}{L}$ and

$$t_k \in \left[\delta', \frac{2-\delta}{L}\right] \ \ for\ all \ \ k \in \mathbb{N}. \tag{5.10}$$

Assume that $\inf f(x^k) > -\infty$. *Then, all the conclusions of Theorem 5.2 hold. Moreover, if $\{t_k\}$ is chosen as (5.10), then $\nabla f(x^k) \to 0$ as $k \to \infty$.*

Proof We know from Remark 4.7 that the direction sequence $\{d^k\}$ is *gradient associated* with $\{x^k\}$. Furthermore, Proposition 4.3 tells us that $\{d^k\}$ satisfies the *sufficient descent* condition (3.7) with the constant $\kappa = 1$. Note that if $\{t_k\}$ is chosen as either (5.9) or (5.10), then we always get that

$$\sum_{k=1}^{\infty} t_k = \infty \quad \text{and} \quad t_k \le \frac{2-\delta}{L} \quad \text{for sufficiently large} \ \ k \in \mathbb{N}.$$

Combining these facts with the imposed L-descent condition on f yields the fulfillment of assumptions (a), (b), (c) in Corollary 3.5. Therefore, conclusions (ii)–(iv) of Theorem 5.2 hold. The proof of Corollary 3.5 also ensures that 0 is an accumulation point of $\{d^k\}$. Thus, it follows from Proposition 4.6 that $\varepsilon_k \downarrow 0$. Using Proposition 4.5(ii), we have $r_k \downarrow 0$, which verifies conclusion (i) of Theorem 5.2. If $\{t_k\}$ is chosen as (5.10), its boundedness away from 0 is guaranteed, and so Corollary 3.5 yields $\nabla f(x^k) \to 0$ as $k \to \infty$ and thus completes the proof of the theorem. □

The final part of our convergence analysis of the proposed IRG methods applies the *KL property* to establishing the *global* convergence of the *entire sequence* of iterations to a *stationary point* with deriving *convergence rates*. We start with the following simple albeit useful lemma.

Lemma 5.4 *Let $\{x^k\}$ be the sequence generated by Algorithm 1 with $\theta < \mu$. Assume that $\varepsilon_k \downarrow 0$ and $r_k \downarrow 0$ as $k \to \infty$. Then, there exists some $N \in \mathbb{N}$ such that*

$$\left\|\nabla f(x^k)\right\| \leq 3\left\|d^k\right\| \quad \textit{for all} \;\; k \notin \mathcal{N},\;\; k \geq N, \tag{5.11}$$

where the set $\mathcal{N}$ is taken from Definition 4.4.

Proof It follows directly from the assumptions of the lemma that there exists a natural number N such that $\varepsilon_k \leq r_k$ for all $k \geq N$. By Step 2 of the algorithm, for any $k \geq N$ with $k \notin \mathcal{N}$ we have $\|g^k\| > r_k + \varepsilon_k$ with the direction d^k calculated in (4.2). Thus, for such k we get the estimates

$$\left\|d^k\right\| = \left\|g^k\right\| - \varepsilon_k > r_k + \varepsilon_k - \varepsilon_k = r_k \geq \varepsilon_k. \tag{5.12}$$

It follows from (4.1) in Step 1 and from (5.12) that

$$\left\|\nabla f(x^k)\right\| \leq \left\|g^k\right\| + \varepsilon_k = \left\|d^k\right\| + 2\varepsilon_k \leq 3\left\|d^k\right\|,$$

which verifies the conclusion of the lemma. □

The following two theorems provide conditions ensuring the global convergence of iterative sequences generated by Algorithm 1 with different stepsize selections to a stationary point of f. The first theorem concerns the IRG methods with the *backtracking* stepsize.

Theorem 5.5 *Let $\{x^k\}$ be the iterative sequence generated by Algorithm 1 with the backtracking linesearch under the condition $\theta < \mu$. Suppose in addition to the inexact gradient condition (4.1) that $\|g^k - \nabla f(x^k)\| \leq \rho_k$ with $\rho_k \downarrow 0$ as $k \to \infty$. Assume furthermore that $\{x^k\}$ has an accumulation point $\bar{x}$, and f satisfies the KL property at $\bar{x}$. Then, $\bar{x}$ is a stationary point of f, and $x^k \to \bar{x}$ as $k \to \infty$.*

Proof Since $\bar{x}$ is an accumulation point of $\{x^k\}$, we can find some infinite set $J \subset \mathbb{N}$ such that $x^k \overset{J}{\to} \bar{x}$. It follows from the choice of $\{t_k\}$ in (5.1) that $\{f(x^k)\}$ is nonincreasing, which implies that

$$\inf_{k\in\mathbb{N}} f(x^k) = \inf_{k\in J} f(x^k) = f(\bar{x}) > -\infty.$$

Therefore, the results of Theorem 5.2 tell us that $\bar{x}$ is a stationary point of f and that $\varepsilon_k \downarrow 0$, $r_k \downarrow 0$ as $k \to \infty$. We employ Proposition 2.4 to verify that $x^k \to \bar{x}$ along the entire sequence of iterations. Indeed, the imposed assumptions and the convergence $\varepsilon_k \downarrow 0$, $r_k \downarrow 0$ as $k \to \infty$ guarantee that all the requirements of Lemma 5.4 are satisfied. Pick $N \in \mathbb{N}$ such that (5.11) holds. The choice of $\{t_k\}$ in (5.1) ensures the lower estimate

$$f(x^k) - f(x^{k+1}) \geq \beta t_k \left\|d^k\right\|^2 \quad \text{for all} \;\; k \in \mathbb{N}. \tag{5.13}$$

Combining this with (5.11) and the relation $x^{k+1} = x^k + t_k d^k$ yields

$$f(x^k) - f(x^{k+1}) \geq \frac{\beta}{3} \left\| \nabla f(x^k) \right\| \cdot \left\| x^{k+1} - x^k \right\| \quad \text{for all } \ k \notin \mathcal{N}, \ k \geq N. \tag{5.14}$$

Observe that when $k \in \mathcal{N}$, both sides of (5.14) reduce to zero, and so (5.14) is satisfied. Therefore, assumption (H1) in Proposition 2.4 holds. Moreover, for $k \geq N$ the conditions $f(x^{k+1}) = f(x^k)$ and (5.13) imply that $d^k = 0$, and hence, $x^{k+1} = x^k$. Thus, assumption (H2) in Proposition 2.4 is satisfied as well. Applying the latter proposition, we arrive at $x^k \to \bar{x}$ as $k \to \infty$ and complete the proof. □

The second theorem of the above type addresses the IRG methods with *diminishing* and *constant* selections of the stepsize sequence $\{t_k\}$.

Theorem 5.6 *Let the objective function f satisfy the L-descent condition (2.1) for some $L > 0$, and let $\left\{x^k\right\}$ be the sequence generated by Algorithm 1 with $\theta < \mu$, and either diminishing (5.9) or constant stepsizes (5.10). Assume in addition that $\bar{x}$ is an accumulation point of the sequence $\left\{x^k\right\}$ and that f satisfies the KL property at $\bar{x}$. Then, $\bar{x}$ is a stationary point of f, and we have the convergence $x^k \to \bar{x}$ as $k \to \infty$.*

Proof Observe first that the assumptions imposed here yield those in Theorem 5.3 and Corollary 3.5 but $\inf f(x^k) > -\infty$. Similarly to the proof of Corollary 3.5, we can show that $\left\{f(x^k)\right\}_{k \geq K}$ is nonincreasing for some $K \in \mathbb{N}$. Since $\bar{x}$ is an accumulation point of $\left\{x^k\right\}$, similarly to the proof of Theorem 5.5, we deduce that $\inf f(x^k) = f(\bar{x}) > -\infty$, which verifies the remaining assumption. Therefore, $\bar{x}$ is a stationary point of f and $\varepsilon_k \downarrow 0$, $r_k \downarrow 0$ as $k \to \infty$. The latter convergence together with the imposed assumptions guarantees the fulfillment of all the conditions of Lemma 5.4. Let $N \in \mathbb{N}$ be such that (5.11) holds. Let $\delta > 0$ be the constant given in (5.10). From the proof of Corollary 3.5, we find some $N_1 \geq N$ such that

$$f(x^k) - f(x^{k+1}) \geq \frac{\delta}{2} t_k \left\| d^k \right\|^2 \quad \text{for all } \ k \geq N_1. \tag{5.15}$$

The relation $x^{k+1} = x^k + t_k d^k$ and (5.11), (5.15) tell us that

$$f(x^k) - f(x^{k+1}) \geq \frac{\delta}{6} \left\| \nabla f(x^k) \right\| \left\| x^{k+1} - x^k \right\| \quad \text{whenever } \ k \notin \mathcal{N}, \ k \geq N_1. \tag{5.16}$$

Similarly to the proof of Theorem 5.5, we get $x^k \to \bar{x}$ as $k \to \infty$ and thus complete the proof. □

We will see below that the boundedness of stepsizes away from 0 plays a crucial role in establishing the *rate of convergence* of the IRG methods. This property automatically holds for constant stepsizes while may fail for diminishing ones. The next proposition shows that the property is satisfied for the backtracking stepsize selection provided that the gradient of the objective function is *locally Lipschitzian* around accumulation points of iterative sequence. Observe that this property is strictly weaker than the (global) Lipschitz continuous of ∇f. Indeed, $\mathcal{C}^2$-smooth functions have locally Lipschitzian gradients but do not need to have a globally Lipschitzian one as, e.g., for $f(x) := x^4$.

Proposition 5.7 *Let $\{x^k\}$ be a sequence generated by Algorithm 1 with the backtracking stepsize. Suppose in addition to the inexact gradient condition (4.1) that $\|g^k - \nabla f(x^k)\| \leq \rho_k$ with $\rho_k \downarrow 0$ as $k \to \infty$. Assume moreover that there exists an infinite set $J \subset \mathbb{N}$ such that $\{x^k\}_{k\in J}$ converges to some $\bar{x} \in \mathbb{R}^n$ and that ∇f is locally Lipschitzian around $\bar{x}$. Then, the stepsize sequence $\{t_k\}_{k\in J}$ is bounded away from zero.*

Proof Assume on the contrary that $\{t_k\}_{k\in J}$ is not bounded away from zero. Then, we find an infinite set $\bar{J} \subset J$ such that $t_k \xrightarrow{\bar{J}} 0$. Let $\tau \in (0, 1)$ be an artificial stepsize of null iterations. Since $t_k \xrightarrow{\bar{J}} 0$, there exists a number $N \in \mathbb{N}$ such that

$$t_k < \tau < 1 \ \text{ for all } \ k \geq N, \ k \in \bar{J}. \tag{5.17}$$

This means that $k \notin \mathcal{N}$ whenever $k \geq N, \ k \in \bar{J}$. By Proposition 4.5(i), we have $d^k \neq 0$ for all $k \geq N, \ k \in \bar{J}$. Then, condition (4.2) in Step 2 leads us to

$$\left\|d^k\right\| = \left\|g^k\right\| - \varepsilon_k \leq \left\|g^k\right\| \ \text{ for all } \ k \geq N, \ k \in \bar{J}. \tag{5.18}$$

Since $x^k \overset{\bar{J}}{\to} \bar{x}$, the continuity of ∇f and the estimate $\|\nabla f(x^k) - g^k\| \leq \rho_k \to 0$ yield that $g^k \overset{\bar{J}}{\to} \nabla f(\bar{x})$. Using (5.18), we get that the sequence $\{d^k\}_{k\in\bar{J}}$ is bounded, and thus,

$$x^k + \gamma^{-1}t_k d^k \to \bar{x} \ \text{ as } \ k \overset{\bar{J}}{\to} \infty. \tag{5.19}$$

Since ∇f is locally Lipschitzian around $\bar{x}$, there exists a positive number δ such that ∇f is Lipschitz continuous on $\mathbb{B}(\bar{x}, \delta)$ with some modulus $L > 0$. By (5.19) and $x^k \overset{\bar{J}}{\to} \bar{x}$, we find $N_1 \geq N$ with $x^k, x^k + \gamma^{-1}t_k d^k \in \mathbb{B}(\bar{x}, \delta)$ for all $k \geq N_1, \ k \in \bar{J}$. The Lipschitz continuity of ∇f on $\mathbb{B}(\bar{x}, \delta)$ with modulus L yields by [10, Proposition A.24] the L-descent condition, i.e.,

$$f(x) - f(y) \leq \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 \ \text{ for all } \ x, y \in \mathbb{B}(\bar{x}, \delta). \tag{5.20}$$

Fixing $k \in \bar{J}, \ k \geq N_1$, we deduce from the above that $d^k \neq 0$, and $t_k < 1$. The exit condition for the backtracking linesearch implies that

$$-\gamma^{-1}\beta t_k \left\|d^k\right\|^2 < f(x^k + \gamma^{-1}t_k d^k) - f(x^k). \tag{5.21}$$

Applying (5.20) for $x = x^k + \gamma^{-1} t_k d^k$ and $y = x^k$, we have that

$$f(x^k + \gamma^{-1} t_k d^k) - f(x^k) \leq \gamma^{-1} t_k \left\langle \nabla f(x^k), d^k \right\rangle + \frac{L\gamma^{-2} t_k^2}{2} \left\| d^k \right\|^2.$$

Combining this with (5.21) leads us to

$$-\gamma^{-1} \beta t_k \left\| d^k \right\|^2 < \gamma^{-1} t_k \left\langle \nabla f(x^k), d^k \right\rangle + \frac{L\gamma^{-2} t_k^2}{2} \left\| d^k \right\|^2,$$

or equivalently to the inequality

$$0 < \gamma^{-1} \beta t_k \left\| d^k \right\|^2 + \gamma^{-1} t_k \left\langle \nabla f(x^k), d^k \right\rangle + \frac{L\gamma^{-2} t_k^2}{2} \left\| d^k \right\|^2. \tag{5.22}$$

Proposition 4.3 and $d^k \neq 0$ tell us that $0 < \left\| d^k \right\|^2 \leq \left\langle \nabla f(x^k), -d^k \right\rangle$. Then, we deduce from (5.22) the fulfillment of the estimate

$$0 < \gamma^{-1} \beta t_k \left\langle \nabla f(x^k), -d^k \right\rangle + \gamma^{-1} t_k \left\langle \nabla f(x^k), d^k \right\rangle + \frac{L\gamma^{-2} t_k^2}{2} \left\langle \nabla f(x^k), -d^k \right\rangle.$$

Dividing both sides above by $\gamma^{-1} t_k \left\langle \nabla f(x^k), -d^k \right\rangle > 0$, we get $0 < \beta - 1 + \frac{L\gamma^{-1} t_k}{2}$. Letting $k \xrightarrow{J} \infty$ yields $\beta \geq 1$, which contradicts the choice of $\beta \in (0, 1)$. Thus, we verify that the sequence $\{t_k\}_{k \in J}$ is bounded away from zero, which completes the proof of the proposition. □

The last two theorems establish sufficient conditions ensuring the *convergence rates* in Algorithm 1 under different stepsize selections. Having the sequence of iterations $\left\{x^k\right\}$ generated by this algorithm, we obtain first from Proposition 4.5(iii) that if $\mathbb{N} \setminus \mathcal{N}$ is finite, then $\left\{x^k\right\}$ stops after a finite number of iterations. Thus, we consider the case where the set $\mathbb{N} \setminus \mathcal{N}$ is infinite and can be numerated as $\{j_1, j_2, \ldots\}$. Construct the sequence $\left\{z^k\right\}$ by

$$z^k := x^{j_k} \quad \text{for all} \quad k \in \mathbb{N}. \tag{5.23}$$

We have $j_{k+1} \geq j_k + 1$ whenever $k \in \mathbb{N}$. If the equality holds therein, then $z^{k+1} = x^{j_k+1}$. Otherwise, by taking into account that the indices $j_k + 1, \ldots, j_{k+1} - 1$ correspond to null iterations, we get that

$$x^{j_k+1} = x^{j_k+2} = \ldots = x^{j_{k+1}-1} = x^{j_{k+1}} = z^{k+1}. \tag{5.24}$$

Therefore, it follows from $j_k \notin \mathcal{N}$ that

$$z^{k+1} = x^{j_k+1} \neq x^{j_k} = z^k \quad \text{for all} \quad k \in \mathbb{N}. \tag{5.25}$$

Furthermore, Proposition 4.5(iv) tells us that $\bar{x}$ is an accumulation point of $\left\{z^k\right\}$ if and only if $\bar{x}$ is also an accumulation point of $\left\{x^k\right\}$.

The first theorem about the convergence rates concerns Algorithm 1 with the *backtracking stepsize*.

Theorem 5.8 *Consider Algorithm 1 with the backtracking stepsize selections under the condition $\theta < \mu$. Let $\left\{x^k\right\}$ be the iterative sequence generated by this algorithm. Suppose in addition to the inexact gradient condition (4.1) that $\left\|g^k - \nabla f(x^k)\right\| \le \rho_k$ with $\rho_k \downarrow 0$ as $k \to \infty$. Assume further that $\left\{x^k\right\}$ has an accumulation point $\bar{x}$, that f satisfies the KL property at $\bar{x}$ with $\psi(t) = Mt^q$ for some $M > 0$ and $q \in (0, 1)$, and that ∇f is locally Lipschitzian around $\bar{x}$. The following convergence rates are guaranteed for the sequence $\left\{z^k\right\}$ defined in (5.23):*

(i) *If $q \in (0, 1/2]$, then the sequence $\left\{z^k\right\}$ converges linearly to $\bar{x}$.*
(ii) *If $q \in (1/2, 1)$, then there exists a positive constant ϱ such that*

$$\left\|z^k - \bar{x}\right\| \le \varrho k^{-\frac{1-q}{2q-1}} \text{ for all large } k \in \mathbb{N}.$$

Proof The imposed assumptions yield the fulfillment of those in Theorem 5.5, and so lead us to the convergence $x^k \to \bar{x}$ as $k \to \infty$. Then, the local Lipschitz continuity of ∇f around $\bar{x}$ and Proposition 5.7 ensure that the sequence $\{t_k\}$ is bounded away from zero.

To deduce now the claimed convergence rates in (i)–(iii) from Theorem 2.5, define $\tau_k := t_{j_k}$ for all $k \in \mathbb{N}$. Then, $\{\tau_k\}$ is also bounded away from zero as a subsequence of $\{t_k\}$. Furthermore, using (5.24) and the linesearch conditions, we have

$$\begin{aligned} f(z^k) - f(z^{k+1}) &= f(x^{j_k}) - f(x^{j_k+1}) \ge \beta t_{j_k} \left\|d^{j_k}\right\|^2 \\ &= \frac{\beta}{t_{j_k}} \left\|x^{j_k+1} - x^{j_k}\right\|^2 = \frac{\beta}{\tau_k} \left\|z^{k+1} - z^k\right\|^2 \end{aligned} \tag{5.26}$$

for all $k \in \mathbb{N}$. Note that all the assumptions of Theorem 5.2 are satisfied, and so Lemma 5.4 holds. Pick any $N \in \mathbb{N}$ from (5.11) and fix $k \ge N$. Then, using (5.24) and (5.11) with taking into account that $j_k \notin \mathcal{N}$ for $j_k \ge k$ leads us to

$$\left\|\nabla f(z^k)\right\| = \left\|\nabla f(x^{j_k})\right\| \le \frac{3}{t_{j_k}} \left\|x^{j_k+1} - x^{j_k}\right\| = \frac{3}{\tau_k} \left\|z^{k+1} - z^k\right\|.$$

Apply finally Theorem 2.5 to $\left\{z^k\right\}$ and $\{\tau_k\}$ while remembering that $z^{k+1} \ne z^k$ for all $k \in \mathbb{N}$ from (5.25). This verifies the convergence rates (i)–(iii) claimed in the theorem. □

The next theorem on the convergence rates addresses Algorithm 1 with the *constant stepsizes*.

Theorem 5.9 *Let f satisfy the L-descent condition for some $L > 0$, and let $\left\{x^k\right\}$ be the iterative sequence generated by Algorithm 1 with the constant stepsizes (5.10)* *under the condition* $\theta < \mu$. *Suppose that* $\{x^k\}$ *has an accumulation point* $\bar{x}$ *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 for the iterative sequence* $\{z^k\}$ *defined in* (5.23)*:*

(i) *If* $q \in (0, 1/2]$, *then the sequence* $\{z^k\}$ *converges linearly to* $\bar{x}$.
(ii) *If* $q \in (1/2, 1)$, *then there exists a positive constant* ϱ *such that*

$$\left\|z^k - \bar{x}\right\| \le \varrho k^{-\frac{1-q}{2q-1}} \text{ for all large } k \in \mathbb{N}.$$

Proof Note that our assumptions yield the fulfillment of those in Theorem 5.6, and thus, we have that $x^k \to \bar{x}$ as $k \to \infty$. Defining $\tau_k := t_{j_k}$ for all $k \in \mathbb{N}$ ensures that the stepsize sequence $\{\tau_k\}$ is bounded away from zero. Note that all the assumptions in Corollary 3.5 hold, and let $\delta > 0$ be the constant taken from in (5.10). By the L-descent property of f and the constant stepsize selection, we find by arguing similarly to the proof of Corollary 3.5 a number $N \in \mathbb{N}$ such that

$$f(x^k) - f(x^{k+1}) \ge \frac{\delta}{2} t_k \left\|d^k\right\|^2 \quad \text{whenever} \quad k \ge N. \tag{5.27}$$

Since $j_k \ge k \ge N$ for such k, it follows that

$$\begin{aligned} f(z^k) - f(z^{k+1}) = f(x^{j_k}) - f(x^{j_k+1}) &\ge \frac{\delta}{2} t_{j_k} \left\|d^{j_k}\right\|^2 \\ &= \frac{\delta}{2t_{j_k}} \left\|x^{j_k+1} - x^{j_k}\right\|^2 = \frac{\delta}{2\tau_k} \left\|z^{k+1} - z^k\right\|^2. \end{aligned}$$

Note that all the assumptions of Theorem 5.3 are satisfied. Using this result together with Lemma 5.4 and then arguing as in the proof of Theorem 5.8, we complete the proof of this theorem. □

6 Applications and Numerical Experiments

In this section, we present efficient implementations of the developed IRG methods to solving particular classes of optimization problems that appear in practical modeling. We conduct numerical experiments and compare the results of computations by using our algorithms with those obtained by applying some other well-known methods. This section is split into two subsections addressing different classes of problems with the usage of different algorithms.

6.1 Comparison with Classical Inexact Proximal Point Method

This subsection addresses the *Least Absolute Deviations* (LAD) *Curve-Fitting problem* which is formulated as follows:

$$\text{minimize} \quad g(x) := \|Ax - b\|_1 \quad \text{over} \quad x \in \mathbb{R}^n, \tag{6.1}$$

where A is an $m \times n$ matrix, b is a vector in $\mathbb{R}^m$, and $\|u\|_1 := \sum_{k=1}^m |u_k|$ for any $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$. Problem (6.1) exhibits robustness in outliers resistance and appears in many applied areas; see, e.g., [11] for more discussions. Observe that (6.1) is a problem of *nonsmooth convex optimization*, but we can reduce it to a smooth problem by using a regularization procedure. In this way, we solve (6.1) by using our *IRG method with constant stepsize* and compare our approach with the usage of the *inexact proximal point method* (IPPM) proposed by Rockafellar in [46].

To proceed, recall that the *Moreau envelope* and the *proximal mapping* of g are defined by

$$e_g(x) := \inf_{y \in \mathbb{R}^n} \varphi_x(y) \text{ and } \operatorname{Prox}_g(x) := \operatorname*{argmin}_{y \in \mathbb{R}^n} \varphi_x(y), \quad x \in \mathbb{R}^n, \tag{6.2}$$

where the minimization mapping $\varphi_x : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\varphi_x(y) := g(y) + \frac{1}{2}\|y - x\|^2, \quad y \in \mathbb{R}^n. \tag{6.3}$$

Since g is convex, it follows from [7, Propositions 12.28 and 12.30] that e_g is $\mathcal{C}^1$-smooth and that its gradient is Lipschitz continuous with constant 1 being represented by

$$\nabla e_g(x) = x - \operatorname{Prox}_g(x) \text{ for all } x \in \mathbb{R}^n. \tag{6.4}$$

Moreover, the set of minimizers of g coincides with the set of zeros of the gradient mapping ∇e_g.

This tells us that problem (6.1) can be equivalently transformed into the problem of finding stationary points of the smooth function $f := e_g$. Therefore, it is possible to solve (6.1) by using Algorithm 1 with constant stepsize, where an inexact gradient g^k of $\nabla f(x^k)$ in Step 1 satisfying the condition (4.1) can be chosen from the conditions

$$g^k := x^k - p^k \text{ with } \left\| p^k - \operatorname{Prox}_g(x^k) \right\| \le \varepsilon_k. \tag{6.5}$$

Meanwhile, the iterative procedure of IPPM in [46, page 878] for solving (6.1) is given by

$$x^{k+1} = p^k \text{ with } \left\| p^k - \operatorname{Prox}_g(x^k) \right\| \le \delta_k, \text{ where } \sum_{k=1}^{\infty} \delta_k < \infty. \tag{6.6}$$

Since the function φ_{x^k} in (6.3) is strongly convex with constant 1 [38, Definition 2.1.3], the error bound for the distance between the inexact proximal point p^k and the exact one $\operatorname{Prox}_g(x^k)$ in (6.5) and (6.6) is satisfied if

$$\varphi_{x^k}(p^k) \le \inf \varphi_{x^k} + \omega_k, \tag{6.7}$$

where $\omega_k := \frac{\varepsilon_k^2}{2}$ for (6.5) and $\omega_k := \frac{\delta_k^2}{2}$ for (6.6) by using [38, Theorem 2.1.8]. In this numerical experiment, we run the *Fast Iterative Shrinkage-Thresholding Algorithm* (FISTA) of Beck and Teboulle [9] for the dual function of φ_{x^k} until the duality gap is below ω_k, which therefore ensures (6.7).

The initial points are chosen as $x^1 := 0_{\mathbb{R}^n}$ for both algorithms, while the detailed settings of each algorithm are given as follows:

- IRG: $\varepsilon_1 = 10, \theta = \mu = 0.5$. Two selections of the initial radius r_1 are 20 and 5, which correspond to versions IRG-20 and IRG-5, respectively. To simplify the iterative sequence of Algorithm 1 when $\|g^k\| \leq r_k + \varepsilon_k$, we put $x^{k+1} := p^k$, which corresponds to the choice of stepsize $t_k = \frac{\|g^k\|}{\|g^k\| - \varepsilon_k}$.
- IPPM: $\omega_k = \frac{1}{k^p}$ for all $k \in \mathbb{N}$, where $p = 4$ or $p = 2.1$. These selections together with the definition of ω_k in (6.7) ensure that $\sum_{k=1}^{\infty} \delta_k < \infty$ as required for IPPM in (6.6). We also use the labels IPPM-4 and IPPM-2.1 for these versions of IPPM, respectively.

In this numerical experiment, we let IPPM-2.1 run for 200 iterations and record the function value obtained by this method. Then, other methods run until their function values are lower than the recorded one of IPPM-2.1. We stop the methods when the time reaches the limit of 4000 seconds. The data A, b are generated randomly with i.i.d. (identically and independent distributed) standard Gaussian entries. To avoid algorithms from reaching the solution promptly, we consider only the cases where $m \leq n$ in (6.1).

The numerical experiment is conducted on a computer with 10th Gen Intel(R) Core(TM) i5-10400 (6-Core 12 M Cache, 2.9–4.3 GHz) and 16GB RAM memory. The codes are written in MATLAB R2021a. Detailed information for the results is presented in Table 1, where 'Test #,' '*iter*,' '*fval*,' '*time*' mean test number, the number of iterations, value of the objective function at the last iteration, and the computational time, respectively. The bold text indicates the running time of the fastest algorithms in each test. The errors ω_k in the inexact proximal point calculations (6.7) and the function values obtained by the algorithms over the duration of time are also graphically illustrated in Figs. 3 and 4.

It can be seen from Table 1 that IRG-5 has the best performance in this numerical experiment. IRG-20 is the second fastest algorithm in Tests 1, 3, 5, while it is slightly slower than IPPM-4 in Tests 2, 4. In Test 5 with the largest dimensions $m = n = 1200$, IRG-5 is around 4 times faster than IPPM-2.1, while IPPM-4 even cannot reach the value obtained by IPPM-2.1 within the time limit. In this test, IRG-20 is also around 2.5 times faster than IPPM-2.1.

The graphs in Figs. 3 and 4 show that the errors (in inexact proximal point calculations) of IRG are automatically adjusted to be suitable for different problems:

- In Tests 1, 3, 5 with $m = n$, it can be seen from Fig. 4 that IPPM-2.1 is faster than IPPM-4, which means that the use of larger errors is preferred in this case. Then, Fig. 3 shows that the errors used in IRG stagnate at most of the iterations. As a result, the IRG methods use the errors larger than that of the IPPM methods and thus achieve better performances.

Table 1 Results for LAD curve-fitting problem

Test #	m	n	IPPM-2.1			IPPM-4			IRG-5			IRG-20		
			Iter	fval	Time	Iter	fval	Time	Iter	fval	Time	Iter	fval	Time
1	300	300	200	3.44	11.79	200	3.44	100.63	209	3.44	**5.51**	210	3.43	*7.91*
2	300	600	200	0.00	1.56	17	0.00	*0.38*	16	0.00	**0.33**	19	0.00	0.44
3	600	600	200	0.44	140.35	201	0.44	863.39	210	0.44	**78.09**	211	0.44	*119.78*
4	600	1200	200	0.00	5.26	16	0.00	*1.72*	17	0.00	**1.42**	20	0.00	1.77
5	1200	1200	67	8.00	4000	14	14.76	4000	75	7.98	**949.46**	75	8.00	*1672.52*

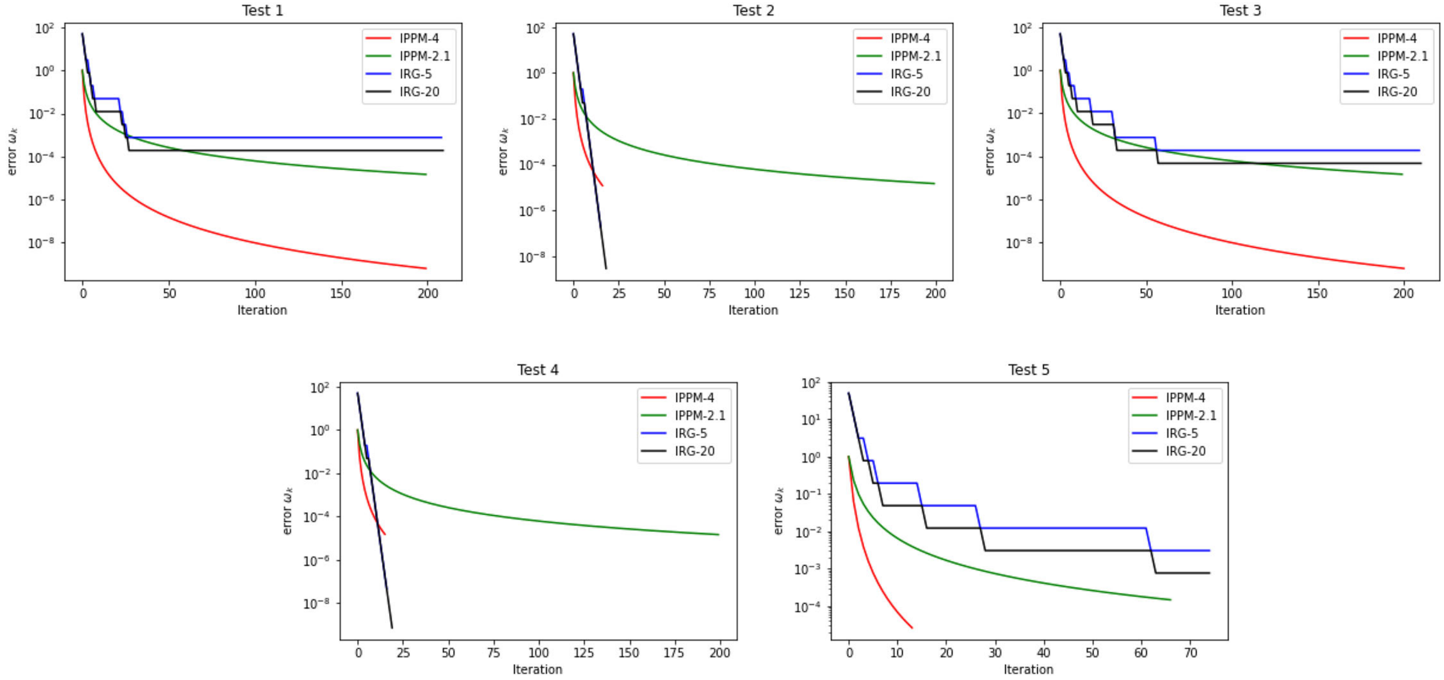

Fig. 3 Errors in proximal points calculation in iterations

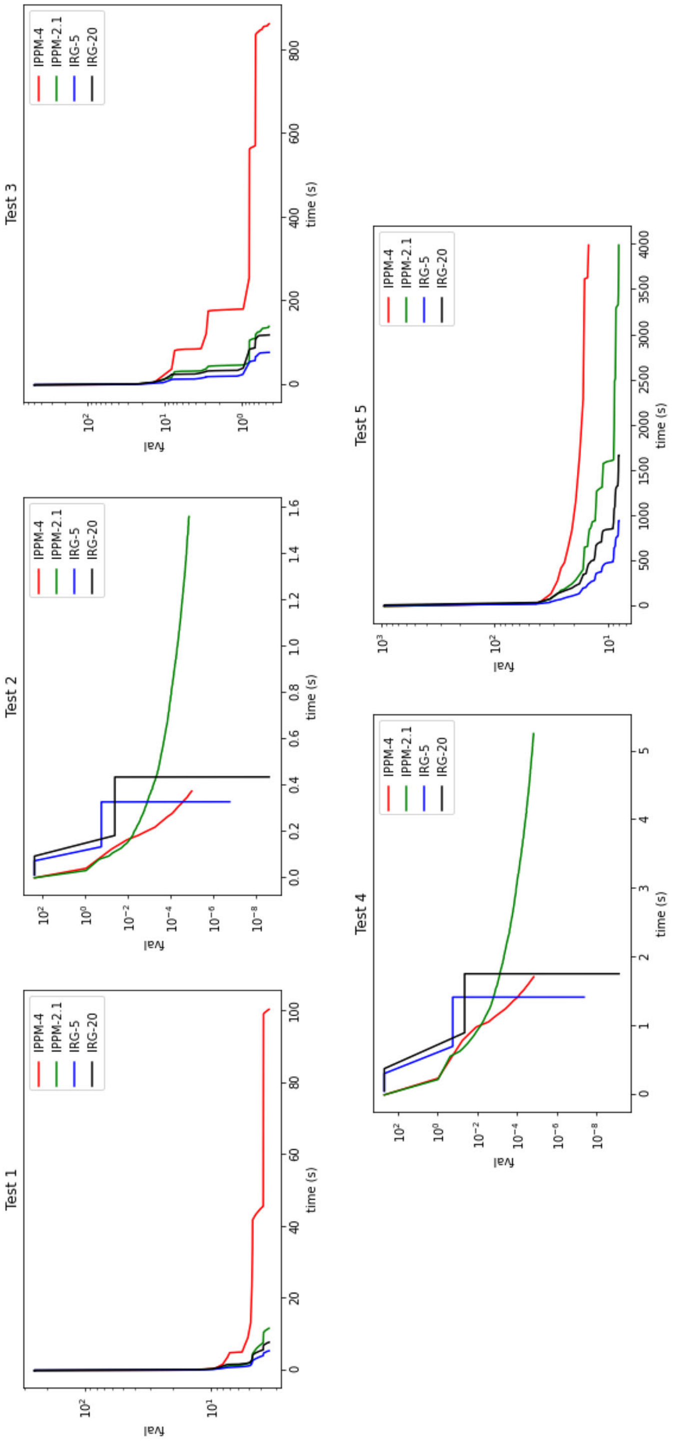

Fig. 4 Value of the objective function with respect to the computational time

- In Tests 2, 4 with $m < n$, IPPM-4 with smaller errors performs better than IPPM-2.1. In this case, the IRG methods decrease in almost every iteration and achieve smaller errors in comparison with IPPM-4.

6.2 Comparison with Exact Gradient Descent Methods

In the numerical experiments presented in this subsection, we show that our IRG method with backtracking stepsize, based on the usage of inexact gradients, performs well compared with the famous methods employing the exact gradient calculation, which are the *reduced gradient* (RG) method and *gradient descent* (GD) method in the following setting:

1. The *accuracy* of the inexact gradient g^k is *low*, i.e., $\left\| g^k - \nabla f(x^k) \right\| \leq \delta_k$, where δ_k is not too small relative to $\left\| \nabla f(x^k) \right\|$.
2. The *accuracy* required for the solution is *increasing*.

To demonstrate this, we choose the following two well-known smooth benchmark functions in global optimization taken from the survey paper [25].

- The *Dixon and Price* function is defined by

$$f_{\text{dixon}}(\mathbf{x}) := (x_1 - 1)^2 + \sum_{i=2}^{n} i \left(2x_i^2 - x_{i-1} \right)^2, \quad x \in \mathbb{R}^n.$$

  The global minimum of this function is $\bar{f}_{\text{dixon}} = 0$, and the two solutions $x^*, y^* \in \mathbb{R}^n$ are

$$\begin{cases} x_1^* = 1, \\ x_k^* = \sqrt{\dfrac{x_{k-1}}{2}} \quad \text{for } k = 2, \ldots, n \end{cases}$$

  and by $y_k^* = x_k^*$ for all $k = 1, \ldots, n-1$, $y_n^* = -x_n^*$.
- The *Rosenbrock* 1 function defined by

$$f_{\text{rosen}}(\mathbf{x}) := \sum_{i=1}^{n-1} \left[100 \left(x_{i+1} - x_i^2 \right)^2 + (x_i - 1)^2 \right], \quad x \in \mathbb{R}^n.$$

  The global minimum of this function is $\bar{f}_{\text{rosen}} = 0$, and the unique solution is $(1, \ldots, 1) \in \mathbb{R}^n$.

Since the information about the convexity and the Lipschitz continuity of gradients of the chosen objective functions is *unknown*, our experiments are conducted by algorithms, where stepsizes are obtained from the corresponding linesearches. We use the following abbreviations:

- GD: *Gradient descent method with the backtracking linesearch.*

- RGB and IRGB: *Reduced gradient method with the backtracking linesearch and Inexact reduced gradient method with the backtracking linesearch*; see (5.1).

To generate the inexactness for testing purposes, given the gradient error $\delta_k := \min\{\varepsilon_k, \rho_k\}$ as in (4.1), we create an inexact gradient g^k by adding a random vector with the norm $0.5\delta_k$ to the exact gradient $\nabla f(x^k)$. To ensure manually controlled errors between the exact gradients and inexact ones that do not decrease so fast, we choose $\rho_k := 1/\log(k+1)$. For all the methods in our experiments, the linesearch parameters are chosen as $\beta = 0.7$ and $\gamma = 0.5$. The initial radii $\varepsilon_1 = r_1 = 5$ and the radius reduction factors $\theta = 0.7$, $\mu = 0.7$ are also used for the RG and IRG methods. To avoid the initial points from being identical with the solutions, we choose $x^1 := 0_{\mathbb{R}^n}$ on tests using the Rosenbrock 1 function. In the tests using the Dixon and Price functions, we choose $x^1 := 1_{\mathbb{R}^n}$ to avoid the algorithms from going to different solutions. The condition

$$\|\nabla f(x)\| \le \nu, \quad \text{where either } \ \nu = 0.01 \text{ or } \nu = 0.001.$$

is used as the stopping criterion for all the tests. The detailed information of the numerical experiments and the achieved numerical results is presented in Table 2. The problem names are given in the forms Dn and Rn, where D stands for Dixon and Price, R stands for Rosenbrock 1, and n is the dimension of the tested problem. In these tables, ‘Iter,’ ‘fval’ stand for the number of iterations and the function value at the last iteration.

It can be seen that the performance of the IRG and RG methods in Tests D200 with $\nu = 0.01$ and $\nu = 0.001$ is better than that of the GD method, while the latter is more efficient in the other tests. It is reasonable that GD usually performs better since it uses the exact gradient, while RGB uses the reduced gradient and IRGB uses even the inexact one. In the worst case in Test R1000 with $\nu = 0.01$, the number of iterations of IRGB is equal around 1.3 times that of GD. It shows that IRGB does not suffer much from the use of inexact gradient compared with the performance of GD using the exact gradient. Table 2 also shows that the decrease of ν in 10 times results in the increase in the number of iterations in IRGB with the rate at most 1.7, where the worst case corresponds to the tests D500. This rate is similar to the rate obtained by the GD method in these tests, which confirms that our IRG method with the backtracking stepsize does not suffer from error accumulation.

The graphs below show that the errors δ_k of the inexact gradient used in IRGB are automatically adjusted to be not too small or too large compared with $\left\|\nabla f(x^k)\right\|$. This confirms the intuitive conclusion on the IRG methods discussed in Remark 4.1(ii). Figure 5 shows that the selections of errors $\delta_k = k^{-p}$, $p \ge 1$ in the existing methods [10, 20, 23] do not fit the unexpected fluctuations in the norm of the exact gradient given in the tests using the Rosenbrock function, which may lead to the over approximation and under approximation issues discussed in Sect. 1.

Table 2 Comparison with Dixon and Price and Rosenbrock 1

Name	ν	GD		IRGB				RGB		
		Iter	fval	Iter	fval	ε_k	δ_k	Iter	fval	ε_k
D200	0.01	1928	2.8E−05	1110	2.6E−05	4.0E−03	2.0E−03	998	2.76E−05	4.0E−03
D500	0.01	3831	2.8E−05	4782	2.7E−05	4.0E−03	2.0E−03	5012	2.43E−05	4.0E−03
D1000	0.01	7655	2.8E−05	9347	2.6E−05	4.0E−03	2.0E−03	9271	2.58E−05	4.0E−03
R200	0.01	20,357	9.5E−05	24,966	9.2E−05	4.0E−03	2.0E−03	25,162	8.65E−05	4.0E−03
R500	0.01	46,135	9.3E−05	59,768	7.3E−05	4.0E−03	2.0E−03	59,604	9.20E−05	4.0E−03
R1000	0.01	89,130	9.4E−05	117,754	8.8E−05	4.0E−03	2.0E−03	117,845	9.09E−05	4.0E−03
D200	0.001	3294	2.7E−07	1870	2.7E−07	4.7E−04	2.3E−04	1704	2.8E−07	4.7E−04
D500	0.001	6543	2.8E−07	8151	2.4E−07	4.7E−04	2.3E−04	7933	2.2E−07	4.7E−04
D1000	0.001	13,078	2.8E−07	15,946	2.7E−07	4.7E−04	2.3E−04	15,598	2.3E−07	4.7E−04
R200	0.001	22,664	9.6E−07	26,958	8.9E−07	4.7E−04	2.3E−04	27,395	9.7E−07	4.7E−04
R500	0.001	48,442	9.5E−07	61,998	7.8E−07	4.7E−04	2.3E−04	61,875	9.6E−07	4.7E−04
R1000	0.001	91,431	9.6E−07	119,687	9.0E−07	4.7E−04	2.3E−04	120,321	9.6E−07	4.7E−04

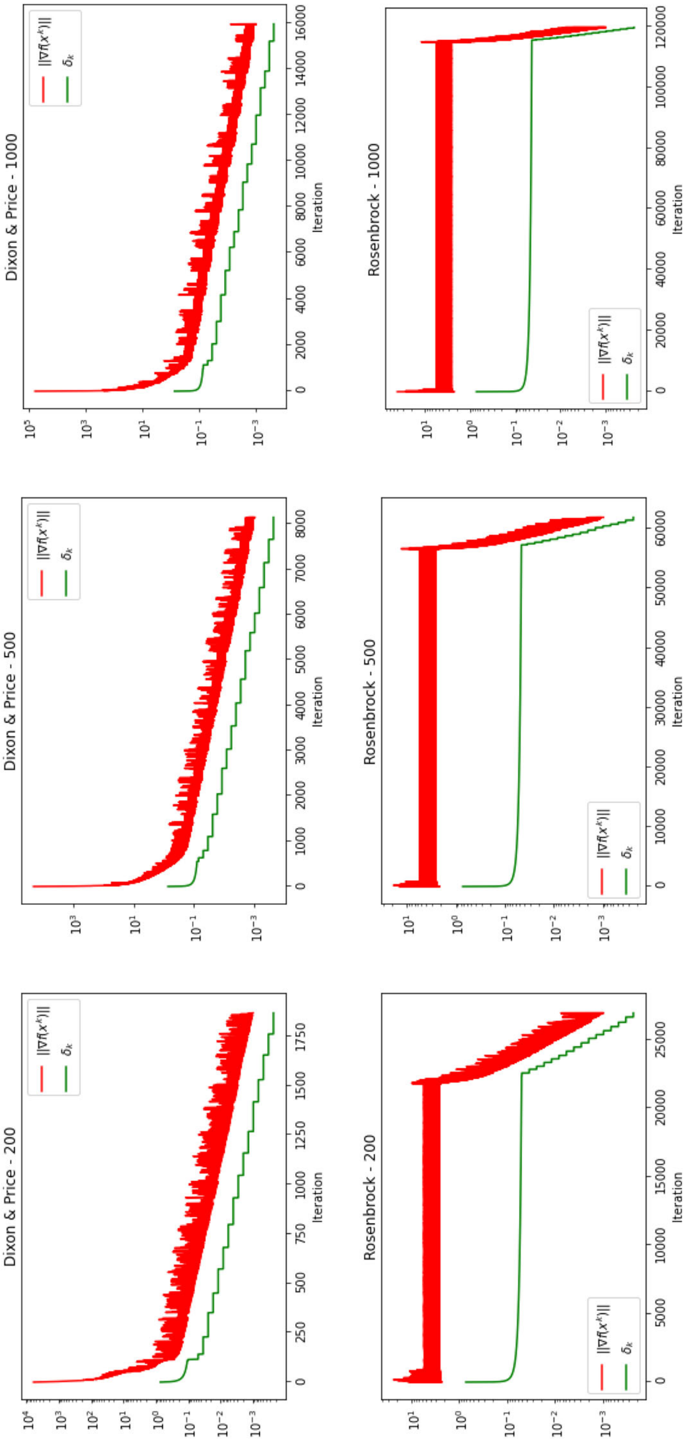

Fig. 5 Errors of IRGB compared with the norm of exact gradient

7 Conclusions and Further Research

In this paper, we propose and develop the inexact reduced gradient methods with different stepsize selections to solve problems of nonconvex optimization. These methods achieve stationary accumulation points and, under additional assumptions on the KL property of the objective functions, the global linear convergence. The convergence analysis of the developed algorithms is based on novel convergence results established for general linesearch methods. The theoretical and numerical comparisons show that our methods do not suffer much from the error accumulation and are able to automatically adjust the errors in the exact gradient approximations to get a better performance than the existing methods using common selections of errors.

In our future research, we aim at developing the IRG methods in different directions, which include designing zeroth-order algorithms by using practical methods for approximating gradients, designing inexact versions of methods frequently used in nonconvex nonsmooth optimization, e.g., the proximal point and proximal gradient methods, and also designing appropriate IRG methods for problems of constrained optimization. The obtained results would allow us to develop new applications to important classes of models in machine learning, statistics, and related disciplines.

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

References

1. Absil, P.-A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM J. Optim. **16**, 531–547 (2005)
2. Addis, A., Cassioli, A., Locatelli, M., Schoen, F.: A global optimization method for the design of space trajectories. Comput. Optim. Appl. **48**, 635–652 (2011)
3. Aragón Artacho, F.J., Fleming, R.M.T., Vuong, P.T.: Accelerating the DC algorithm for smooth functions. Math. Program. **169**, 95–118 (2018)
4. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**, 5–16 (2009)
5. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka–Łojasiewicz property. Math. Oper. Res. **35**, 438–457 (2010)
6. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**, 330–348 (2017)
7. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, Cham (2017)
8. Beck, A.: First-Order Methods in Optimization. SIAM, Philadelphia (2017)
9. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
10. Bertsekas, D.P.: Nonlinear Programming, 3rd edn. Athena Scientific, Belmont (2016)
11. Bloomfield, P., Steiger, W.: Least absolute deviations curve fitting. SIAM J. Sci. Stat. Comput. **1**, 290–301 (1980)
12. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**, 1205–1223 (2006)
13. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**, 223–311 (2018)
14. Burke, J.V., Lewis, A.S., Overton, M.L.: Two numerical methods for optimizing matrix stability. Linear Algebra Appl. **351–352**, 147–184 (2002)

15. Burke, J.V., Lin, Q.: Convergence of the gradient sampling algorithm on directionally Lipschitz functions. Set-Valued Var. Anal. **29**, 949–966 (2021)
16. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. MOS-SIAM Optimization Series, Philadelphia (2008)
17. Crockett, J.B., Chernoff, H.: Gradient methods of maximization. Pac. J. Math. **5**, 33–50 (1955)
18. Curry, H.B.: The method of steepest descent for non-linear minimization problems. Q. Appl. Math. **2**, 258–261 (1944)
19. Dan, H., Yamashita, N., Fukushima, M.: Convergence properties of the inexact Levenberg–Marquardt method under local error bound conditions. Optim. Methods Softw. **17**, 605–626 (2002)
20. Devolder, O., Glineur, F., Nesterov, Yu.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**, 37–75 (2014)
21. Facchinei, F., Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, vol. II. Springer, New York (2003)
22. Gannot, O.: A frequency-domain analysis of inexact gradient methods. Math. Program. **194**, 975–1016 (2022)
23. Gilmore, P., Kelley, C.T.: An implicit filtering algorithm for optimization of functions with many local minima. SIAM J. Optim. **5**, 269–285 (1995)
24. Izmailov, A.F., Solodov, M.V.: Newton-Type Methods for Optimization and Variational Problems. Springer, New York (2014)
25. Jamil, M.: A literature survey of benchmark functions for global optimization problems. Int. J. Math. Model. Numer. Optim. **4**, 150–194 (2013)
26. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In: Frasconi, P. et al. (eds.) Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, Part 1. Springer, Cham, pp. 795–811 (2016)
27. Khanh, P.D., Mordukhovich, B.S., Phat, V.T., Tran, D.B.: Generalized damped Newton algorithms in nonsmooth optimization via second-order subdifferentials. J. Glob. Optim. **86**, 93–122 (2023)
28. Khanh, P.D., Mordukhovich, B.S., Phat, V.T., Tran, D.B.: Globally convergent coderivative-based generalized Newton methods in nonsmooth optimizations. Math. Program. (2023). https://doi.org/10.1007/s10107-023-01980-2
29. Khanh, P.D., Mordukhovich, B.S., Phat, V.T., Tran, D.B.: A new inexact gradient descent method with applications to nonsmooth convex optimization. arXiv:2303.08785
30. Kiwiel, K.C.: Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. SIAM J. Optim. **18**, 379–388 (2007)
31. Kiwiel, K.C.: A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. SIAM J. Optim. **20**, 1983–1994 (2010)
32. Kiwiel, K.C.: Improved convergence result for the discrete gradient and secant methods for nonsmooth optimization. J. Optim. Theory Appl. **144**, 69–75 (2010)
33. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. Inst. Fourier **48**, 769–783 (1998)
34. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence for alternating and averaged nonconvex projections. Found. Comput. Math. **9**, 485–513 (2009)
35. Lobanov, A., Gasnikov, A., Stonyakin, F.: Highly smoothness zero-order methods for solving optimization problems under PL condition. arXiv:2305.15828 (2023)
36. Łojasiewicz, S.: Ensembles Semi-analytiques. Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette (Seine-et-Oise) (1965)
37. Nesterov, Yu.: Universal gradient methods for convex optimization problems. Math. Program. **152**, 381–404 (2015)
38. Nesterov, Yu.: Lectures on Convex Optimization, 2nd edn. Springer, Cham (2018)
39. Nielsen, M.A.: Neural Networks and Deep Learning. Determination Press, New York (2015)
40. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2016)
41. Noll, D.: Convergence of non-smooth descent methods using the Kurdyka–Łojasiewicz inequality. J. Optim. Theory Appl. **160**, 553–572 (2014)
42. Ostrowski, A.: Solution of Equations and Systems of Equations, 2nd edn. Academic Press, New York (1966)
43. Polyak, B.T.: Gradient methods for minimizing functionals. USSR Comput. Math. Math. Phys. **3**, 864–878 (1963)

44. Polyak, B.T.: Iterative Algorithms for Singular Minimization Problems. Nonlinear Programming, vol. 4, pp. 147–166. Academic Press, London (1981)
45. Polyak, B.T.: Introduction to Optimization. Optimization Software, New York (1987)
46. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control. Optim. **14**, 877–898 (1976)
47. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
48. Rotaru, T., Glineur, F., Patrinos, P.: Tight convergence rates of the gradient method on hypoconvex functions. https://doi.org/10.48550/arXiv.2203.00775
49. Ruder, S.: An overview of gradient descent optimization algorithms. https://doi.org/10.48550/arXiv:1609.04747
50. Themelis, A., Stella, L., Patrinos, P.: Forward-backward quasi-Newton methods for nonsmooth optimization problems. Comput. Optim. Appl. **67**, 443–487 (2017)
51. Vasin, A., Gasnikov, A., Dvurechensky, P., Spokoiny, V.: Accelerated gradient methods with absolute and relative noise in the gradient. Optim. Methods Softw. (2023). https://doi.org/10.1080/10556788.2023.2212503
52. Xingyu, Z.: On the Fenchel duality between strong convexity and Lipschitz continuous gradient. https://doi.org/10.48550/arXiv.1803.06573

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Pham Duy Khanh[1] · **Boris S. Mordukhovich**[2] · **Dat Ba Tran**[2]

Dat Ba Tran
tranbadat@wayne.edu

[1] Group of Analysis and Applied Mathematics, Department of Mathematics, Ho Chi Minh City University of Education, Ho Chi Minh City, Vietnam

[2] Department of Mathematics, Wayne State University, Detroit, MI, USA